Learning Survival Models with Right-Censored Reporting Delays

Yuta Shikuri
The Graduate University for Advanced Studies
Tokio Marine Holdings, Inc.
shikuriyuta@gmail.com

Hironori Fujisawa
The Institute of Statistical Mathematics
The Graduate University for Advanced Studies
RIKEN
fujisawa@ism.ac.jp

Abstract

Survival analysis provides statistical methods to model the time until an event occurs. Reporting delays arise when event times are not observed at their occurrence but are only revealed upon reporting. This issue is particularly critical for timely risk evaluation when the observation window is short due to administrative censoring. In this study, we incorporate right-censored reporting delays by jointly modeling parametric hazards for the event and reporting processes. We then construct a consistent estimator for the model parameters and develop a Monte Carlo expectation-maximization algorithm to compute it. To address the challenges posed by administrative censoring, we leverage these findings and propose a transfer-learning procedure. Experimental results demonstrate that our method improves the accuracy of timely risk evaluation under administrative censoring.

1 Introduction

Survival analysis is a statistical framework for modeling and estimating the time until an event occurs [Kalbfleisch and Prentice, 2002]. Its applications span a wide range of fields, including health sciences, marketing, reliability engineering, and finance. Beyond the statistics community, machine-learning methods have been applied to survival analysis [Wang et al., 2019]. The proportional hazards model decomposes the hazard function into a time-dependent baseline hazard and a time-independent component [Cox, 1972, 1975, Breslow, 1972]. Modern approaches in the machine-learning community have extended this model with neural networks to improve predictive performance [Kvamme et al., 2019, Zhong et al., 2021, Nagpal et al., 2021]. Appendix B.1 reviews the proportional hazards framework and related developments.

Censoring is a key concept in survival analysis; for some individuals, the exact event time is unknown and only partial information is available. For timely risk evaluation, follow-up is often right-censored at a fixed assessment date, a phenomenon known as administrative censoring. When the resulting enrollment-to-evaluation window is short, the identifiability condition required for the asymptotic consistency of the parameter estimator may be violated. In such settings, borrowing information from related cohorts can supplement the limited information available in the target domain. Motivated by this idea, recent studies have explored transfer learning and domain adaptation for survival models [Li et al., 2016, Bellot and van der Schaar, 2019, Shaker and Lawrence, 2023, Steinberg et al., 2024, Zhao et al., 2025].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Reporting delays are also important in survival analysis; a key difference from the standard setting is that the event occurrence time is revealed only after reporting. The biostatistical literature emphasizes accounting for these reporting delays, since ignoring them can lead to underestimation of the hazard of event occurrence. In typical clinical trials, vital status is often ascertained only at scheduled follow-up visits, resulting in a time lag between death occurrence and its reporting. To account for such delays, several studies have proposed estimators as alternatives to the naive Kaplan-Meier estimator [Hu and Tsiatis, 1996, Laan and Hubbard, 1998]. In infectious disease research, individuals may remain unobserved until infection is detected and reported, so the at-risk cohort may not be known in advance. A series of studies addresses this issue under right truncation [Lagakos et al., 1988, Kalbfleisch and Lawless, 1989, Harris, 1990, Kalbfleisch and Lawless, 1991, Pagano et al., 1994]. Similar delay structures have also been studied beyond survival analysis; see Appendix B.2.

As illustrated in Figure 1, we consider survival analysis with right-censored reporting delays: individuals are known to be at risk, but the event occurrence time is observed only when it is reported at unscheduled times. This observation scheme is plausible in certain real-world applications, where an event of interest triggers an unscheduled report that provides an opportunity to collect detailed information about its occurrence. For instance, individuals may be followed from enrollment until death, while disease onset is ascertained only at death based on the recorded cause. This example is closely related to survival-sacrifice experiments [Kodell and Nelson, 1980, Turnbull and Mitchell, 1984, Laan et al., 1997], with one key distinction: we assume that the exact onset time becomes observable at death. Although Buchardt et al. [2025] considered the same observation scheme, they estimated the event-occurrence and reporting processes separately without examining whether the parameters governing these processes are jointly identifiable.

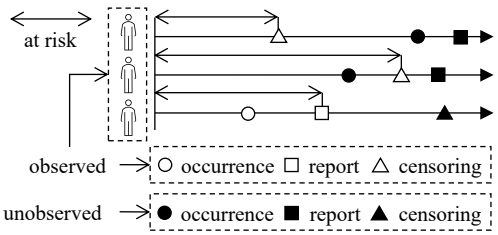

Figure 1: Illustration of right-censored reporting delays. The figure shows three possible observation patterns over the study period. The detailed setting is provided in Section 3.1.

In this study, we develop a survival analysis framework for jointly estimating the hazards of event occurrence and reporting under right-censored reporting delays. In the absence of administrative censoring, we establish the asymptotic consistency of the resulting parameter estimator under a mild identifiability condition. However, when the enrollment-to-evaluation window is short, administrative censoring makes identification more challenging. To address this issue, we adopt the transfer-learning procedure illustrated in Figure 2. Specifically, we consider two domains: a source domain without administrative censoring and a target domain subject to it. In the first stage, we estimate all parameters using observations from the source domain. In the second stage, we transfer selected components of the source-domain parameter estimates to the target domain and estimate the remaining components using target-domain observations. Under appropriate transfer conditions, the resulting estimator is consistent. As a specialization of this transfer-learning procedure, we derive an approximate estimating equation for parametric proportional hazards models [Holford, 1980, Royston and Parmar, 2002]. Experiments demonstrate that our approach improves the accuracy of timely risk evaluation under administrative censoring. Appendix D provides the proofs of the theoretical results.

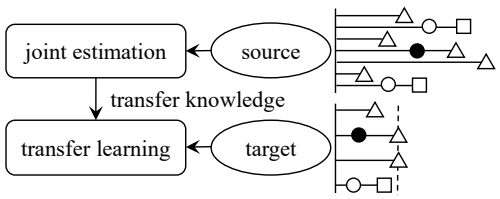

Figure 2: Illustration of the transfer-learning procedure. The procedure enables consistent estimation of the survival-model parameters in the target domain under administrative censoring. The formal setting is given in Section 4.1.

2 Standard Survival Analysis

This section reviews standard survival analysis in the absence of reporting delays. The derivations of the estimators and the proofs of their consistency are provided in Appendix C, which serves as background for the proofs of our main results.

Throughout this paper, random variables are denoted by uppercase letters and their realizations by lowercase letters, with indices and accents omitted when no confusion arises. Let P_X denote the distribution of the covariates $X \in \mathcal{X} \subset \mathbb{R}^d$ in the domain under consideration. All probability density functions are assumed to be strictly positive and bounded. The true parameter values are assumed to lie in the interiors of their respective parameter spaces. We also assume the standard regularity conditions required for the asymptotic arguments below, including uniform convergence and appropriate conditions for the consistency of the maximizers. An asterisk superscript denotes a true parameter value, e.g., $\boldsymbol{\theta}^*$ for $\boldsymbol{\theta}$, whereas a prime superscript denotes another element of the same parameter space, e.g., $\boldsymbol{\theta}'$. A semicolon explicitly specifies parameters, as in $f(y; \boldsymbol{\theta})$, while a vertical bar indicates conditioning on observed covariates, as in $f(y \mid \boldsymbol{x})$.

Survival analysis models the time until an event occurs as a non-negative random variable T. For $t \in [0, \infty)$ and $\boldsymbol{x} \in \mathcal{X}$, the hazard function of T is defined as $h(t \mid \boldsymbol{x}) \coloneqq f(t \mid \boldsymbol{x}) S(t \mid \boldsymbol{x})^{-1}$, where $f(t \mid \boldsymbol{x})$ denotes the conditional probability density function of T given $\boldsymbol{x}$, and $S(t \mid \boldsymbol{x}) \coloneqq 1 - \int_0^t f(s \mid \boldsymbol{x}) ds$ is the corresponding survival function. The censoring time is modeled as a non-negative random variable C, assumed to be conditionally independent of T given X. Define the observation time as $Y \coloneqq \min\{T, C\}$ and the event indicator as $V \coloneqq \mathbb{I}(T \leq C)$, where $\mathbb{I}$ denotes the indicator function. For $y \in [0, \infty)$, $v \in \{0, 1\}$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Y, V) is given by

$$g(y, v \mid \boldsymbol{x}) \coloneqq \begin{cases} \frac{\partial}{\partial y} \Pr(Y \leq y, V = 0 \mid X = \boldsymbol{x}) = S(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) & (v = 0), \\ \frac{\partial}{\partial y} \Pr(Y \leq y, V = 1 \mid X = \boldsymbol{x}) = f(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}) & (v = 1), \end{cases} \tag{1}$$

where f_c and S_c denote the conditional probability density function and the survival function of C, respectively. Administrative censoring refers to situations where observations are censored at a predetermined time point $\tau \in [0, \infty)$. Let $\bar{Y} \coloneqq \min\{Y, \tau\}$ and $\bar{V} \coloneqq \mathbb{I}(T \leq C, T \leq \tau)$. For $y \in [0, \tau]$, $v \in \{0, 1\}$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$\bar{g}(y, v \mid \boldsymbol{x}) \coloneqq \begin{cases} \Pr(\bar{Y} = \tau, \bar{V} = 0 \mid X = \boldsymbol{x}) = S(\tau \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}) & (y = \tau, v = 0), \\ \frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 0 \mid X = \boldsymbol{x}) = S(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) & (y \neq \tau, v = 0), \\ \frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 1 \mid X = \boldsymbol{x}) = f(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}) & (v = 1). \end{cases} \tag{2}$$

In this study, we parameterize the hazard functions of T and C by unknown real-valued vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$, respectively. Parametric proportional hazards models are widely used across many fields. They specify the hazard function as $h_b(t; \boldsymbol{\alpha}) \phi(\boldsymbol{x}; \boldsymbol{\beta})$ for $t \in [0, \infty)$ and $\boldsymbol{x} \in \mathcal{X}$, where the baseline hazard h_b and the time-independent component ϕ are strictly positive functions parameterized by unknown real-valued vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively.

The parameters characterizing the hazard functions of T and C are estimated from independent and identically distributed observations $\mathcal{D} \coloneqq (\boldsymbol{x}_i, y_i, v_i)_{i=1}^n$. Each observation is characterized by the covariates $\boldsymbol{x}_i \in \mathcal{X}$, the time $y_i \in [0, \infty)$ until the event or censoring, and the event indicator $v_i \in \{0, 1\}$. Then the log likelihood in the absence of administrative censoring is given by $\mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c(\boldsymbol{\theta}_c)$, where the two components are defined as

$$\mathcal{L}(\boldsymbol{\theta}) \coloneqq \sum_{i=1}^n \Big((1 - v_i) \log S(y_i \mid \boldsymbol{x}_i) + v_i \log f(y_i \mid \boldsymbol{x}_i) \Big), \tag{3}$$

$$\mathcal{L}_c(\boldsymbol{\theta}_c) \coloneqq \sum_{i=1}^n \Big((1 - v_i) \log f_c(y_i \mid \boldsymbol{x}_i) + v_i \log S_c(y_i \mid \boldsymbol{x}_i) \Big). \tag{4}$$

Based on the discrepancy between Equations (1) and (2), we derive the corresponding likelihood function under administrative censoring. For simplicity, we do not distinguish between realizations of (Y, V) and $(\bar{Y}, \bar{V})$, and collectively denote them by (y_i, v_i). In the presence of administrative censoring, y_i is understood as the censored observation taking values in $[0, \tau]$. As an alternative to $\mathcal{L}_c$, we introduce

$$\mathcal{L}_c^\dagger(\boldsymbol{\theta}_c) \coloneqq \sum_{i=1}^n \Big((1 - v_i^\dagger) \log f_c(y_i \mid \boldsymbol{x}_i) + v_i^\dagger \log S_c(y_i \mid \boldsymbol{x}_i) \Big), \tag{5}$$

where $v_i^\dagger \coloneqq \mathbb{I}(y_i \geq \tau(1 - v_i))$. The log likelihood in the presence of administrative censoring is given by $\mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c^\dagger(\boldsymbol{\theta}_c)$. The estimator of $\boldsymbol{\theta}$ is obtained by maximizing $\mathcal{L}(\boldsymbol{\theta})$, regardless of whether

administrative censoring is absent or present. Provided that the density model $f(t;\boldsymbol{\theta})$ is identifiable on $[0,\infty)$ in the absence of administrative censoring, or on $[0,\tau]$ in its presence, this estimator is asymptotically consistent. This property is useful when the primary focus is on the hazard of T, and that of C is regarded as a nuisance. However, when the administrative censoring time τ is short, the identifiability condition can be restrictive for certain survival models. For instance, in piecewise-proportional hazards models [Friedman, 1982], the baseline hazard parameters corresponding to intervals beyond τ are not identifiable.

3 Survival Analysis under Right-Censored Reporting Delays

This section presents a general framework for survival analysis under right-censored reporting delays.

3.1 Problem Setting

In this study, we consider a setting in which two events occur sequentially over time. The earlier and later events correspond to the event of interest and its reporting, respectively. We jointly model the time to the earlier event and the subsequent time from the earlier to the later event.

We use subscripts 1 and 2 for the symbols $(T, h, f, S, \boldsymbol{\theta})$ to distinguish the earlier and later events, respectively. We assume that T_1, T_2, and C are mutually independent conditional on X, and that $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_c$ are unknown. For the earlier event, we let $Z \coloneqq \min\{T_1, C\}, W \coloneqq \mathbb{I}(T_1 \leq C), \bar{Z} \coloneqq \min\{Z, \tau\}$, and $\bar{W} \coloneqq \mathbb{I}(T_1 \leq C, T_1 \leq \tau)$. For the later event, we define $Y, V, \bar{Y}$, and $\bar{V}$ by substituting $T = T_1 + T_2$ into their respective expressions. Each observation is characterized by the covariates $\boldsymbol{x}_i \in \mathcal{X}$, the time $z_i \in [0, y_i]$ until the earlier event or censoring, the event indicator $w_i \in \{0,1\}$ for the earlier event, the time y_i until the later event or censoring, and the event indicator v_i for the later event. For simplicity, the same notation (z_i, y_i, w_i, v_i) is used to represent the realizations of both (Z, Y, W, V) and $(\bar{Z}, \bar{Y}, \bar{W}, \bar{V})$. A similar setting involving two successive events under right censoring is also discussed by Visser [1996]. We additionally consider the setting in which the earlier event is observed only if the later event occurs, as illustrated in Figure 1. The triplet $(\boldsymbol{x}_i, y_i, v_i)$ is observed for $1 \leq i \leq n$, while the pair (z_i, w_i) is observed only if $v_i = 1$. The probability of observing $(w, v) = (0, 1)$ is zero by the definitions of (W, V) and $(\bar{W}, \bar{V})$. Figure 3 illustrates how our notation corresponds to the three possible observation patterns under this extended setting. For simplicity, we set $v_i = 0$ for $1 \leq i \leq m$ and $v_i = 1$ for $m+1 \leq i \leq n$.

