Institute for Economic Studies, Keio University

Keio-IES Discussion Paper Series

Positive-Unlabelled Survival Analysis

Tomoki Toyabe, Takahiro Hoshino, Taisuke Otsu

25 September, 2026
DP2026-021
https://ies.keio.ac.jp/en/publications/28064/

Institute for Economic Studies, Keio University
2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan
ies-office-group@keio.jp
25 September, 2026

Positive-Unlabelled Survival Analysis
Tomoki Toyabe, Takahiro Hoshino, Taisuke Otsu
Keio-IES DP2026-021
25 September, 2026
JEL Classification: C41, C14, C24
Keywords: conditional likelihood, positive–unlabelled learning, reverse hazard, semiparametric efficiency, survival analysis

Abstract

We study survival analysis with positive–unlabelled event-time data, where censoring times and covariates are observed for all subjects but event times only for labelled positives. Under conditionally independent censoring and allowing the labelling probability to depend on censoring time and covariates, conditioning on these variables eliminates the unknown labelling probability from the event-time density of labelled positives. We show that this conditional density identifies the reverse hazard and the relative event-time distribution over the observed follow-up range, while absolute event probabilities require additional information or restrictions. We develop parametric and semiparametric estimators based on the conditional likelihood and establish their large-sample properties. In regular parametric models, the estimator is semiparametrically efficient when the labelling mechanism is locally unrestricted. We also show that information about scale parameters can vanish under limited follow-up and propose a sensitivity analysis for event-timedependent labelling. In simulations, we examine finite-sample performance under varying follow-up and label selection. In an empirical application, we study the time from clinical-trial completion to the first results disclosure recorded through a registry posting or qualifying journal publication.

Tomoki Toyabe
Kanazawa Gakuin University
toyabe@kanazawa-gu.ac.jp

Takahiro Hoshino
Keio University
bayesian@keio.jp

Taisuke Otsu
London School of Economics and Political Science
T.Otsu@lse.ac.uk

Positive-Unlabelled Survival Analysis

Tomoki Toyabe* Takahiro Hoshino† Taisuke Otsu‡

25 September 2026

Abstract

We study survival analysis with positive–unlabelled event-time data, where censoring times and covariates are observed for all subjects but event times only for labelled positives. Under conditionally independent censoring and allowing the labelling probability to depend on censoring time and covariates, conditioning on these variables eliminates the unknown labelling probability from the event-time density of labelled positives. We show that this conditional density identifies the reverse hazard and the relative event-time distribution over the observed follow-up range, while absolute event probabilities require additional information or restrictions. We develop parametric and semiparametric estimators based on the conditional likelihood and establish their large-sample properties. In regular parametric models, the estimator is semiparametrically efficient when the labelling mechanism is locally unrestricted. We also show that information about scale parameters can vanish under limited follow-up and propose a sensitivity analysis for event-time-dependent labelling. In simulations, we examine finite-sample performance under varying follow-up and label selection. In an empirical application, we study the time from clinical-trial completion to the first results disclosure recorded through a registry posting or qualifying journal publication.

Keywords: conditional likelihood; positive–unlabelled learning; reverse hazard; semiparametric efficiency; survival analysis.
JEL classification: C41, C14, C24.

1 Introduction

In survival analysis with right-censored data, we typically observe whether an event occurred before censoring for every subject. In many applications, however, event times are observed only for a subset of subjects known to have experienced the event. For the remaining subjects, follow-up times and covariates are observed, but whether the event occurred before the end of follow-up is unknown. This unlabelled group thus contains both subjects who experienced

*Kanazawa Gakuin University.
†Keio University and RIKEN. Email: hoshino@econ.keio.ac.jp.
‡London School of Economics and Political Science.

the event before censoring and subjects who did not. Examples include disease registries with incomplete linkage, electronic health records, and digital platforms where outcomes are observed only through selected reporting channels. We refer to this setting as positive–unlabelled (PU) survival analysis.

Let T be the event time and C the censoring time. For labelled positives ($S = 1$), we observe (T, C, X) and know that $T \leq C$. For unlabelled subjects ($S = 0$), we observe only (C, X) and do not know whether $T \leq C$. Thus the conventional likelihood function for right-censored data cannot be applied.