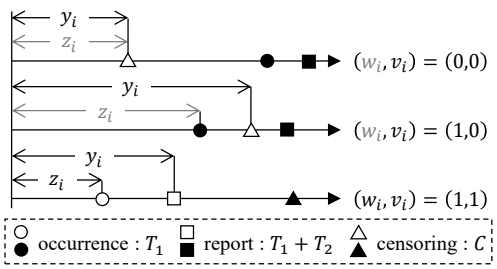

Figure 3: Notation for observations. The figure presents the notation for the three observation patterns shown in Figure 1. In the top two patterns with $v_i = 0$, the pair (z_i, w_i) remains unobserved.

3.2 Estimator Formulation

We first construct the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in the absence of administrative censoring. The corresponding density model can be derived through a straightforward calculation.

Lemma 1. *For $0 \leq y$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Y, V) is given by*

$$g_{\blacksquare}(y \mid \boldsymbol{x}) \coloneqq \frac{\partial}{\partial y} \Pr(Y \leq y, V = 0 \mid X = \boldsymbol{x}) = S_{\blacksquare}(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}), \tag{6}$$

where $S_{\blacksquare}(y \mid \boldsymbol{x}) \coloneqq S_1(y \mid \boldsymbol{x}) + \int_0^y f_1(t \mid \boldsymbol{x}) S_2(y - t \mid \boldsymbol{x}) dt$. For $0 \leq z \leq y$, $w = v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Z, Y, W, V) is given by

$$g_{\square}(z, y \mid \boldsymbol{x}) \coloneqq \frac{\partial^2}{\partial y \partial z} \Pr(Z \leq z, Y \leq y, W = 1, V = 1 \mid X = \boldsymbol{x}) = f_{\square}(z, y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}), \tag{7}$$

where $f_{\square}(z, y \mid \boldsymbol{x}) \coloneqq f_1(z \mid \boldsymbol{x}) f_2(y - z \mid \boldsymbol{x})$.

In Lemma 1, we specify the joint distribution only for the observed variables. Accordingly, we can write the likelihood for the observed data as follows.

Proposition 1. *The log likelihood in the absence of administrative censoring is $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c(\boldsymbol{\theta}_c)$, where $\mathcal{E}$ is defined as*

$$\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \coloneqq \sum_{i=1}^{n} \Big((1 - v_i) \log S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) + v_i \log f_{\square}(z_i, y_i \mid \boldsymbol{x}_i) \Big). \tag{8}$$

The parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ can be estimated by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, since $\mathcal{L}_c(\boldsymbol{\theta}_c)$ does not depend on $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Although the structure of the joint distribution resembles that of Equation (1), it involves a newly introduced pair of functions, $(S_{\blacksquare}, f_{\square})$. In this case, the identifiability of (f_1, f_2) with respect to $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is not straightforward, since (Z, W) is observed only when $V = 1$. As a result, proving the consistency of the estimator is nontrivial. Nevertheless, it can be ensured under the standard identifiability condition.

Assumption 1. If $f_1(t \mid \boldsymbol{x}; \boldsymbol{\theta}_1) = f_1(t \mid \boldsymbol{x}; \boldsymbol{\theta}_1')$ and $f_2(t \mid \boldsymbol{x}; \boldsymbol{\theta}_2) = f_2(t \mid \boldsymbol{x}; \boldsymbol{\theta}_2')$ for Lebesgue-almost every $t \in [0, \infty)$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$.

Theorem 1. *Under Assumption 1, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in the absence of administrative censoring is asymptotically consistent.*

As in survival analysis without reporting delays, censoring can be ignored when estimating the event hazards. Appendix E shows that this property remains valid even when the censoring times for the earlier and later events differ. There is another situation in which the hazard function of C is also of interest, and $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ shares some components with $\boldsymbol{\theta}_c$. Under the following assumption, we can jointly estimate $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)$.

Assumption 2. If $f_1(t \mid \boldsymbol{x}; \boldsymbol{\theta}_1) = f_1(t \mid \boldsymbol{x}; \boldsymbol{\theta}_1')$, $f_2(t \mid \boldsymbol{x}; \boldsymbol{\theta}_2) = f_2(t \mid \boldsymbol{x}; \boldsymbol{\theta}_2')$, and $f_c(t \mid \boldsymbol{x}; \boldsymbol{\theta}_c) = f_c(t \mid \boldsymbol{x}; \boldsymbol{\theta}_c')$ for Lebesgue-almost every $t \in [0, \infty)$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_c')$.

Proposition 2. *Under Assumption 2, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c(\boldsymbol{\theta}_c)$ in the absence of administrative censoring is asymptotically consistent.*

Next, we consider the case with administrative censoring. The structure of the probability model differs slightly from that described in Lemma 1.

Lemma 2. *For $0 \leq y \leq \tau$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by*

$$\bar{g}_{\blacksquare}(y \mid \boldsymbol{x}) \coloneqq \begin{cases} \Pr(\bar{Y} = \tau, \bar{V} = 0 \mid X = \boldsymbol{x}) = S_{\blacksquare}(\tau \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}) & (y = \tau), \\ \frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 0 \mid X = \boldsymbol{x}) = S_{\blacksquare}(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) & (y \neq \tau). \end{cases} \tag{9}$$

For $0 \leq z \leq y \leq \tau$, $w = v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Z}, \bar{Y}, \bar{W}, \bar{V})$ is given by

$$\bar{g}_{\square}(z, y \mid \boldsymbol{x}) \coloneqq \frac{\partial^2}{\partial y \partial z} \Pr(\bar{Z} \leq z, \bar{Y} \leq y, \bar{W} = 1, \bar{V} = 1 \mid X = \boldsymbol{x}) = f_{\square}(z, y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}). \tag{10}$$

Proposition 3. *The log likelihood in the presence of administrative censoring is $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c^{\dagger}(\boldsymbol{\theta}_c)$.*

The presence of administrative censoring results in the replacement of the density f_c in Equation (6) at $y = \tau$ by the survival function S_c. Then the log likelihood factorization allows us to estimate the hazard functions of T_1 and T_2 without requiring the censoring time distribution. The following theorem shows that the consistency requires a more restrictive condition than Assumption 1.

Assumption 3. If $f_1(t_1 \mid \boldsymbol{x}; \boldsymbol{\theta}_1) f_2(t_2 \mid \boldsymbol{x}; \boldsymbol{\theta}_2) = f_1(t_1 \mid \boldsymbol{x}; \boldsymbol{\theta}_1') f_2(t_2 \mid \boldsymbol{x}; \boldsymbol{\theta}_2')$ for Lebesgue-almost every $t_1, t_2 \in [0, \tau]$ satisfying $t_1 + t_2 \leq \tau$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$.

Theorem 2. *Under Assumption 3, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in the presence of administrative censoring is asymptotically consistent.*

Because observations are restricted to the finite time interval $[0, \tau]$, we can no longer exploit the normalization property that a density integrates to one over its entire support. Consequently, Assumption 3 cannot be decomposed into separate identifiability assumptions for f_1 and f_2, as in Assumption 1; instead, it is stated in terms of the identifiability of their product. This product-form

requirement does not substantially change the nature of the identifiability difficulty: it mainly excludes special classes of hazard functions, such as the example in Appendix F.1. Rather, similarly to the setting without reporting delays, the essential difficulty of Assumption 3 lies in establishing identifiability when the administrative censoring time τ is short. The transfer-learning approach proposed in Section 4 is designed to address this difficulty induced by administrative censoring.

3.3 Optimization

One approach to jointly estimating the earlier and later hazard functions is to directly maximize $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. However, computing its gradient with respect to the parameters is often challenging, as $\log S_{\blacksquare}(y \mid \boldsymbol{x})$ does not necessarily admit a closed-form expression. To overcome this problem, we maximize $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ using a Monte Carlo expectation-maximization (EM) algorithm [Wei and Tanner, 1990]. Specifically, we employ a lower bound on $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ derived from Jensen's inequality [Billingsley, 1995] and approximate the resulting expectations via Monte Carlo integration. This yields a tractable optimization procedure. We begin by establishing the following theorem, which forms the basis of an EM algorithm [Dempster et al., 1977].

Theorem 3. *Let $\mathcal{L}_1(\boldsymbol{\theta}_1 \mid \mathcal{D}_1)$ denote the log-likelihood obtained from Equation* (3) *by setting $f = f_1$ and $\mathcal{D} = \mathcal{D}_1 \coloneqq (\boldsymbol{x}_i, z_i, w_i)_{i=1}^n$. Similarly, let $\mathcal{L}_2(\boldsymbol{\theta}_2 \mid \mathcal{D}_2)$ denote the corresponding log-likelihood obtained by setting $f = f_2$ and $\mathcal{D} = \mathcal{D}_2 \coloneqq (\boldsymbol{x}_i, y_i - z_i, v_i)_{i=1}^n$. Then $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is bounded below by*

$$\mathcal{LB}(q, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \coloneqq \mathcal{I}\big[\big(\mathcal{L}_1(\boldsymbol{\theta}_1 \mid \mathcal{D}_1) + \mathcal{L}_2(\boldsymbol{\theta}_2 \mid \mathcal{D}_2) - \log q\big)q\big], \tag{11}$$

where $\mathcal{I} \coloneqq \mathcal{I}_1 \circ \cdots \circ \mathcal{I}_m$, $\mathcal{I}_i[\,\cdot\,] \coloneqq [\,\cdot\,]\big|_{z_i = y_i, w_i = 0} + \int_0^{y_i} [\,\cdot\,]\big|_{w_i = 1} dz_i$, $q \coloneqq \prod_{i=1}^m q_i(z_i, w_i)$, and q_i is a strictly positive function such that $\mathcal{I}_i[q_i(z_i, w_i)] = 1$. Given fixed parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, this lower bound on $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ attains its maximum when, for each $1 \leq i \leq m$,

$$q_i(z_i, w_i) = \begin{cases} S_{\blacksquare}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{-1} S_1(z_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1) & (w_i = 0), \\ S_{\blacksquare}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)^{-1} f_1(z_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1) S_2(y_i - z_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_2) & (w_i = 1). \end{cases} \tag{12}$$

We iteratively update the density q and optimize the parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Let $(q^{(\kappa)}, \boldsymbol{\theta}_1^{(\kappa)}, \boldsymbol{\theta}_2^{(\kappa)})$ denote the iterate after the κ-th iteration, starting from the initial values at $\kappa = 0$. In the κ-th iteration, we first set the density $q = q^{(\kappa)}$ as in Equation (12), which maximizes the lower bound $\mathcal{LB}(q, \boldsymbol{\theta}_1^{(\kappa-1)}, \boldsymbol{\theta}_2^{(\kappa-1)})$. We then obtain the updated parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1^{(\kappa)}, \boldsymbol{\theta}_2^{(\kappa)})$ by maximizing the lower bound $\mathcal{LB}(q^{(\kappa)}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. In this iterative scheme, $\mathcal{E}(\boldsymbol{\theta}_1^{(\kappa)}, \boldsymbol{\theta}_2^{(\kappa)})$ is monotonically nondecreasing in κ.

The integral in Equation (11) cannot be evaluated in closed form. To resolve this issue, we adopt a Monte Carlo approximation. The missing values in $\mathcal{D}_1$ and $\mathcal{D}_2$ are imputed using samples generated from q. This procedure is repeated s times, and the resulting pseudo-complete datasets are then combined. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ denote the combined pseudo-complete datasets corresponding to $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. Equation (11) is approximated by

$$s^{-1}\mathcal{L}_1(\boldsymbol{\theta}_1 \mid \mathcal{P}_1) + s^{-1}\mathcal{L}_2(\boldsymbol{\theta}_2 \mid \mathcal{P}_2) - \mathcal{I}[q \log q]. \tag{13}$$

The formulation in Equation (13) enables the independent optimization of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ using standard full-likelihood methods, thereby facilitating the reuse of existing resources.

The remaining problem is how to efficiently generate samples from q for use in Monte Carlo integration. As stated at the end of Theorem 3, the normalization constant $S_{\blacksquare}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ appears in the sampling distribution $q_i(z_i, w_i)$. To avoid computing this constant explicitly, we adopt the rejection sampling scheme: For each $1 \leq i \leq m$, repeatedly draw $t \sim f_1(t \mid \boldsymbol{x}_i)$ until one of the following conditions is met. If the sampled earlier-event time t exceeds the censoring time y_i, we set $(z_i, w_i) = (y_i, 0)$. Otherwise, we generate an independent $U \sim \mathrm{Uniform}(0, 1)$. If $U \leq S_2(y_i - t \mid \boldsymbol{x}_i)$, we accept the sample and set $(z_i, w_i) = (t, 1)$. If $U > S_2(y_i - t \mid \boldsymbol{x}_i)$, we reject the sampled t and redraw.

4 Transfer Learning for Administratively Censored Reporting Delays

As discussed in Section 3.2, asymptotic consistency of the parameter estimator can be established under a moderate assumption in the absence of administrative censoring. In contrast, under administrative censoring, consistency requires a restrictive identifiability condition. This section presents a transfer-learning procedure designed to address this issue.

4.1 Transfer Setting

We consider a setting in which the target domain is subject to administrative censoring, and supplementary observations unaffected by administrative censoring are available from the source domain. We assume that the observations from the source and target domains are independent. We introduce a two-stage estimation procedure by decomposing the parameter vectors in the earlier and later hazards as $\boldsymbol{\theta}_1 = (\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{1,2})$ and $\boldsymbol{\theta}_2 = (\boldsymbol{\theta}_{2,1}, \boldsymbol{\theta}_{2,2})$, respectively. Let $\mathcal{E}_s$ and $\mathcal{E}_t$ be defined as $\mathcal{E}$ evaluated on the observations from the source and target domains, respectively. In the first stage, we jointly estimate $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ by maximizing $\mathcal{E}_s(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, and denote the corresponding first-stage estimator of $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1})$ by $(\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1})$. In the second stage, we estimate the remaining components $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ by maximizing $\mathcal{E}_t((\hat{\boldsymbol{\theta}}_{1,1}, \boldsymbol{\theta}_{1,2}), (\hat{\boldsymbol{\theta}}_{2,1}, \boldsymbol{\theta}_{2,2}))$, and denote the resulting estimator by $(\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2})$. Both stages can be carried out by applying the procedure described in Section 3.3.