Under conditionally independent censoring and allowing the probability of $S = 1$ to depend on C and X, the conditional density of T for labelled positives given (C, X) can be written as

$$q(t \mid c, x) = \frac{f_T(t \mid x)}{F_T(c \mid x)}, \qquad 0 < t < c,$$

where the unknown labelling probability cancels. Thus, variation in c allows us to identify the reverse hazard and relative event-time distribution. With additional information or structural restrictions, the event-time distribution itself can also be identified.

We make three contributions. First, we establish identification based on the conditional density q. Second, we develop parametric and semiparametric estimators and study their efficiency and properties under limited follow-up, where censoring occurs early relative to the event-time distribution. Third, we propose a sensitivity analysis for event-time-dependent labelling and examine its performance in simulations and an empirical example.

This paper is related to the literature on PU learning and survival analysis. PU learning considers settings where positively labelled observations are available but labelled negatives are absent (Elkan and Noto, 2008; Bekker and Davis, 2020). Existing work has focused primarily on classification and binary outcomes, where identification relies on assumptions on the labelling mechanism or the class prior. In survival analysis, the comparison between the event and censoring times provides additional information for identification. Estimation with PU survival data was considered in an early preprint by Toyabe et al. (2020), but identification results were not developed there.

This paper is also related to the literature on right-censored survival data (Kaplan and Meier, 1958; Cox, 1972; Andersen et al., 1993), semiparametric efficiency in event-time models (Zeng and Lin, 2007), truncation (Lynden-Bell, 1971; Woodroofe, 1985), and biased sampling (Vardi, 1989). The conditional density q has the same form as the likelihood contribution under classical random right truncation (Kalbfleisch and Lawless, 1989), but the sampling schemes are different. In PU survival data, labelled positives are selected from subjects with $T \leq C$ through an unknown labelling mechanism, while the censoring-time distribution is observed from the full cohort or a separate population sample. These differences lead to a different identification problem.

The remainder of the paper is organized as follows. Section 2 defines the observation structure, assumptions, and estimands. Section 3 presents the identification results. Section 4 develops estimation and inference procedures. Section 5 reports simulation evidence, Section 6 presents the empirical application, and Section 7 concludes. Proofs of all theoretical results are provided in the Online Appendix.

2 Setup

2.1 Data and estimands

Let $T \in \mathbb{R}_+$ be the event time, $C \in \mathbb{R}_+$ be the censoring time, and $X \in \mathcal{X} \subseteq \mathbb{R}^d$ be a vector of baseline covariates. Define $Y = \mathbf{1}\{T \leq C\}$ and let $S \in \{0, 1\}$ be a positive label satisfying $S = 1 \Rightarrow Y = 1$. Thus $S = 1$ indicates that the event occurred before censoring, whereas $S = 0$ does not reveal whether $Y = 1$ or not. We assume that C is observed for all subjects, including labelled positives. The observed data are

$$O = \begin{cases} (S, T, C, X), & S = 1, \\ (S, C, X), & S = 0. \end{cases}$$

Observing C for labelled positives is important for identification, as shown in Proposition 1. In contrast, C need not be observed for every unlabelled subject (see Section S7.1 of the Online Appendix for this extension).

We consider two sampling frames. In the *single-cohort frame*, a target cohort is sampled, (C, X) is observed for every subject, and a subset of subjects with $Y = 1$ receives a positive label with an observed event time. In the *case-population frame*, labelled positives are sampled from the event-before-censoring population and unlabelled observations independently from the target population. The identification results below apply to both frames.

We consider three estimands. Let $F_T(t \mid x) = \mathbb{P}(T \leq t \mid X = x)$ and let $G_C(\cdot \mid x)$ be the conditional distribution of C. The first is the relative event-time distribution

$$\frac{F_T(t \mid x)}{F_T(\tau_x \mid x)},$$

where τ_x is the upper endpoint of the support of C given $X = x$. The second is the event-time distribution $F_T(t \mid x)$. The third is the event-before-censoring probability

$$\pi(x) = \mathbb{P}(T \leq C \mid X = x) = \int F_T(c \mid x)\, dG_C(c \mid x).$$

We will show that the relative distribution is nonparametrically identified, whereas the latter two require additional information that fixes the terminal level $F_T(\tau_x \mid x)$. We also allow F_T to be defective, so that some subjects may never experience the event.

2.2 Assumptions

For our identification analysis, we impose the following assumptions.

Assumption 1 (Conditionally independent censoring). *$T \perp C \mid X = x$ for all $x \in \mathcal{X}$. The event and censoring times admit conditional densities $f_T(\cdot \mid x)$ and $g_C(\cdot \mid x)$.*

Assumption 1 is the standard independent-censoring condition in survival analysis. It is natural when censoring is determined independently of the event time conditional on X, although relevant cohort, site, or calendar effects may need to be included in X.

Assumption 2 (Positive-label selection among event subjects). *For all (t, c, x) with $t \leq c$,*

$$\mathbb{P}(S = 1 \mid T = t, C = c, X = x, Y = 1) = \mathbb{P}(S = 1 \mid C = c, X = x, Y = 1).$$

Moreover, $\mathbb{P}(S = 1 \mid C = c, X = x, Y = 1) > 0$ on the target support.

Assumption 2 allows the probability of receiving a positive label to depend on the censoring time and covariates, but rules out additional dependence on the event time conditional on (C, X) among event subjects. This assumption is weaker than selected-completely-at-random (SCAR) sampling of positives, since the labelling probability may vary with (C, X). This assumption is not testable from the labelled positives alone.

Let $\mathcal{T}_x$ denote the set of $t > 0$ having a possibly one-sided neighbourhood on which $g_C(\cdot \mid x)$ is positive.

Assumption 3 (Support and terminal level). *For each $x \in \mathcal{X}$, (a) $\mathcal{T}_x = (0, \tau_x]$ for some $\tau_x > 0$; and (b) $F_T(t \mid x) > 0$ for all $t \in (0, \tau_x]$ and $F_T(\cdot \mid x)$ is continuous at τ_x.*

Throughout, τ_x is treated as known, as under an administrative follow-up horizon. Whenever absolute event-time probabilities are considered, we will state separately the additional information used to identify $F_T(\tau_x \mid x)$. We show that observing C for labelled positives is necessary for identification.

Proposition 1 (Lack of identification when C is missing for labelled positives). *Fix x. Suppose $T \perp C \mid X = x$, $S = 1 \Rightarrow T \leq C$, and $\mathbb{P}(S = 1 \mid T = t, C = c, X = x, Y = 1) = \rho(c, x)$ with $\rho(\cdot, x)$ unrestricted. Suppose we observe only*

$$O^\dagger = (S,\ T\,\mathbf{1}\{S = 1\},\ C\,\mathbf{1}\{S = 0\},\ X),$$

and $G_C(\cdot \mid x)$ is unknown. Then $F_T(\cdot \mid x)$ is not nonparametrically identified from the distribution of $O^\dagger \mid X = x$.

Without C for labelled positives, the truncation point associated with each observed event time is unknown. We therefore maintain the design in which C is observed for all labelled positives.

3 Identification

Conditioning on the realized censoring time eliminates the unknown labelling probability from the conditional distribution of event times among labelled positives.

Theorem 1 (Conditional density identity). *Under Assumptions 1 and 2, for $P_{C,X|S=1}$-almost every (c, x) and Lebesgue-almost every $t \in (0, c)$, the conditional density satisfies*

$$q(t \mid c, x) := f_{T|C,X,S=1}(t \mid c, x) = \frac{f_T(t \mid x)}{F_T(c \mid x)}. \tag{1}$$

For the same (c, x), $f_{T|C,X,S=1}(t \mid c, x) = 0$ for Lebesgue-almost every $t > c$.