We next establish the asymptotic consistency of the estimator obtained by the above two-stage procedure. This consistency relies on the following transfer condition.

Assumption 4. The true values of the transferable components $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1})$, denoted by $(\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})$, are shared by the source and target domains. Moreover, the remaining components $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ are identifiable in the following sense: if $f_1(t_1 \mid \boldsymbol{x}; (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}_{1,2}))f_2(t_2 \mid \boldsymbol{x}; (\boldsymbol{\theta}^*_{2,1}, \boldsymbol{\theta}_{2,2})) = f_1(t_1 \mid \boldsymbol{x}; (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}'_{1,2}))f_2(t_2 \mid \boldsymbol{x}; (\boldsymbol{\theta}^*_{2,1}, \boldsymbol{\theta}'_{2,2}))$ for Lebesgue-almost every $t_1, t_2 \in [0, \tau]$ satisfying $t_1 + t_2 \leq \tau$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) = (\boldsymbol{\theta}'_{1,2}, \boldsymbol{\theta}'_{2,2})$.

Intuitively, Assumption 4 means that Assumption 3 holds for the remaining components $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ once the transferable components $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1})$ are known. In the two-stage procedure, these true values are replaced by their first-stage estimates. The following result formalizes this intuition.

Theorem 4. *Let n_s and n_t denote the sample sizes in the source and target domains, respectively. Then, under Assumptions 1 and 4, the two-stage estimator $((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{1,2}), (\hat{\boldsymbol{\theta}}_{2,1}, \hat{\boldsymbol{\theta}}_{2,2}))$ is asymptotically consistent as $n_s, n_t \to \infty$.*

While the two-stage procedure ensures consistency under administrative censoring, the first-stage estimation error propagates into the second stage and affects the asymptotic covariance of the resulting estimator. The following proposition quantifies this effect.

Proposition 4. *Let $(\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})$ denote the target-domain true values of $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$. Then, under Assumptions 1 and 4, the covariance matrix of $(\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2}) - (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})$ admits the expansion*

$$n_t^{-1} I_1^{-1} + n_s^{-1} I_1^{-1} I_2 I_3^{-1} I_2^\top I_1^{-1} + o(n_s^{-1} + n_t^{-1}), \tag{14}$$

where I_1 is the Fisher information matrix for the remaining components $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ in the target-domain objective $n_t^{-1}\mathcal{E}_t(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, I_2 is the corresponding cross-information matrix between the remaining components $(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ and the transferable components $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1})$ in the target-domain objective, and I_3 is the Schur-complement information matrix for the transferable components $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1})$ in the source-domain objective $n_s^{-1}\mathcal{E}_s(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. All matrices are evaluated at the true parameter values.

We illustrate the above transfer-learning procedure using a parametric proportional hazards model for the earlier event time T_1. To allow a more general specification, we write the earlier hazard as $h_1(t \mid \boldsymbol{x}; \boldsymbol{\theta}_1) = h_b(t \mid \boldsymbol{x}; \boldsymbol{\alpha})\phi(\boldsymbol{x}; \boldsymbol{\beta})$, where the baseline component h_b is allowed to depend on the covariates. As a simple setting in which Assumption 4 can be verified, we assume that the baseline component h_b and the later hazard h_2 are shared by the source and target domains, whereas the time-independent component ϕ in the earlier hazard h_1 is domain-specific. Accordingly, we set $\boldsymbol{\theta}_{1,1} = \boldsymbol{\alpha}$, $\boldsymbol{\theta}_{1,2} = \boldsymbol{\beta}$, $\boldsymbol{\theta}_{2,1} = \boldsymbol{\theta}_2$, and $\boldsymbol{\theta}_{2,2} = \emptyset$. Under this assumption, Assumption 4 is satisfied provided that the domain-specific parameter $\boldsymbol{\beta}$ is identifiable. This setting is suitable when domains differ in disease occurrence levels through ϕ but share the onset-time pattern h_b and the post-onset reporting process h_2.

4.2 Optimization for Parametric Proportional Hazards Models

The procedure described in Section 3.3 can be used for parameter estimation in the second stage. However, the repeated Monte Carlo approximation required by this procedure may induce substantial

fluctuations in the second-stage estimates at each update as the administrative censoring time τ increases and new target-domain data accumulate. Such instability is undesirable for operational decision-making, where practitioners require stable estimates to monitor risk and adapt their actions over time. We next show that, in the transfer-learning setting described at the end of Section 4.1, this issue can be alleviated by constructing a more tractable estimating equation. For clarity, we denote the target-domain coefficient by $\boldsymbol{\gamma}$ instead of $\boldsymbol{\beta}$, while writing the transferred source-domain estimates as $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}_2$. In the following, $(\boldsymbol{x}_i, z_i, y_i, w_i, v_i)_{i=1}^n$ denotes the target-domain observations. We further assume that the baseline hazard function h_b is piecewise continuous in time and bounded on every compact time interval.

We aim to construct a more tractable approximation to the original estimating equation $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) = \mathbf{0}$, where $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) \coloneqq \frac{\partial}{\partial \boldsymbol{\gamma}} \mathcal{E}_t((\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}), \hat{\boldsymbol{\theta}}_2)$. Let $\hat{\boldsymbol{\gamma}}$ denote a solution to this equation. The evaluation of $\boldsymbol{\zeta}_0(\boldsymbol{\gamma})$ involves $S_{\blacksquare}(y \mid \boldsymbol{x})$, whose second term may not admit a closed-form expression. As a simple approximation, we ignore this second term. Then the equation $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) = \mathbf{0}$ becomes

$$\boldsymbol{\zeta}(\boldsymbol{\gamma}) \coloneqq \sum_{i=1}^{n} \Big(v_i \frac{\partial \log \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} - r_i \frac{\partial \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \Big) = \mathbf{0}, \tag{15}$$

where $r_i \coloneqq \int_0^{\tilde{y}_i} h_b(t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) dt$, with $\tilde{y}_i \coloneqq y_i$ for $1 \le i \le m$ and $\tilde{y}_i \coloneqq z_i$ for $m < i \le n$. Although solving Equation (15) often yields stable estimates, it can lead to substantial bias due to the omitted term; see Appendix F.2. To mitigate this bias, we focus on the regime in which only a small proportion of individuals experience the event over $[0, \tau]$. This regime is plausible when the early event hazard is low or the administrative censoring time τ is short. In this regime, we derive a more accurate approximation to the original estimating equation $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) = \mathbf{0}$ by applying a Taylor expansion to the difference $\boldsymbol{\zeta}_0(\hat{\boldsymbol{\gamma}}) - \boldsymbol{\zeta}(\hat{\boldsymbol{\gamma}})$.

Assumption 5. Let $\nu_i \coloneqq \phi(\boldsymbol{x}_i; \hat{\boldsymbol{\gamma}}) \sup_{t \in [0, y_i]} h_b(t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}})$. Assume that $\nu_i y_i$ is bounded above by a common finite constant for all $1 \le i \le m$.

Theorem 5. *Let $\tilde{\boldsymbol{\zeta}}$ be defined analogously to $\boldsymbol{\zeta}$ by replacing r_i with*

$$\tilde{r}_i \coloneqq \int_0^{\tilde{y}_i} h_b(t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\alpha}}) \big(1 - S_2(y_i - t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2)\big)^{1 - v_i} dt. \tag{16}$$

Let δ_i denote the fixed point satisfying

$$\delta_i = \int_{\delta_i}^{y_i} \exp\Big(\int_0^t \big(\nu_i - h_2(u \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2)\big) du \Big) dt. \tag{17}$$

Let $\boldsymbol{\pi}_i$ denote the vector of componentwise absolute values of $\frac{\partial}{\partial \boldsymbol{\gamma}} \log \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}}$. Then, under Assumption 5, we have $\tilde{\boldsymbol{\zeta}}(\hat{\boldsymbol{\gamma}}) = \mathcal{O}(\sum_{i=1}^m \nu_i^2 \delta_i y_i \boldsymbol{\pi}_i)$.

Theorem 5 shows that $\tilde{\boldsymbol{\zeta}}(\boldsymbol{\gamma}) = \mathbf{0}$ can be a reasonable approximation of the original estimating equation $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) = \mathbf{0}$. Note that the proof provides the multiplicative constant in the Landau notation. For $v_i = 0$, the coefficient $\tilde{r}_i$ can be viewed as a weighted version of r_i, where the weight represents the probability that an event occurring at time t is reported within the remaining observation period. This weight arises from the approximation of $\boldsymbol{\zeta}_0(\hat{\boldsymbol{\gamma}}) - \boldsymbol{\zeta}(\hat{\boldsymbol{\gamma}})$. Aside from a one-time precomputation of $\tilde{r}_i$, solving the approximate equation avoids repeated numerical integration and can therefore improve numerical stability compared with the procedure described in Section 3.3.

We can see that $\tilde{\boldsymbol{\zeta}}(\hat{\boldsymbol{\gamma}})$ has a smaller error when the later hazard h_2 is substantially larger than the earlier hazard h_1. This regime can arise, for example, when disease onset is rare but subsequent reporting, such as death after onset, occurs relatively frequently. In this regime, the right-hand side of Equation (17) is strictly decreasing and rapidly decays to zero as δ_i increases. Therefore, the fixed-point solution of Equation (17) is close to zero. Consequently, the equation $\tilde{\boldsymbol{\zeta}}(\boldsymbol{\gamma}) = \mathbf{0}$ provides a reasonable approximation to the original estimating equation $\boldsymbol{\zeta}_0(\boldsymbol{\gamma}) = \mathbf{0}$.

5 Numerical Study

In this section, we present experimental evidence supporting the validity of our approach. The experiment is motivated by timely risk evaluation for individuals newly enrolled in an at-risk monitoring program, where disease-onset times are retrospectively observed only at death.

Survival Model. The earlier hazard is modeled by a piecewise-proportional hazards model. Specifically, the baseline hazard is defined as $h_b(t;\boldsymbol{\alpha}) = \alpha_1\mathbb{I}(t \in [0, 0.5]) + \alpha_2\mathbb{I}(t \in (0.5, 1]) + \alpha_3\mathbb{I}(t \in (1, \infty))$, where $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^3 \in (0, \infty)^3$. The earlier hazards in the source and target domains are modeled as $h_b(t;\boldsymbol{\alpha})\exp(\boldsymbol{\beta}^\top \boldsymbol{x})$ and $h_b(t;\boldsymbol{\alpha})\exp(\boldsymbol{\gamma}^\top \boldsymbol{x})$, respectively, where $\boldsymbol{\beta} = (\beta_i)_{i=1}^d \in \mathbb{R}^d$ and $\boldsymbol{\gamma} = (\gamma_i)_{i=1}^d \in \mathbb{R}^d$. The later hazard is assumed to be constant over time, with $h_2(t) = \lambda \in (0, \infty)$.

Estimation. The parameters for the source domain were estimated using the procedure described in Section 3.3, with 100 iterations and $s = 10$. Increasing these values had little effect on the resulting estimates. The maximization of $\mathcal{L}_1(\boldsymbol{\theta}_1 \mid \mathcal{P}_1)$ was performed using the lifelines library, version 0.27.8, with the penalizer parameter fixed at 10^{-4} for numerical stability. All other hyperparameters in lifelines were set to their default values. The parameters were initialized using estimates obtained by substituting the missing values $(z_i, w_i)_{i=1}^m$ with $(y_i, v_i)_{i=1}^m$. For the target domain, we computed $\boldsymbol{\gamma}$ by solving the approximate estimating equation $\tilde{\boldsymbol{\zeta}}(\boldsymbol{\gamma}) = \mathbf{0}$ via an L-BFGS-based minimization procedure, initialized at $\boldsymbol{\gamma} = \mathbf{0}$. The L-BFGS optimization was implemented using SciPy, version 1.2.1, with default solver options. The computational environment is provided in Appendix G.1. We refer to this transfer-based estimation procedure as *Ours*.

Baselines. We compare our approach with *Standard*, a survival model that does not explicitly model reporting delays and is characterized only by the event hazard h_1. This baseline uses the same transfer procedure as *Ours*, except that the missing values $(z_i, w_i)_{i=1}^m$ are replaced with $(y_i, v_i)_{i=1}^m$ in the relevant domains. For both *Standard* and *Ours*, we also consider variants that learn directly from the target observations, denoted by *Standard Direct* and *Ours Direct*, respectively. To satisfy the identifiability conditions, we impose $\alpha_1 = \alpha_2 = \alpha_3$ when $\tau \leq 0.5$, and $\alpha_2 = \alpha_3$ when $0.5 < \tau \leq 1$. Moreover, we consider a variant of *Ours* that estimates the target-domain parameter $\boldsymbol{\gamma}$ using the procedure described in Section 3.3, with 100 iterations and $s = 10$, instead of the approximate estimating equation. We refer to this variant as *Ours MCEM*.

Toy Data. For each of the source and target domains, we generated 1000 observations with covariate dimension $d = 10$ from the survival model specified by the following parameters. The value of $(\alpha_1, \alpha_2, \alpha_3, \beta_1, \gamma_1)$ was set to $(0.1, 0.2, 0.3, 1, 2)$. For each $2 \leq i \leq d$, β_i and γ_i were independently drawn from the uniform distribution on $[-1, 1]$. To assess the impact of reporting delays, we considered two values of λ, 1 and 5, denoted by *Toy1* and *Toy2*, respectively. The censoring time was generated from a distribution with constant hazard rate 1. For both the source and target domains, each covariate vector was generated uniformly over $[-1, 1]^d$.

Real Data. We adopted two benchmark datasets, `Colon` and `Rotterdam`, from the Rdatasets repository, version 1.0.0. These datasets contain 929 and 2982 individuals, respectively, and provide time-to-event information on disease recurrence and death. To introduce right-censored reporting delays, we treat death occurring after disease recurrence as the reporting event. We rescale z_i and y_i by the median observation time within each dataset. Both datasets contain the `age` variable. For each dataset, the youngest 25% of individuals are assigned to the target domain, whereas the remaining 75% form the source domain. We refer to these settings as *Age1* and *Age2*, respectively. We also consider an alternative split for each dataset. For the `Colon` dataset, we define the target domain using the `surg` variable, which indicates whether the interval from surgery to registration is short or long. Patients with long intervals are assigned to the registration-delayed target cohort, whereas those with shorter intervals form the source domain. For the `Rotterdam` dataset, we define the target domain using the `year` variable, which indicates the year of surgery. After ordering patients by surgery year, the latest 25% are assigned to the temporally later target cohort, whereas the remaining 75% form the source domain. We refer to these settings as *Surg* and *Year*, respectively. Once the split-defining variables are excluded, the `Colon` and `Rotterdam` datasets contain 10 and 9 covariates, respectively. These covariates were used in the survival models, and the preprocessing steps are described in Appendix G.2.

Evaluation. We evaluate the estimation performance under varying administrative censoring times $\tau \in \{0.25, 0.5, 0.75, 1\}$. Because $\tau \leq 1$ in all cases, identifiability of the survival model cannot be ensured from the target-domain observations alone, which motivates the transfer-learning procedure. We conducted 100 trials. In each trial, a randomly selected half of the source-domain observations was used to estimate the parameters in the first stage of the two-stage procedure. The target-domain observations were randomly shuffled and evenly split into two subsets. One subset was administratively censored at each τ and used for parameter estimation in the second stage. The other subset served as a hold-out set to evaluate the estimated disease-onset hazard using the integrated

Table 1: Performance of risk evaluation. The mean and standard deviation of the integrated Brier score over 100 trials are reported; lower values indicate better performance.

Dataset	Method	$\tau = 0.25$	$\tau = 0.5$	$\tau = 0.75$	$\tau = 1$
Toy1	Standard Direct	0.407 (0.051)	0.342 (0.043)	0.334 (0.044)	0.284 (0.036)
	Standard	0.221 (0.030)	0.198 (0.027)	0.187 (0.025)	0.180 (0.023)
	Ours Direct	0.342 (0.087)	0.238 (0.062)	0.176 (0.057)	0.155 (0.034)
	Ours MCEM	0.187 (0.044)	0.149 (0.031)	0.137 (0.025)	0.132 (0.024)
	Ours	0.179 (0.043)	0.147 (0.031)	0.136 (0.024)	0.132 (0.022)
Toy2	Standard Direct	0.264 (0.040)	0.211 (0.031)	0.186 (0.031)	0.154 (0.023)
	Standard	0.143 (0.021)	0.131 (0.019)	0.128 (0.018)	0.127 (0.019)
	Ours Direct	0.199 (0.039)	0.170 (0.028)	0.137 (0.022)	0.132 (0.020)
	Ours MCEM	0.142 (0.025)	0.130 (0.021)	0.127 (0.020)	0.125 (0.021)
	Ours	0.140 (0.024)	0.129 (0.021)	0.126 (0.019)	0.125 (0.020)
Surg	Standard Direct	0.298 (0.037)	0.252 (0.022)	0.246 (0.017)	0.237 (0.016)
	Standard	0.320 (0.035)	0.272 (0.024)	0.245 (0.018)	0.238 (0.016)
	Ours Direct	0.289 (0.035)	0.250 (0.023)	0.240 (0.018)	0.235 (0.016)
	Ours MCEM	0.291 (0.037)	0.250 (0.023)	0.239 (0.018)	0.234 (0.016)
	Ours	0.276 (0.034)	0.246 (0.021)	0.237 (0.017)	0.233 (0.016)
Age1	Standard Direct	0.300 (0.040)	0.253 (0.028)	0.238 (0.022)	0.229 (0.019)
	Standard	0.323 (0.040)	0.269 (0.031)	0.235 (0.021)	0.229 (0.019)
	Ours Direct	0.299 (0.046)	0.260 (0.029)	0.232 (0.024)	0.230 (0.020)
	Ours MCEM	0.292 (0.046)	0.249 (0.032)	0.232 (0.023)	0.228 (0.019)
	Ours	0.280 (0.042)	0.246 (0.029)	0.230 (0.020)	0.227 (0.019)
Year	Standard Direct	0.238 (0.017)	0.200 (0.012)	0.199 (0.011)	0.192 (0.010)
	Standard	0.249 (0.017)	0.213 (0.013)	0.195 (0.011)	0.191 (0.010)
	Ours Direct	0.213 (0.019)	0.180 (0.010)	0.185 (0.009)	0.177 (0.009)
	Ours MCEM	0.209 (0.019)	0.181 (0.010)	0.174 (0.008)	0.172 (0.008)
	Ours	0.204 (0.017)	0.181 (0.009)	0.175 (0.008)	0.173 (0.008)
Age2	Standard Direct	0.370 (0.028)	0.277 (0.018)	0.289 (0.016)	0.268 (0.013)
	Standard	0.404 (0.022)	0.319 (0.017)	0.279 (0.016)	0.264 (0.014)
	Ours Direct	0.334 (0.041)	0.253 (0.019)	0.242 (0.017)	0.240 (0.013)
	Ours MCEM	0.318 (0.042)	0.250 (0.018)	0.233 (0.015)	0.229 (0.013)
	Ours	0.311 (0.039)	0.251 (0.017)	0.233 (0.014)	0.228 (0.013)

Brier score [Graf et al., 1999]. The evaluation horizon was set to the larger of the maximum onset time and the maximum censoring time in the hold-out set. For inverse probability of censoring weighting, we estimated the censoring survival function using the Kaplan–Meier estimator fitted on the hold-out set.

Result. Table 1 shows that *Ours* performs competitively with the baselines and achieves better performance in many settings. These results support the effectiveness of both modeling reporting delays and leveraging transfer learning. For $\tau = 0.25$, *Ours* consistently outperforms *Ours MCEM*, supporting the advantage of using the approximate estimating equation. The parameter estimation results are reported in Appendix G.3.

6 Conclusion

This study addressed survival modeling with right-censored reporting delays and proposed a transfer-learning procedure to mitigate the identifiability challenges posed by administrative censoring. The setting considered in this study is not limited to biostatistical applications. Future work should further examine the validity of the transfer assumptions in application-specific settings, since misspecified shared components may lead to biased risk estimates and inappropriate decision-making.

References

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley-Interscience, second edition, 2002.

Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):1–36, 2019.

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

D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

Norman E. Breslow. Discussion on professor Cox's paper. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):216–217, 1972.

Håvard Kvamme, Ørnulf Borgan, and Ida Scheel. Time-to-event prediction with neural networks and Cox regression. *Journal of Machine Learning Research*, 20(129):1–30, 2019.

Qixian Zhong, Jonas Mueller, and Jane-Ling Wang. Deep extended hazard models for survival analysis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 15111–15124, 2021.

Chirag Nagpal, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller. Deep Cox mixtures for survival regression. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, pages 674–708, 2021.

Yan Li, Lu Wang, Jie Wang, Jieping Ye, and Chandan K. Reddy. Transfer learning for survival analysis via efficient L2,1-norm regularized cox regression. In *Proceedings of the 16th IEEE International Conference on Data Mining*, pages 231–240, 2016.

Alexis Bellot and Mihaela van der Schaar. Boosting transfer learning with survival data from heterogeneous domains. In *Proceedings of the 32nd International Conference on Artificial Intelligence and Statistics*, pages 57–65, 2019.

Ammar Shaker and Carolin Lawrence. Multi-source survival domain adaptation. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pages 9752–9762, 2023.

Ethan Steinberg, Jason A. Fries, Yizhe Xu, and Nigam H. Shah. MOTOR: A time-to-event foundation model for structured medical records. In *International Conference on Learning Representations*, 2024.

Pan Zhao, Julie Josse, and Shu Yang. Efficient and robust transfer learning of optimal individualized treatment regimes with right-censored survival data. *Journal of Machine Learning Research*, 26(48):1–54, 2025.

Ping Hu and Anastasios A. Tsiatis. Estimating the survival distribution when ascertainment of vital status is subject to delay. *Biometrika*, 83(2):371–380, 1996.

Mark J. Van Der Laan and Alan E. Hubbard. Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. *Biometrika*, 85(4):771–783, 1998.

Stephen W. Lagakos, Leila M. Barraj, and V. De Gruttola. Nonparametric analysis of truncated survival data, with application to AIDS. *Biometrika*, 75(3):515–523, 1988.

John D. Kalbfleisch and Jerald F. Lawless. Inference based on retrospective ascertainment: An analysis of the data on transfusion-related AIDS. *Journal of the American Statistical Association*, 84(406):360–372, 1989.

Jeffrey E. Harris. Reporting delays and the incidence of AIDS. *Journal of the American Statistical Association*, 85(412):915–924, 1990.

John D. Kalbfleisch and Jerald F. Lawless. Regression models for right truncated data with application to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32, 1991.

Marcello Pagano, Xin Ming Tu, Victor De Gruttola, and Samantha MaWhinney. Regression analysis of censored and truncated data: Estimating reporting-delay distributions and AIDS incidence from surveillance data. *Biometrics*, 50(4):1203–1214, 1994.

R. L. Kodell and C. J. Nelson. An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics*, 36(2):267–277, 1980.

Bruce W. Turnbull and Toby J. Mitchell. Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments. *Biometrics*, 40(1):41–50, 1984.

Mark J. Van Der Laan, Nicholas P. Jewell, and Derick R. Peterson. Efficient estimation of the lifetime and disease onset distribution. *Biometrika*, 84(3):539–554, 1997.

Kristian Buchardt, Christian Furrer, and Oliver Sandqvist. Estimation for multistate models subject to reporting delays and incomplete event adjudication with application to disability insurance. *Annals of Applied Statistics*, 19(4):2782–2802, 2025.

Theodore R. Holford. The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2): 299–305, 1980.

Patrick Royston and Mahesh Parmar. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15):2175–2197, 2002.

Michael Friedman. Piecewise exponential models for survival data with covariates. *The Annals of Statistics*, 10 (1):101–113, 1982.

Michael Visser. Nonparametric estimation of the bivariate survival function with an application to vertically transmitted AIDS. *Biometrika*, 83(3):507–518, 1996.

Greg C. G. Wei and Martin A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

Patrick Billingsley. *Probability and Measure*. Wiley, third edition, 1995.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18:2529–2545, 1999.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

Michael LeBlanc and John Crowley. Relative risk trees for censored survival data. *Biometrics*, 48(2):411–425, 1992.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in Medicine*, 14(1): 73–82, 1995.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(24), 2018.

Jiachang Liu, Rui Zhang, and Cynthia Rudin. Fastsurvival: Hidden computational blessings in training Cox proportional hazards models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 87712–87765, 2024.

Michael Höhle and Matthias an der Heiden. Bayesian nowcasting during the STEC O104:H4 outbreak in Germany, 2011. *Biometrics*, 70(4):993–1002, 2014.

Oliver Stoner and Theo Economou. Multivariate hierarchical frameworks for modeling delayed reporting in count data. *Biometrics*, 76(3):789–798, 2020.

Simon N. Wood. Inferring UK COVID-19 fatal infection trajectories from daily mortality data: Were infections already in decline before the UK lockdowns? *Biometrics*, 78(3):1127–1140, 2022.

Oliver Stoner, Alba Halliday, and Theo Economou. Correcting delayed reporting of COVID-19 using thegeneralized-dirichlet-multinomial method. *Biometrics*, 79(3):2537–2550, 2023.

C. P. Farrington, N. J. Andrews, A. D. Beale, and M. A. Catchpole. A statistical algorithm for the early detection of outbreaks of infectious disease. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159 (3):547–563, 1996.

Angela Noufaily, Doyo G. Enki, Paddy Farrington, Paul Garthwaite, Nick Andrews, and André Charlett. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, 32(7): 1206–1222, 2013.

Pooria Joulani, Andras Gyorgy, and Csaba Szepesvari. Online learning under delayed feedback. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1453–1461, 2013.

Kent Quanrud and Daniel Khashabi. Online learning with adversarial delays. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*, pages 1270–1278, 2015.

Timothy A. Mann, Sven Gowal, Andras Gyorgy, Huiyi Hu, Ray Jiang, Balaji Lakshminarayanan, and Prav Srinivasan. Proxies with applications to recommender systems. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4324–4332, 2019.

Ilai Bistritz, Zhengyuan Zhou, Xi Chen, Nicholas Bambos, and Jose Blanchet. Online EXP3 learning in adversarial bandits with delayed feedback. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 11349–11358, 2019.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brueckne. Linear bandits with stochastic delayed feedback. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9712–9721, 2020.

Julian Zimmert, Saeed Masoudian, and Yevgeny Seldin. A best-of-both-worlds algorithm for bandits with delayed feedback with robustness to excessive delays. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 141071–141102, 2024.

A Notations

Notation	Description
$\mathcal{D}$	$\mathcal{D} \equiv (\boldsymbol{x}_i, y_i, v_i)_{i=1}^n$
$\mathcal{D}_1$	$\mathcal{D}_1 \equiv (\boldsymbol{x}_i, z_i, w_i)_{i=1}^n$
$\mathcal{D}_2$	$\mathcal{D}_2 \equiv (\boldsymbol{x}_i, y_i - z_i, v_i)_{i=1}^n$
$\mathcal{P}_1$	the combined pseudo-complete dataset corresponding to the earlier process
$\mathcal{P}_2$	the combined pseudo-complete dataset corresponding to the later process
n	the number of observations
m	the number of observations with $v_i = 0$
s	the number of pseudo-complete datasets
Y	$\min\{T, C\}$
$\bar{Y}$	$\min\{Y, \tau\}$, where τ denotes the administrative censoring time
V	$\mathbb{I}(T \leq C)$
$\bar{V}$	$\mathbb{I}(T \leq C, T \leq \tau)$
Z	$\min\{T_1, C\}$
$\bar{Z}$	$\min\{Z, \tau\}$
W	$\mathbb{I}(T_1 \leq C)$
$\bar{W}$	$\mathbb{I}(T_1 \leq C, T_1 \leq \tau)$
$\boldsymbol{x}_i$	$\boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ represents the individual covariates
y_i	the time from the start of the earlier process to the later event or censoring
v_i	the later event indicator
z_i	the time until the earlier event or censoring
w_i	the earlier event indicator
$\mathcal{L}(\boldsymbol{\theta})$	$\log \prod_{i=1}^n f(y_i \mid \boldsymbol{x}_i)^{v_i} S(y_i \mid \boldsymbol{x}_i)^{1-v_i}$
$\mathcal{L}_c(\boldsymbol{\theta}_c)$	$\log \prod_{i=1}^n S_c(y_i \mid \boldsymbol{x}_i)^{v_i} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i}$
$\mathcal{L}_c^\dagger(\boldsymbol{\theta}_c)$	$\log \prod_{i=1}^n S_c(y_i \mid \boldsymbol{x}_i)^{v_i^\dagger} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i^\dagger}$ with $v_i^\dagger \equiv \mathbb{I}(y_i \geq \tau(1 - v_i))$
$\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$	$\log \prod_{i=1}^n S_\blacksquare(y_i \mid \boldsymbol{x}_i)^{1-v_i} f_\square(z_i, y_i \mid \boldsymbol{x}_i)^{v_i}$
$f(t \mid \boldsymbol{x})$	the conditional density function of T at t given $X = \boldsymbol{x}$
$f_c(t \mid \boldsymbol{x})$	the conditional density function of C at t given $X = \boldsymbol{x}$
$f_1(t \mid \boldsymbol{x})$	the conditional density function of T_1 at t given $X = \boldsymbol{x}$
$f_2(t \mid \boldsymbol{x})$	the conditional density function of T_2 at t given $X = \boldsymbol{x}$
$f_\square(z, y \mid \boldsymbol{x})$	$f_1(z \mid \boldsymbol{x}) f_2(y - z \mid \boldsymbol{x})$
$S(t \mid \boldsymbol{x})$	$\int_t^\infty f(s \mid \boldsymbol{x}) ds$
$S_c(t \mid \boldsymbol{x})$	$\int_t^\infty f_c(s \mid \boldsymbol{x}) ds$
$S_1(t \mid \boldsymbol{x})$	$\int_t^\infty f_1(s \mid \boldsymbol{x}) ds$
$S_2(t \mid \boldsymbol{x})$	$\int_t^\infty f_2(s \mid \boldsymbol{x}) ds$
$S_\blacksquare(y \mid \boldsymbol{x})$	$S_1(y \mid \boldsymbol{x}) + \int_0^y f_1(t \mid \boldsymbol{x}) S_2(y - t \mid \boldsymbol{x}) dt$
$\boldsymbol{\theta}$	the real-valued vector that parameterizes f
$\boldsymbol{\theta}_c$	the real-valued vector that parameterizes f_c
$\boldsymbol{\theta}_1$	the real-valued vector that parameterizes f_1
$\boldsymbol{\theta}_2$	the real-valued vector that parameterizes f_2

B Related Work

B.1 Proportional Hazards Model

The Cox proportional hazards model [Cox, 1972] is widely used to represent time dependence and covariate effect, assuming a constant hazard ratio over time. The regression coefficients capturing covariate effect are estimated via partial likelihood [Cox, 1975], after which the baseline hazard characterizing time dependence is estimated nonparametrically [Breslow, 1972]. Alternatively, the baseline hazard can be modeled parametrically using piecewise constant functions [Holford, 1980]. A large body of work has focused on enhancing the flexibility of covariate effect, replacing the linear predictor with nonlinear structures such as additive splines [Hastie and Tibshirani, 1986], decision trees [LeBlanc and Crowley, 1992], and neural networks [Faraggi and Simon, 1995, Katzman et al., 2018]. More recently, extensions beyond the proportional hazards assumption have been studied in the machine learning community [Kvamme et al., 2019, Zhong et al., 2021, Nagpal et al.,

2021]. Additionally, efficient and stable optimization for Cox regression has become increasingly important in large-scale and high-dimensional settings [Liu et al., 2024].

B.2 Delay Structures in Related Fields

Delay structures arising from operational constraints have been studied in fields such as nowcasting [Höhle and an der Heiden, 2014, Stoner and Economou, 2020, Wood, 2022, Stoner et al., 2023], outbreak detection [Farrington et al., 1996, Noufaily et al., 2013], and online learning [Joulani et al., 2013, Quanrud and Khashabi, 2015, Mann et al., 2019, Bistritz et al., 2019, Vernade et al., 2020, Zimmert et al., 2024]. Nowcasting and outbreak detection typically correct reporting lags in aggregated case counts for real-time estimation or surveillance. In online learning, delayed feedback similarly complicates sequential decision-making, since the learner must update its model or policy before outcomes from earlier actions are revealed.

C Preliminaries

C.1 Standard Survival Model without Administrative Censoring

For $y \in [0, \infty)$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Y, V) is given by

$$\begin{aligned}\frac{\partial}{\partial y}\Pr(Y \le y, V = 0 \mid X = \boldsymbol{x}) &= \frac{\partial}{\partial y}\Pr(C \le y, C < T \mid X = \boldsymbol{x}) \\ &= \frac{\partial}{\partial y}\int_0^y dc \int_c^\infty dt f(t \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\ &= \int_y^\infty f(t \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) dt \\ &= S(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}). \end{aligned} \tag{18}$$

For $y \in [0, \infty)$, $v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Y, V) is given by

$$\begin{aligned}\frac{\partial}{\partial y}\Pr(Y \le y, V = 1 \mid X = \boldsymbol{x}) &= \frac{\partial}{\partial y}\Pr(T \le y, T \le C \mid X = \boldsymbol{x}) \\ &= \frac{\partial}{\partial y}\int_0^y dt \int_t^\infty dc f(t \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\ &= \int_y^\infty f(y \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) dc \\ &= f(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}). \end{aligned} \tag{19}$$

Consequently, the conditional joint distribution of (Y, V) is given by $g(y, v \mid \boldsymbol{x})$. Then the log likelihood without administrative censoring is given by

$$\begin{aligned}\log \prod_{i=1}^n g(y_i, v_i \mid \boldsymbol{x}_i) &= \log \prod_{i=1}^n \big(S(y_i \mid \boldsymbol{x}_i) f_c(y_i \mid \boldsymbol{x}_i)\big)^{1-v_i} \big(f(y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{v_i} \\ &= \log \prod_{i=1}^n S(y_i \mid \boldsymbol{x}_i)^{1-v_i} f(y_i \mid \boldsymbol{x}_i)^{v_i} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i} S_c(y_i \mid \boldsymbol{x}_i)^{v_i} \\ &= \mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c(\boldsymbol{\theta}_c). \end{aligned} \tag{20}$$

From the definition of g, for any $y \in [0, \infty)$ and $\boldsymbol{x} \in \mathcal{X}$, we have

$$S(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}) = 1 - \int_0^y \big(f(t \mid \boldsymbol{x}) S_c(t \mid \boldsymbol{x}) + S(t \mid \boldsymbol{x}) f_c(t \mid \boldsymbol{x})\big) dt = 1 - \sum_{v \in \{0,1\}} \int_0^y g(t, v \mid \boldsymbol{x}) dt, \tag{21}$$

$$S(y \mid \boldsymbol{x}) = \exp\big(\log S(y \mid \boldsymbol{x})\big) = \exp\Big(-\int_0^y \frac{f(t \mid \boldsymbol{x})}{S(t \mid \boldsymbol{x})} dt\Big) = \exp\Big(-\int_0^y \frac{g(t, 1 \mid \boldsymbol{x})}{S(t \mid \boldsymbol{x}) S_c(t \mid \boldsymbol{x})} dt\Big), \tag{22}$$

$$S_c(y \mid \boldsymbol{x}) = \exp\big(\log S_c(y \mid \boldsymbol{x})\big) = \exp\Big(-\int_0^y \frac{f_c(t \mid \boldsymbol{x})}{S_c(t \mid \boldsymbol{x})} dt\Big) = \exp\Big(-\int_0^y \frac{g(t, 0 \mid \boldsymbol{x})}{S(t \mid \boldsymbol{x}) S_c(t \mid \boldsymbol{x})} dt\Big). \tag{23}$$

Therefore, if (f, f_c) is identifiable with respect to $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$, then g is also identifiable with respect to $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$. By the uniform law of large numbers and Jensen's inequality, the following holds:

$$\begin{aligned}
\frac{1}{n}\big(\mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c(\boldsymbol{\theta}_c)\big) &= \frac{1}{n}\sum_{i=1}^{n} \log g(y_i, v_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}, \boldsymbol{\theta}_c) \\
&\xrightarrow{p} \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\theta}_c) \\
&= \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log \frac{g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}, \boldsymbol{\theta}_c)}{g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*)} \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \\
&\leq \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*), \qquad (24)
\end{aligned}$$

where $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}_c^*$ are the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_c$, respectively. Consider the equality condition in Jensen's inequality. If (f, f_c) is identifiable with respect to $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$, then the estimator of $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$ obtained by maximizing $\mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c(\boldsymbol{\theta}_c)$ is asymptotically consistent. Similarly, the following holds:

$$\begin{aligned}
\frac{1}{n}\mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})^{v_i} S(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta})^{1-v_i} \\
&\xrightarrow{p} \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log f(y \mid \boldsymbol{x}; \boldsymbol{\theta})^{v} S(y \mid \boldsymbol{x}; \boldsymbol{\theta})^{1-v} \\
&= \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log \frac{f(y \mid \boldsymbol{x}; \boldsymbol{\theta})^{v} S(y \mid \boldsymbol{x}; \boldsymbol{\theta})^{1-v}}{f(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{v} S(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{1-v}} \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log f(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{v} S(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{1-v} \\
&\leq \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \sum_{v\in\{0,1\}} g(y, v \mid \boldsymbol{x}; \boldsymbol{\theta}^*, \boldsymbol{\theta}_c^*) \log f(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{v} S(y \mid \boldsymbol{x}; \boldsymbol{\theta}^*)^{1-v}. \qquad (25)
\end{aligned}$$

Therefore, if f is identifiable with respect to $\boldsymbol{\theta}$, then the estimator of $\boldsymbol{\theta}$ obtained by maximizing $\mathcal{L}(\boldsymbol{\theta})$ is asymptotically consistent.

C.2 Standard Survival Model with Administrative Censoring

For $y = \tau$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$\begin{aligned}
\Pr(\bar{Y} = \tau, \bar{V} = 0 \mid X = \boldsymbol{x}) &= \Pr(\tau \leq C, \tau < T \mid X = \boldsymbol{x}) \\
&= \int_\tau^\infty dc \int_\tau^\infty dt f(t \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\
&= S(\tau \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}). \qquad (26)
\end{aligned}$$

For $y \in [0, \tau)$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$\frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 0 \mid X = \boldsymbol{x}) = \frac{\partial}{\partial y} \Pr(Y \leq y, V = v \mid X = \boldsymbol{x}) = S(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}). \qquad (27)$$

For $y \in [0, \tau)$, $v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$\frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 1 \mid X = \boldsymbol{x}) = \frac{\partial}{\partial y} \Pr(Y \leq y, V = v \mid X = \boldsymbol{x}) = f(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}). \qquad (28)$$

Consequently, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by $\bar{g}(y, v \mid \boldsymbol{x})$. Then the log likelihood with administrative censoring is given by

$$
\begin{aligned}
&\log \prod_{i=1}^{n} \bar{g}(y_i, v_i \mid \boldsymbol{x}_i) \\
&= \log \prod_{i=1}^{n} \big(S(y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{(1-v_i)\mathbb{I}(y_i=\tau)} \big(S(y_i \mid \boldsymbol{x}_i) f_c(y_i \mid \boldsymbol{x}_i)\big)^{(1-v_i)\mathbb{I}(y_i<\tau)} \big(f(y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{v_i} \\
&= \log \prod_{i=1}^{n} S(y_i \mid \boldsymbol{x}_i)^{1-v_i} f(y_i \mid \boldsymbol{x}_i)^{v_i} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i^\dagger} S_c(y_i \mid \boldsymbol{x}_i)^{v_i^\dagger} \\
&= \mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c^\dagger(\boldsymbol{\theta}_c). \qquad (29)
\end{aligned}
$$

Consider the setting obtained by replacing the time interval $[0, \infty)$ with $[0, \tau]$ and the observed-data distribution characterized by g with that characterized by $\bar{g}$. If (f, f_c) over the time interval $[0, \tau]$ is identifiable with respect to $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$, then the estimator of $(\boldsymbol{\theta}, \boldsymbol{\theta}_c)$ obtained by maximizing $\mathcal{L}(\boldsymbol{\theta}) + \mathcal{L}_c^\dagger(\boldsymbol{\theta}_c)$ is asymptotically consistent. If f over the time interval $[0, \tau]$ is identifiable with respect to $\boldsymbol{\theta}$, the estimator of $\boldsymbol{\theta}$ obtained by maximizing $\mathcal{L}(\boldsymbol{\theta})$ is asymptotically consistent.

D Proofs

D.1 Proof of Lemma 1

For $0 \leq y$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Y, V) is given by

$$
\begin{aligned}
&\frac{\partial}{\partial y} \Pr(Y \leq y, V = 0 \mid X = \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \Pr(Y \leq y, W = 0, V = 0 \mid X = \boldsymbol{x}) + \frac{\partial}{\partial y} \Pr(Y \leq y, W = 1, V = 0 \mid X = \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \Pr(C \leq y, C < T_1 \mid X = \boldsymbol{x}) + \frac{\partial}{\partial y} \Pr(C \leq y, T_1 \leq C, C < T_1 + T_2 \mid X = \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_0^y dc \int_c^\infty dt_1 f_1(t_1 \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) + \frac{\partial}{\partial y} \int_0^y dc \int_0^c dt_1 \int_{c-t_1}^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\
&= \int_y^\infty dt_1 f_1(t_1 \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) + \int_0^y dt_1 \int_{y-t_1}^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) \\
&= S_1(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) + \int_0^y f_1(t_1 \mid \boldsymbol{x}) S_2(y - t_1 \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}) dt_1 \\
&= S_\blacksquare(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}). \qquad (30)
\end{aligned}
$$

For $0 \leq z \leq y$, $w = v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of (Z, Y, W, V) is given by

$$
\begin{aligned}
&\frac{\partial^2}{\partial y \partial z} \Pr(Z \leq z, Y \leq y, W = 1, V = 1 \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \Pr(T_1 \leq z, T_1 + T_2 \leq y, T_1 \leq C, T_1 + T_2 \leq C \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dt_1 \int_0^{y-t_1} dt_2 \int_{t_1+t_2}^\infty dc f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dt_1 \int_0^{y-t_1} dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) S_c(t_1 + t_2 \mid \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_0^{y-z} f_1(z \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) S_c(z + t_2 \mid \boldsymbol{x}) dt_2 \\
&= f_1(z \mid \boldsymbol{x}) f_2(y - z \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}) \\
&= f_\square(z, y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}). \qquad (31)
\end{aligned}
$$

D.2 Proof of Proposition 1

Applying Lemma 1, the log likelihood without administrative censoring is given by

$$\begin{aligned}\log \prod_{i=1}^{n} g_{\blacksquare}(y_i \mid \boldsymbol{x}_i)^{1-v_i} g_{\square}(z_i, y_i \mid \boldsymbol{x}_i)^{v_i} &= \log \prod_{i=1}^{n} \big(S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) f_c(y_i \mid \boldsymbol{x}_i)\big)^{1-v_i} \big(f_{\square}(z_i, y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{v_i} \\ &= \log \prod_{i=1}^{n} S_{\blacksquare}(y_i \mid \boldsymbol{x}_i)^{1-v_i} f_{\square}(z_i, y_i \mid \boldsymbol{x}_i)^{v_i} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i} S_c(y_i \mid \boldsymbol{x}_i)^{v_i} \\ &= \mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c(\boldsymbol{\theta}_c). \end{aligned} \tag{32}$$

D.3 Proof of Theorem 1

For any $t_1, t_2 \in [0, \infty)$ and $\boldsymbol{x} \in \mathcal{X}$, the following holds:

$$f_1(t_1 \mid \boldsymbol{x}) = \int_0^\infty f_{\square}(t_1, t_1 + t \mid \boldsymbol{x}) dt, \qquad f_2(t_2 \mid \boldsymbol{x}) = \int_0^\infty f_{\square}(t, t + t_2 \mid \boldsymbol{x}) dt. \tag{33}$$

Therefore, under Assumption 1, if $f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ for Lebesgue-almost every $0 \le z \le y$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$. From Lemma 1, the uniform law of large numbers, and Jensen's inequality, we have

$$\begin{aligned}\frac{1}{n}\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \frac{1}{n} \sum_{i=1}^{n} \big((1 - v_i) \log S_{\blacksquare}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + v_i \log f_{\square}(z_i, y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\big) \\ &\xrightarrow{p} \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log S_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \int_0^y dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log \frac{S_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{S_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)} \\ &\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \int_0^y dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log \frac{f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*)} \\ &\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log S_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) \\ &\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \int_0^y dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) \\ &\le \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log S_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*) \\ &\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^\infty dy \int_0^y dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*). \end{aligned} \tag{34}$$

Consequently, from the equality condition in Jensen's inequality, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is asymptotically consistent.

D.4 Proof of Proposition 2

By the definitions of $g_{\blacksquare}$ and $g_{\square}$, for $0 \le z \le y$ and $\boldsymbol{x} \in \mathcal{X}$, we have

$$\begin{aligned}S_{\blacksquare}(y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}) &= 1 - \int_0^y \Big(S_c(t \mid \boldsymbol{x}) \int_0^t f_1(u \mid \boldsymbol{x}) f_2(t - u \mid \boldsymbol{x}) du + S_{\blacksquare}(t \mid \boldsymbol{x}) f_c(t \mid \boldsymbol{x})\Big) dt \\ &= 1 - \int_0^y \Big(\int_0^t g_{\square}(u, t \mid \boldsymbol{x}) du + g_{\blacksquare}(t \mid \boldsymbol{x})\Big) dt, \end{aligned} \tag{35}$$

$$\begin{aligned}S_c(y \mid \boldsymbol{x}) &= \exp\big(\log S_c(y \mid \boldsymbol{x})\big) \\ &= \exp\Big(-\int_0^y \frac{f_c(t \mid \boldsymbol{x})}{S_c(t \mid \boldsymbol{x})} dt\Big) \\ &= \exp\Big(-\int_0^y \frac{g_{\blacksquare}(t \mid \boldsymbol{x})}{S_{\blacksquare}(t \mid \boldsymbol{x}) S_c(t \mid \boldsymbol{x})} dt\Big), \end{aligned} \tag{36}$$

$$f_{\square}(z, y \mid \boldsymbol{x}) = \frac{g_{\square}(z, y \mid \boldsymbol{x})}{S_c(y \mid \boldsymbol{x})}. \tag{37}$$

Therefore, under Assumption 2, if $g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) = g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_c')$ and $g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) = g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_c')$ for Lebesgue-almost every $0 \leq z \leq y$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \boldsymbol{\theta}_c')$. By Lemma 1, the uniform law of large numbers, and Jensen's inequality, we have

$$
\begin{aligned}
\frac{1}{n}\big(\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c(\boldsymbol{\theta}_c)\big) &= \frac{1}{n}\sum_{i=1}^{n}\big((1 - v_i)\log g_{\blacksquare}(y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) + v_i \log g_{\square}(z_i, y_i \mid \boldsymbol{x}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)\big) \\
&\xrightarrow{p} \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy \int_0^{y} dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c) \\
&= \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log \frac{g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)}{g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*)} \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy \int_0^{y} dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log \frac{g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)}{g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*)} \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy \int_0^{y} dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \\
&\leq \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\blacksquare}(y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \\
&\quad + \int_{\mathcal{X}} dP_X(\boldsymbol{x}) \int_0^{\infty} dy \int_0^{y} dz g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*) \log g_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1^*, \boldsymbol{\theta}_2^*, \boldsymbol{\theta}_c^*).
\end{aligned} \tag{38}
$$

Consequently, applying the equality condition in Jensen's inequality, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_c)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c(\boldsymbol{\theta}_c)$ is asymptotically consistent.

D.5 Proof of Lemma 2

For $y = \tau$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$
\begin{aligned}
&\Pr(\bar{Y} = \tau, \bar{V} = 0 \mid X = \boldsymbol{x}) \\
&= \Pr(\bar{Y} = \tau, \bar{W} = 0, \bar{V} = 0 \mid X = \boldsymbol{x}) + \Pr(\bar{Y} = \tau, \bar{W} = 1, \bar{V} = 0 \mid X = \boldsymbol{x}) \\
&= \Pr(\tau \leq C, \tau < T_1 \mid X = \boldsymbol{x}) + \Pr(\tau \leq C, T_1 \leq \tau, \tau < T_1 + T_2 \mid X = \boldsymbol{x}) \\
&= \int_{\tau}^{\infty} dc \int_{\tau}^{\infty} dt_1 f_1(t_1 \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) + \int_{\tau}^{\infty} dc \int_0^{\tau} dt_1 \int_{\tau - t_1}^{\infty} dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_c(c \mid \boldsymbol{x}) \\
&= S_1(\tau \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}) + \int_0^{\tau} f_1(t_1 \mid \boldsymbol{x}) S_2(\tau - t_1 \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}) dt_1 \\
&= S_{\blacksquare}(\tau \mid \boldsymbol{x}) S_c(\tau \mid \boldsymbol{x}).
\end{aligned} \tag{39}
$$

For $0 \leq y < \tau$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Y}, \bar{V})$ is given by

$$
\begin{aligned}
\frac{\partial}{\partial y} \Pr(\bar{Y} \leq y, \bar{V} = 0 \mid X = \boldsymbol{x}) &= \frac{\partial}{\partial y} \Pr(Y \leq y, V = 0 \mid X = \boldsymbol{x}) \\
&= S_{\blacksquare}(y \mid \boldsymbol{x}) f_c(y \mid \boldsymbol{x}).
\end{aligned} \tag{40}
$$

For $0 \leq z \leq y \leq \tau$, $w = v = 1$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\bar{Z}, \bar{Y}, \bar{W}, \bar{V})$ is given by

$$
\begin{aligned}
\frac{\partial^2}{\partial y \partial z} \Pr(\bar{Z} \leq z, \bar{Y} \leq y, \bar{W} = 1, \bar{V} = 1 \mid X = \boldsymbol{x}) &= \frac{\partial^2}{\partial y \partial z} \Pr(Z \leq z, Y \leq y, V = 1 \mid X = \boldsymbol{x}) \\
&= f_{\square}(z, y \mid \boldsymbol{x}) S_c(y \mid \boldsymbol{x}).
\end{aligned} \tag{41}
$$

D.6 Proof of Proposition 3

Applying Lemma 2, the log likelihood with administrative censoring is given by

$$
\begin{aligned}
&\log \prod_{i=1}^{n} \bar{g}_{\blacksquare}(y_i \mid \boldsymbol{x}_i)^{1-v_i} g_{\square}(z_i, y_i \mid \boldsymbol{x}_i)^{v_i} \\
&= \log \prod_{i=1}^{n} \big(S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{(1-v_i)\mathbb{I}(y_i=\tau)} \big(S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) f_c(y_i \mid \boldsymbol{x}_i)\big)^{(1-v_i)\mathbb{I}(y_i<\tau)} \big(f_{\square}(z_i, y_i \mid \boldsymbol{x}_i) S_c(y_i \mid \boldsymbol{x}_i)\big)^{v_i} \\
&= \log \prod_{i=1}^{n} S_{\blacksquare}(y_i \mid \boldsymbol{x}_i)^{1-v_i} f_{\square}(z_i, y_i \mid \boldsymbol{x}_i)^{v_i} f_c(y_i \mid \boldsymbol{x}_i)^{1-v_i^{\dagger}} S_c(y_i \mid \boldsymbol{x}_i)^{v_i^{\dagger}} \\
&= \mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \mathcal{L}_c^{\dagger}(\boldsymbol{\theta}_c).
\end{aligned} \tag{42}
$$

D.7 Proof of Theorem 2

Under Assumption 3, if $f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = f_{\square}(z, y \mid \boldsymbol{x}; \boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$ for Lebesgue-almost every $0 \le z \le y \le \tau$ and P_X-almost every $\boldsymbol{x} \in \mathcal{X}$, then $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$. The proof proceeds as in Theorem 1, with the time interval $[0, \infty)$ replaced by $[0, \tau]$ and $(g_{\blacksquare}, g_{\square})$ replaced by $(\bar{g}_{\blacksquare}, \bar{g}_{\square})$, accounting for the probability mass at $y = \tau$. Then the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by maximizing $\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is asymptotically consistent.

D.8 Proof of Theorem 3

The following holds:

$$
\begin{aligned}
&\log \mathcal{I}\big[\exp\big(\mathcal{L}(\boldsymbol{\theta}_1 \mid \mathcal{D}_1) + \mathcal{L}(\boldsymbol{\theta}_2 \mid \mathcal{D}_2)\big)\big] \\
&= \log \mathcal{I}\Big[\prod_{i=1}^{n} f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} f_2(y_i - z_i \mid \boldsymbol{x}_i)^{v_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)^{1-v_i}\Big] \\
&= \log \mathcal{I}\Big[\prod_{i=1}^{m} f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)\Big] \prod_{i=m+1}^{n} f_1(z_i \mid \boldsymbol{x}_i) f_2(y_i - z_i \mid \boldsymbol{x}_i) \\
&= \log \prod_{i=1}^{m} \Big(S_1(y_i \mid \boldsymbol{x}_i) + \int_0^{y_i} f_1(z_i \mid \boldsymbol{x}_i) S_2(y_i - z_i \mid \boldsymbol{x}_i) dz_i\Big) \prod_{i=m+1}^{n} f_1(z_i \mid \boldsymbol{x}_i) f_2(y_i - z_i \mid \boldsymbol{x}_i) \\
&= \log \prod_{i=1}^{m} S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) \prod_{i=m+1}^{n} f_{\square}(z_i, y_i \mid \boldsymbol{x}_i) \\
&= \mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2).
\end{aligned} \tag{43}
$$

Therefore, applying Jensen's inequality, we have

$$
\begin{aligned}
\mathcal{E}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) &= \log \mathcal{I}\Big[\frac{\exp\big(\mathcal{L}(\boldsymbol{\theta}_1 \mid \mathcal{D}_1) + \mathcal{L}(\boldsymbol{\theta}_2 \mid \mathcal{D}_2)\big)}{q} q\Big] \\
&\ge \mathcal{I}\big[\big(\mathcal{L}(\boldsymbol{\theta}_1 \mid \mathcal{D}_1) + \mathcal{L}(\boldsymbol{\theta}_2 \mid \mathcal{D}_2) - \log q\big) q\big] \\
&= \mathcal{I}\Big[q \log \frac{\prod_{i=1}^{m} f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)}{q}\Big] + \log \prod_{i=m+1}^{n} f_{\square}(z_i, y_i \mid \boldsymbol{x}_i).
\end{aligned} \tag{44}
$$

Additionally, for each $1 \le i \le m$, the following holds:

$$
\begin{aligned}
&\mathcal{I}_i\big[f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)\big] \\
&= \big[f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)\big]\big|_{z_i=y_i, w_i=0} \\
&\quad + \int_0^{y_i} \big[f_1(z_i \mid \boldsymbol{x}_i)^{w_i} S_1(z_i \mid \boldsymbol{x}_i)^{1-w_i} S_2(y_i - z_i \mid \boldsymbol{x}_i)\big]\big|_{w_i=1} dz_i \\
&= S_1(y_i \mid \boldsymbol{x}_i) + \int_0^{y_i} f_1(z_i \mid \boldsymbol{x}_i) S_2(y_i - z_i \mid \boldsymbol{x}_i) dz_i \\
&= S_{\blacksquare}(y_i \mid \boldsymbol{x}_i).
\end{aligned} \tag{45}
$$

Consequently, according to the equality condition in Jensen's inequality, the lower bound defined in Equation (11) is maximized when Equation (12) holds for each $1 \le i \le m$.

D.9 Proof of Theorem 4

By Theorem 1, under Assumption 1, we have

$$(\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}) \xrightarrow{p} (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1}) \qquad \text{as } n_s \to \infty. \tag{46}$$

Under the assumed regularity conditions, this implies

$$\sup_{\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}} \left| n_t^{-1} \mathcal{E}_t((\hat{\boldsymbol{\theta}}_{1,1}, \boldsymbol{\theta}_{1,2}), (\hat{\boldsymbol{\theta}}_{2,1}, \boldsymbol{\theta}_{2,2})) - n_t^{-1} \mathcal{E}_t((\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}_{1,2}), (\boldsymbol{\theta}^*_{2,1}, \boldsymbol{\theta}_{2,2})) \right| \xrightarrow{p} 0 \qquad \text{as } n_s, n_t \to \infty. \tag{47}$$

Moreover, the uniform law of large numbers yields

$$\sup_{\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}} \left| n_t^{-1} \mathcal{E}_t((\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}_{1,2}), (\boldsymbol{\theta}^*_{2,1}, \boldsymbol{\theta}_{2,2})) - M_t(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) \right| \xrightarrow{p} 0 \qquad \text{as } n_t \to \infty, \tag{48}$$

where $M_t(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})$ denotes the population target-domain objective with $(\boldsymbol{\theta}_{1,1}, \boldsymbol{\theta}_{2,1}) = (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})$. As in the proof of Theorem 2, Jensen's inequality shows that M_t is maximized at $(\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})$. By the equality condition in Jensen's inequality, under Assumption 4, this maximizer is unique. Consequently, the uniform convergence of the second-stage objective and the argmax consistency theorem imply

$$(\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2}) \xrightarrow{p} (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2}) \qquad \text{as } n_s, n_t \to \infty. \tag{49}$$

This establishes the consistency of the two-stage estimator.

D.10 Proof of Proposition 4

By the standard asymptotic expansion of the first-stage estimator based on the source-domain objective, we have

$$\operatorname{Cov}\big((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}) - (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})\big) = n_s^{-1} I_3^{-1} + o(n_s^{-1}). \tag{50}$$

Next, the second-stage estimator satisfies

$$\begin{aligned} 0 &= n_t^{-1} \nabla_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})} \mathcal{E}_t\big((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}), (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})\big)|_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) = (\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2})} \\ &= n_t^{-1} \nabla_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})} \mathcal{E}_t\big((\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1}), (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})\big)|_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) = (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})} \\ &\quad - I_1\big((\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2}) - (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})\big) - I_2\big((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}) - (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})\big) + o_p(n_t^{-1/2} + n_s^{-1/2}). \end{aligned} \tag{51}$$

Therefore, we have

$$\begin{aligned} \big((\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2}) - (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})\big) &= n_t^{-1} I_1^{-1} \nabla_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})} \mathcal{E}_t\big((\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1}), (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})\big)|_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) = (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})} \\ &\quad - I_1^{-1} I_2\big((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}) - (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})\big) + o_p(n_t^{-1/2} + n_s^{-1/2}). \end{aligned} \tag{52}$$

Consequently, we obtain

$$\begin{aligned} \operatorname{Cov}\big((\hat{\boldsymbol{\theta}}_{1,2}, \hat{\boldsymbol{\theta}}_{2,2}) - (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})\big) &= n_t^{-2} I_1^{-1} \operatorname{Cov}(\nabla_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})} \mathcal{E}_t\big((\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1}), (\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2})\big)|_{(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{2,2}) = (\boldsymbol{\theta}^*_{1,2}, \boldsymbol{\theta}^*_{2,2})}) I_1^{-1} \\ &\quad + I_1^{-1} I_2 \operatorname{Cov}\big((\hat{\boldsymbol{\theta}}_{1,1}, \hat{\boldsymbol{\theta}}_{2,1}) - (\boldsymbol{\theta}^*_{1,1}, \boldsymbol{\theta}^*_{2,1})\big) I_2^\top I_1^{-1} + o(n_t^{-1} + n_s^{-1}) \\ &= n_t^{-1} I_1^{-1} + n_s^{-1} I_1^{-1} I_2 I_3^{-1} I_2^\top I_1^{-1} + o(n_t^{-1} + n_s^{-1}). \end{aligned} \tag{53}$$

D.11 Proof of Theorem 5

Let η be defined as

$$\begin{aligned} \eta(\boldsymbol{\gamma}) \coloneqq &\sum_{i=1}^m \frac{\partial \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \int_0^{y_i} h_b(t \mid \boldsymbol{x}_i) S_2(y_i - t \mid \boldsymbol{x}_i) dt \\ &- \sum_{i=1}^m \frac{\frac{\partial \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \int_0^{y_i} \big(1 + \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) \int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) ds\big) h_b(t \mid \boldsymbol{x}_i) \exp\big(\int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) ds\big) S_2(y_i - t \mid \boldsymbol{x}_i) dt}{1 + \int_0^{y_i} h_b(t \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) \exp\big(\int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) ds\big) S_2(y_i - t \mid \boldsymbol{x}_i) dt}. \end{aligned} \tag{54}$$

Then the gradient of $\mathcal{E}_t((\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma}),\hat{\boldsymbol{\theta}}_2)$ can be represented as

$$
\begin{aligned}
&\frac{\partial}{\partial\boldsymbol{\gamma}}\mathcal{E}_t((\hat{\boldsymbol{\alpha}},\boldsymbol{\gamma}),\hat{\boldsymbol{\theta}}_2)\\
&=\frac{\partial}{\partial\boldsymbol{\gamma}}\Big(\sum_{i=1}^{m}\log S_{\blacksquare}(y_i\mid\boldsymbol{x}_i)+\sum_{i=m+1}^{n}\log\big(f_1(z_i\mid\boldsymbol{x}_i)f_2(y_i-z_i\mid\boldsymbol{x}_i)\big)\Big)\\
&=\frac{\partial}{\partial\boldsymbol{\gamma}}\Big(\sum_{i=1}^{m}\log\Big(S_1(y_i\mid\boldsymbol{x}_i)+\int_0^{y_i}f_1(t\mid\boldsymbol{x}_i)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big)+\sum_{i=m+1}^{n}\log f_1(z_i\mid\boldsymbol{x}_i)\Big)\\
&=\frac{\partial}{\partial\boldsymbol{\gamma}}\Big(\sum_{i=1}^{m}\Big(\log S_1(y_i\mid\boldsymbol{x}_i)+\log\Big(1+\int_0^{y_i}\frac{f_1(t\mid\boldsymbol{x}_i)S_2(y_i-t\mid\boldsymbol{x}_i)}{S_1(y_i\mid\boldsymbol{x}_i)}dt\Big)\Big)+\sum_{i=m+1}^{n}\log f_1(z_i\mid\boldsymbol{x}_i)\Big)\\
&=\frac{\partial}{\partial\boldsymbol{\gamma}}\sum_{i=1}^{m}\Big(-\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})dt+\log\Big(1+\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big)\Big)\\
&\quad+\frac{\partial}{\partial\boldsymbol{\gamma}}\sum_{i=m+1}^{n}\Big(\log\big(h_b(z_i\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\big)-\int_0^{z_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})dt\Big)\\
&=\sum_{i=1}^{n}\Big(v_i\frac{\partial\log\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}-\int_0^{\tilde{y}_i}h_b(t\mid\boldsymbol{x}_i)\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}dt\Big)\\
&\quad+\sum_{i=1}^{m}\frac{\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\int_0^{y_i}\big(1+\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)ds\big)h_b(t\mid\boldsymbol{x}_i)\exp\big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\big)S_2(y_i-t\mid\boldsymbol{x}_i)dt}{1+\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\big)S_2(y_i-t\mid\boldsymbol{x}_i)dt}\\
&=\boldsymbol{\zeta}(\boldsymbol{\gamma})+\sum_{i=1}^{m}\frac{\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\int_0^{y_i}\big(1+\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)ds\big)h_b(t\mid\boldsymbol{x}_i)\exp\big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\big)S_2(y_i-t\mid\boldsymbol{x}_i)dt}{1+\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\big)S_2(y_i-t\mid\boldsymbol{x}_i)dt}\\
&=\tilde{\boldsymbol{\zeta}}(\boldsymbol{\gamma})-\eta(\boldsymbol{\gamma}).
\end{aligned}
\tag{55}
$$

We analyze $\eta(\boldsymbol{\gamma})$ as follows:

$$
\begin{aligned}
|\eta(\boldsymbol{\gamma})|&\le\sum_{i=1}^{m}\Big|\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big(\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big)\\
&\qquad\times\Big(1+\int_0^{y_i}h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big)\\
&\qquad-\int_0^{y_i}\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big(1+\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)ds\Big)h_b(t\mid\boldsymbol{x}_i)\exp\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big|\\
&\le\sum_{i=1}^{m}\Big|\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big|\Big\{\int_0^{y_i}\Big(\exp\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\Big)-1\Big)h_b(t\mid\boldsymbol{x}_i)S_2(y_i-t\mid\boldsymbol{x}_i)dt\\
&\quad+\int_0^{y_i}\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)ds+\int_0^{y_i}h_b(s\mid\boldsymbol{x}_i)S_2(y_i-s\mid\boldsymbol{x}_i)ds\Big)\\
&\qquad\times h_b(t\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\int_t^{y_i}h_b(s\mid\boldsymbol{x}_i)\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})ds\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big\}\\
&\le\sum_{i=1}^{m}\Big|\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big|\Big\{\int_0^{y_i}\Big(\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)-1\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-t\mid\boldsymbol{x}_i)dt\\
&\quad+\int_0^{y_i}\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}+\int_0^{y_i}\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-s\mid\boldsymbol{x}_i)ds\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big\}\\
&\le\sum_{i=1}^{m}\Big|\frac{\partial\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}}\Big|\Big\{\int_0^{y_i-\delta_i}\Big(\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)-1\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-t\mid\boldsymbol{x}_i)dt\\
&\quad+\int_{y_i-\delta_i}^{y_i}\Big(\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)-1\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-t\mid\boldsymbol{x}_i)dt\\
&\quad+\int_0^{y_i-\delta_i}\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}+\int_0^{y_i}\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-s\mid\boldsymbol{x}_i)ds\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\\
&\quad+\int_{y_i-\delta_i}^{y_i}\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}+\int_0^{y_i}\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}S_2(y_i-s\mid\boldsymbol{x}_i)ds\Big)\frac{\nu_i}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\exp\Big(\frac{\nu_i(y_i-t)}{\phi(\boldsymbol{x}_i;\hat{\boldsymbol{\gamma}})}\phi(\boldsymbol{x}_i;\boldsymbol{\gamma})\Big)S_2(y_i-t\mid\boldsymbol{x}_i)dt\Big\}.
\end{aligned}
\tag{56}
$$

By the definition of δ_i, we have

$$\delta_i = \int_{\delta}^{y_i} \frac{S_2(t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2)}{\exp(-\nu_i t)} dt = \int_0^{y_i - \delta} \frac{S_2(y_i - t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2)}{\exp(-(y_i - t)\nu_i)} dt \geq \int_0^{y_i - \delta} S_2(y_i - t \mid \boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_2) dt. \tag{57}$$

By substituting $\boldsymbol{\gamma}$ with $\hat{\boldsymbol{\gamma}}$, we have

$$\begin{aligned}
|\eta(\hat{\boldsymbol{\gamma}})| &\leq \sum_{i=1}^{m} \boldsymbol{\pi}_i \Big\{ \big(\exp(\nu_i y_i) - 1\big)\nu_i\delta_i + \big(\exp(\nu_i\delta_i) - 1\big)\nu_i\delta_i \\
&\qquad + \Big(y_i + \int_0^{y_i} S_2(y_i - s \mid \boldsymbol{x}_i) ds\Big)\nu_i^2\delta_i + \Big(\delta_i + \int_0^{y_i} S_2(y_i - s \mid \boldsymbol{x}_i) ds\Big)\nu_i^2\delta_i \exp\big(\nu_i\delta_i\big) \Big\} \\
&= \sum_{i=1}^{m} \boldsymbol{\pi}_i \Big\{ \big(\exp(\nu_i y_i) - 1\big)\nu_i\delta_i + \big(\exp(\nu_i\delta_i) - 1\big)\nu_i\delta_i + \nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2 \exp\big(\nu_i\delta_i\big) \\
&\qquad + \Big(\int_0^{y_i - \delta_i} S_2(y_i - s \mid \boldsymbol{x}_i) ds + \int_{y_i - \delta_i}^{y_i} S_2(y_i - s \mid \boldsymbol{x}_i) ds\Big)\big(\nu_i^2\delta_i + \nu_i^2\delta_i \exp(\nu_i\delta_i)\big) \Big\} \\
&\leq \sum_{i=1}^{m} \boldsymbol{\pi}_i \Big\{ \big(\exp(\nu_i y_i) - 1\big)\nu_i\delta_i + \big(\exp(\nu_i\delta_i) - 1\big)\nu_i\delta_i + \nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2 \exp\big(\nu_i\delta_i\big) + 2\big(1 + \exp\big(\nu_i\delta_i\big)\big)\nu_i^2\delta_i^2 \Big\}.
\end{aligned} \tag{58}$$

Consequently, under Assumption 5, we obtain

$$\begin{aligned}
\eta(\hat{\boldsymbol{\gamma}}) &= \mathcal{O}\Big(\sum_{i=1}^{m} \Big(\big(\exp(\nu_i y_i) - 1\big)\nu_i\delta_i + \big(\exp(\nu_i\delta_i) - 1\big)\nu_i\delta_i + \nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2 \exp\big(\nu_i\delta_i\big) + 2\big(1 + \exp\big(\nu_i\delta_i\big)\big)\nu_i^2\delta_i^2 \Big)\boldsymbol{\pi}_i \Big) \\
&= \mathcal{O}\Big(\sum_{i=1}^{m} \Big(\nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2 + \nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2\big(1 + \nu_i\delta_i\big) + 2\big(2 + \nu_i\delta_i\big)\nu_i^2\delta_i^2 \Big)\boldsymbol{\pi}_i \Big) \\
&= \mathcal{O}\Big(\sum_{i=1}^{m} \big(\nu_i^2 y_i\delta_i + \nu_i^2\delta_i^2 + \nu_i^3\delta_i^3\big)\boldsymbol{\pi}_i \Big) \\
&= \mathcal{O}\Big(\sum_{i=1}^{m} \nu_i^2\delta_i y_i \boldsymbol{\pi}_i \Big).
\end{aligned} \tag{59}$$

Consequently, we have

$$\tilde{\boldsymbol{\zeta}}(\hat{\boldsymbol{\gamma}}) = \mathcal{O}\Big(\sum_{i=1}^{m} \nu_i^2\delta_i y_i \boldsymbol{\pi}_i \Big). \tag{60}$$

E Alternative Problem Setting

Let $\check{Z} \coloneqq \min\{T_1, C_1\}, \check{Y} \coloneqq \min\{T_1 + T_2, C_1 + C_2\}, \check{W} \coloneqq \mathbb{I}(T_1 \leq C_1)$, and $\check{V} \coloneqq \mathbb{I}(T_1 + T_2 \leq C_1 + C_2)$, where C_1 and C_2 are non-negative random variables. Assume that T_1, T_2, and (C_1, C_2) are mutually independent conditional on X. Define f_{c_1,c_2} as the joint density function of (C_1, C_2).

Proposition 5. *For $0 \leq y$, $v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\check{Y}, \check{V})$ is given by*

$$S_{\blacksquare}(y \mid \boldsymbol{x}) \int_0^y f_{c_1,c_2}(c_1, y - c_1 \mid \boldsymbol{x}) dc_1. \tag{61}$$

For $0 \leq z \leq y$, $(w, v) = (0, 1)$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\check{Z}, \check{Y}, \check{W}, \check{V})$ is given by

$$\int_z^y f_{\square}(t, y \mid \boldsymbol{x}) dt \int_{y-z}^{\infty} f_{c_1,c_2}(z, c_2 \mid \boldsymbol{x}) dc_2. \tag{62}$$

For $0 \leq z \leq y$, $(w, v) = (1, 1)$, and $\boldsymbol{x} \in \mathcal{X}$, the conditional joint distribution of $(\check{Z}, \check{Y}, \check{W}, \check{V})$ is given by

$$f_{\square}(z, y \mid \boldsymbol{x}) \int_z^{\infty} dc_1 \int_{\max\{y - c_1, 0\}}^{\infty} dc_2 f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}). \tag{63}$$

Proof. For $0 \le y, v = 0$, and $\boldsymbol{x} \in \mathcal{X}$, the following holds:

$$
\begin{aligned}
&\frac{\partial}{\partial y} \Pr(\check{Y} \le y, \check{V} = v \mid X = \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \Pr(C_1 + C_2 \le y, C_1 + C_2 < T_1 + T_2 \mid X = \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_0^\infty dc_1 \int_0^{\max\{y-c_1,0\}} dc_2 \int_0^\infty dt_1 \int_{\max\{c_1+c_2-t_1,0\}}^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_0^y dc_1 \int_0^{y-c_1} dc_2 \int_0^{c_1+c_2} dt_1 \int_{c_1+c_2-t_1}^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&\quad + \frac{\partial}{\partial y} \int_0^y dc_1 \int_0^{y-c_1} dc_2 \int_{c_1+c_2}^\infty dt_1 \int_0^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \int_0^y dc_1 \int_0^y dt_1 \int_{y-t_1}^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, y - c_1 \mid \boldsymbol{x}) \\
&\quad + \int_0^y dc_1 \int_y^\infty dt_1 \int_0^\infty dt_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, y - c_1 \mid \boldsymbol{x}) \\
&= S_{\blacksquare}(y \mid \boldsymbol{x}) \int_0^y f_{c_1,c_2}(c_1, y - c_1 \mid \boldsymbol{x}) dc_1.
\end{aligned} \tag{64}
$$

For $0 \le z \le y$, $(w, v) = (0, 1)$, and $\boldsymbol{x} \in \mathcal{X}$, the following holds:

$$
\begin{aligned}
&\frac{\partial^2}{\partial y \partial z} \Pr(\check{Z} \le z, \check{Y} \le y, \check{W} = w, \check{V} = v \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \Pr(C_1 \le z, T_1 + T_2 \le y, C_1 < T_1, T_1 + T_2 \le C_1 + C_2 \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dc_1 \int_{c_1}^\infty dt_1 \int_0^{\max\{y-t_1,0\}} dt_2 \int_{t_1+t_2-c_1}^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dc_1 \int_{c_1}^y dt_1 \int_0^{y-t_1} dt_2 \int_{t_1+t_2-c_1}^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_z^y dt_1 \int_0^{y-t_1} dt_2 \int_{t_1+t_2-z}^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(z, c_2 \mid \boldsymbol{x}) \\
&= \int_z^y f_{\square}(t, y \mid \boldsymbol{x}) dt \int_{y-z}^\infty f_{c_1,c_2}(z, c_2 \mid \boldsymbol{x}) dc_2.
\end{aligned} \tag{65}
$$

For $0 \le z \le y$, $(w, v) = (1, 1)$, and $\boldsymbol{x} \in \mathcal{X}$, the following holds:

$$
\begin{aligned}
&\frac{\partial^2}{\partial y \partial z} \Pr(\check{Z} \le z, \check{Y} \le y, \check{W} = w, \check{V} = v \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \Pr(T_1 \le z, T_1 + T_2 \le y, T_1 \le C_1, T_1 + T_2 \le C_1 + C_2 \mid X = \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dt_1 \int_0^{y-t_1} dt_2 \int_{t_1}^\infty dc_1 \int_{\max\{t_1+t_2-c_1,0\}}^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \frac{\partial^2}{\partial y \partial z} \int_0^z dt_1 \int_0^{y-t_1} dt_2 \int_{t_1}^{t_1+t_2} dc_1 \int_{t_1+t_2-c_1}^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&\quad + \frac{\partial^2}{\partial y \partial z} \int_0^z dt_1 \int_0^{y-t_1} dt_2 \int_{t_1+t_2}^\infty dc_1 \int_0^\infty dc_2 f_1(t_1 \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \frac{\partial}{\partial y} \int_0^{y-z} dt_2 \int_z^{z+t_2} dc_1 \int_{z+t_2-c_1}^\infty dc_2 f_1(z \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&\quad + \frac{\partial}{\partial y} \int_0^{y-z} dt_2 \int_{z+t_2}^\infty dc_1 \int_0^\infty dc_2 f_1(z \mid \boldsymbol{x}) f_2(t_2 \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= \int_z^y dc_1 \int_{y-c_1}^\infty dc_2 f_1(z \mid \boldsymbol{x}) f_2(y - z \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) + \int_y^\infty dc_1 \int_0^\infty dc_2 f_1(z \mid \boldsymbol{x}) f_2(y - z \mid \boldsymbol{x}) f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}) \\
&= f_{\square}(z, y \mid \boldsymbol{x}) \int_z^\infty dc_1 \int_{\max\{y-c_1,0\}}^\infty dc_2 f_{c_1,c_2}(c_1, c_2 \mid \boldsymbol{x}).
\end{aligned} \tag{66}
$$

□

Proposition 6. *Let $\check{\mathcal{E}}$ be defined as*

$$\check{\mathcal{E}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \coloneqq \sum_{i=1}^{m} \log S_{\blacksquare}(y_i \mid \boldsymbol{x}_i) + \sum_{i=m+1}^{n} \Big(w_i \log f_{\square}(z_i, y_i \mid \boldsymbol{x}_i) + (1 - w_i) \log \int_{z_i}^{y_i} f_{\square}(t, y_i \mid \boldsymbol{x}_i) dt \Big). \tag{67}$$

Under Assumption 1, the estimator of $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ obtained by maximizing $\check{\mathcal{E}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ in the absence of administrative censoring is asymptotically consistent.

Proof. Consider the case $(W, V) = (1, 1)$ in the proof of Theorem 1. The proposition follows immediately from the same argument. □

In this setting, the occurrence of an earlier event after the first censoring time is reported, but its timing is not observed. When $\Pr(C_2 = 0) = 1$, this setting is equivalent to that described in Section 3.1.

F Supplementary Theoretical Results

F.1 Restricted Class of Density Functions

Let $f(t; C) = C$ if $t \leq \tau_0$; otherwise, let $f(t; C) = (1 - C\tau_0) f_0(t)$, where τ_0 is a known parameter satisfying $\tau \leq \tau_0$, C is an unknown parameter satisfying $0 < C < \tau_0^{-1}$, and $f_0(t)$ is a known positive function such that $\int_{\tau_0}^{\infty} f_0(t) dt = 1$. A single function from this class is identifiable on $[0, \tau]$, since $f(t; C) = C$ on this interval. In contrast, the product of two functions from this class is not identifiable on $[0, \tau]$: only the product $C_1 C_2$ is observed, so different pairs satisfying $C_1 C_2 = C_1' C_2'$ yield the same product. For example, whenever the admissibility constraints are satisfied, $(C_1, C_2) = (1, 1)$ and $(C_1', C_2') = (0.5, 2)$ yield the same product on $[0, \tau]$. Therefore, this case does not satisfy Assumption 3.

F.2 Bias Induced by Ignoring Reporting Delays

The second term in $S_{\blacksquare}(y \mid \boldsymbol{x})$ captures the probability that the event occurrence is unreported. The following proposition shows that omitting this term can introduce bias in the estimation of $\boldsymbol{\gamma}$.

Assumption 6. For each $1 \leq i \leq m$, the gradient $\frac{\partial}{\partial \boldsymbol{\gamma}} \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})|_{\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}}$ is strictly positive componentwise.

Proposition 7. *Under Assumption 6, $\boldsymbol{\zeta}(\hat{\boldsymbol{\gamma}})$ is strictly negative componentwise for $m \geq 1$.*

Proof. As in Equation (55), the gradient of $\mathcal{E}_t((\hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}), \hat{\boldsymbol{\theta}}_2)$ can be represented as

$$\boldsymbol{\zeta}(\boldsymbol{\gamma}) + \sum_{i=1}^{m} \frac{\frac{\partial \phi(\boldsymbol{x}_i; \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \int_0^{y_i} \big(1 + \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) \int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) ds\big) h_b(t \mid \boldsymbol{x}_i) \exp\big(\int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) ds\big) S_2(y_i - t \mid \boldsymbol{x}_i) dt}{1 + \int_0^{y_i} h_b(t \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) \exp\big(\int_t^{y_i} h_b(s \mid \boldsymbol{x}_i) \phi(\boldsymbol{x}_i; \boldsymbol{\gamma}) ds\big) S_2(y_i - t \mid \boldsymbol{x}_i) dt}. \tag{68}$$

Under Assumption 6, the last factor on the right-hand side is strictly positive componentwise. Consequently, $\boldsymbol{\zeta}(\hat{\boldsymbol{\gamma}})$ is strictly negative componentwise. □

Assumption 6 holds, for example, when $\phi(\boldsymbol{x}; \boldsymbol{\gamma}) = \exp(\boldsymbol{\gamma}^\top \boldsymbol{x})$ and $\boldsymbol{x}$ are strictly positive componentwise.

G Supplementary Experiments

G.1 Computational Environment

We used a 64-bit Windows machine with the Intel Core i9-9900K @ 3.60 GHz processor and 64 GB of RAM. The experiments were implemented in Python 3.7.3 (Python Software Foundation License Version 2), using lifelines 0.27.8 (MIT License) and SciPy 1.2.1 (BSD 3-Clause License).

G.2 Data Preprocessing

For the variable *rx* in the `Colon` dataset, we encoded the levels *Obs*, *Lev*, and *Lev+5FU* as 0, 1, and 2, respectively. For the variable *size* in the `Rotterdam` dataset, we encoded the levels $<= 20$, $20 - 50$, and > 50 as 0, 1, and 2, respectively. All variables were standardized. In the second stage of the transfer-learning procedure, we included an intercept term in the covariate vector.

G.3 Parameter Estimation Performance

The coverage probabilities for the source-domain parameters range from 0.914 to 0.958. Those for the target-domain parameters range from 0.894 to 0.970, except for γ_1 or the Toy1 setting with $\tau = 0.25$. Thus, the estimated standard errors generally yield coverage probabilities reasonably close to the nominal 95% level.

Table 2: Parameter estimation results for *Ours* in the toy model. For each parameter, we report the bias, root mean squared error (RMSE), mean estimated standard error (MeanSE), and empirical coverage probability of the nominal 95% Wald confidence interval. The asymptotic covariance in Equation (14) is modified to account for the approximate estimating equation. We conducted 500 trials.

τ	Parameter	Toy1				Toy2			
		Bias	RMSE	MeanSE	Coverage	Bias	RMSE	MeanSE	Coverage
∞	$\log\lambda$	0.014	0.178	0.166	0.924	0.001	0.100	0.094	0.942
	$\log\alpha_1$	-0.111	0.272	0.239	0.936	-0.050	0.182	0.169	0.944
	$\log\alpha_2$	-0.092	0.259	0.238	0.938	-0.034	0.172	0.164	0.942
	$\log\alpha_3$	-0.029	0.259	0.238	0.926	-0.013	0.143	0.144	0.958
	β_1	0.051	0.236	0.218	0.928	0.005	0.160	0.151	0.932
	β_2	0.008	0.236	0.208	0.914	-0.006	0.154	0.146	0.930
	β_3	0.007	0.218	0.208	0.934	0.002	0.148	0.145	0.948
	β_4	-0.001	0.222	0.208	0.928	-0.006	0.157	0.145	0.920
	β_5	0.009	0.226	0.208	0.936	-0.008	0.152	0.145	0.946
	β_6	-0.017	0.211	0.208	0.954	0.009	0.155	0.145	0.932
	β_7	-0.005	0.220	0.208	0.936	-0.001	0.155	0.145	0.946
	β_8	0.009	0.240	0.209	0.922	0.003	0.156	0.145	0.942
	β_9	-0.005	0.213	0.208	0.948	-0.001	0.157	0.145	0.934
	β_{10}	-0.007	0.219	0.208	0.940	0.012	0.158	0.145	0.934
0.25	γ_1	-0.522	0.799	0.501	0.734	-0.194	0.439	0.363	0.872
	γ_2	-0.009	0.590	0.480	0.868	0.007	0.344	0.343	0.952
	γ_3	-0.028	0.609	0.480	0.828	0.002	0.343	0.343	0.936
	γ_4	0.016	0.575	0.475	0.868	-0.011	0.344	0.345	0.950
	γ_5	0.024	0.613	0.475	0.824	0.011	0.359	0.340	0.936
	γ_6	-0.025	0.610	0.475	0.834	-0.021	0.375	0.341	0.900
	γ_7	0.004	0.591	0.480	0.842	0.009	0.352	0.344	0.940
	γ_8	-0.023	0.592	0.482	0.844	0.006	0.347	0.342	0.942
	γ_9	0.005	0.562	0.477	0.864	-0.003	0.355	0.341	0.936
	γ_{10}	0.018	0.611	0.482	0.848	0.005	0.356	0.344	0.932
1	γ_1	-0.307	0.421	0.280	0.784	-0.045	0.204	0.199	0.946
	γ_2	-0.023	0.266	0.244	0.918	0.000	0.170	0.168	0.958
	γ_3	0.012	0.282	0.241	0.894	0.002	0.168	0.168	0.950
	γ_4	-0.007	0.261	0.243	0.940	0.008	0.172	0.169	0.938
	γ_5	0.005	0.270	0.243	0.922	0.006	0.177	0.168	0.940
	γ_6	-0.016	0.242	0.242	0.940	-0.002	0.173	0.168	0.942
	γ_7	0.002	0.274	0.243	0.920	0.005	0.172	0.168	0.946
	γ_8	0.004	0.265	0.244	0.932	-0.004	0.163	0.168	0.954
	γ_9	-0.007	0.265	0.242	0.908	-0.003	0.165	0.167	0.970
	γ_{10}	0.005	0.269	0.242	0.916	0.002	0.168	0.168	0.958

H Broader Impacts

This work provides a methodological contribution to survival analysis under reporting delays. It may support risk evaluation in applications such as biostatistics and insurance, where timely risk evaluation is important but event information may be delayed. Careful validation of the modeling assumptions is important, since violations of these assumptions may lead to inappropriate decision making.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the paper's scope and contributions, including the proposed transfer-learning procedure for survival modeling with right-censored reporting delays under administrative censoring.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section 6, particularly the dependence of the proposed transfer-learning procedure on the validity of the shared-parameter assumptions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result states its assumptions explicitly, and complete proofs are provided in Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 describes the data, evaluation metric, compared methods, and implementation details needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplemental material includes the code and instructions needed to reproduce the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 specifies all the training and test details necessary to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 5 reports the mean and standard deviation of the evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix G.1 reports the hardware and software environment used for the experiments, including the CPU, memory, operating system, and Python version.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We are not aware of any ethical concerns that would conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix H discusses potential positive impacts of more reliable survival analysis under reporting delays, as well as potential risks arising from model misspecification or inappropriate use in real-world decision-making.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release data or models that pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The software licenses are specified in Appendix G.1, and the sources of the publicly available datasets are described in Section 5.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new dataset, pretrained model, benchmark, or other external asset is introduced. The code used to reproduce the experiments is provided in the supplementary material.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, participant recruitment, or direct interaction with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve participant recruitment, intervention, or direct interaction with human subjects, and uses only existing publicly available datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the core methodology of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.