Density identities are understood in this almost-everywhere sense. When Assumption 3 is imposed, integrating the density gives a conditional CDF ratio with a unique continuous version over the censoring-time support; pointwise ratio statements below refer to that version. This theorem shows that the unknown labelling probability $\rho(c, x) = \mathbb{P}(S = 1 \mid C = c, X = x, Y = 1)$ cancels from the conditional density $q(t \mid c, x)$. Variation in c then provides a family of truncation points that identifies relative event-time probabilities.

Remark 1 (Sensitivity to label-selection violations). *To examine violations of Assumption 2, we employ the exponentially tilted model with a finite positive normalizing integral*

$$q_\gamma(t \mid c, x) = \frac{\exp\{\gamma a(t, c, x)\} f_T(t \mid x)}{\int_0^c \exp\{\gamma a(u, c, x)\} f_T(u \mid x)\, du}, \qquad 0 < t < c,$$

where $a(t, c, x)$ is a specified contrast and γ is a sensitivity parameter. The benchmark model corresponds to $\gamma = 0$. When $a(t, c, x)$ is increasing in t, negative values of γ tilt the labelled-positive distribution toward earlier event times.

Integrating (1) yields a direct restriction on conditional event-time probabilities.

Corollary 1 (Conditional CDF ratio). *Under Assumptions 1 and 2, for $P_{C,X|S=1}$-almost every (c, x) and every $u \in [0, c)$, we have*

$$Q(u \mid c, x) := \mathbb{P}(T \le u \mid C = c, X = x, S = 1) = \frac{F_T(u \mid x)}{F_T(c \mid x)}. \tag{2}$$

The proof follows by integrating (1). Thus $q(t \mid c, x)$ identifies the ratio $F_T(u \mid x)/F_T(c \mid x)$ for supported $u < c$.

Taking the limit $c \downarrow t$ in (1) yields the reverse hazard.

Corollary 2 (Reverse hazard and distribution recovery). *Suppose that for Lebesgue-almost every $t \in \mathcal{T}_x$, the conditional density* (1) *is continuous in c from the right at $c = t$. Then*

(a) *for Lebesgue-almost every $t \in \mathcal{T}_x$,*

$$r_T(t \mid x) := \frac{f_T(t \mid x)}{F_T(t \mid x)} = \lim_{c \downarrow t} f_{T|C,X,S=1}(t \mid c, x), \tag{3}$$

and consequently

$$r_T(t \mid x) = \frac{\partial}{\partial t} \log F_T(t \mid x); \tag{4}$$

(b) *under Assumption 3(a)–(b), for every $t \in (0, \tau_x]$,*

$$F_T(t \mid x) = F_T(\tau_x \mid x) \exp\left\{ -\int_t^{\tau_x} r_T(u \mid x)\, du \right\}. \tag{5}$$

The object $r_T(t \mid x)$ is the reverse hazard rather than the usual hazard $f_T(t \mid x)/(1 - F_T(t \mid x))$. Equation (5) identifies the event-time distribution up to the terminal level $F_T(\tau_x \mid x)$.

The probability of the event occurring before censoring (or class prior) is $\pi(x) = \int F_T(c \mid x)\, dG_C(c \mid x)$. Thus $\pi(x)$ is identified whenever $F_T(\cdot \mid x)$ is identified on the support of $G_C(\cdot \mid x)$ and $G_C(\cdot \mid x)$ is known or estimable from the study design. Here $G_C(\cdot \mid x)$ is the population censoring-time distribution, not its distribution among labelled positives. Indeed, the latter is tilted by $\rho(c, x) F_T(c \mid x)$ relative to the target distribution; see Proposition S1 of the Online Appendix.

Theorem 2 (Identification of relative event-time distribution). *Fix x and suppose Assumptions 1, 2, and 3(a)–(b) hold. Then*

(a) *The relative distribution $\frac{F_T(t|x)}{F_T(\tau_x|x)}$ is identified for every $t \in (0, \tau_x]$.*

(b) *The terminal level $F_T(\tau_x \mid x)$ is not identified. For any $a(x) > 0$ such that $a(x) F_T(\tau_x \mid x) \le 1$, $\widetilde{F}_T(t \mid x) = a(x) F_T(t \mid x)$ generates the same conditional density $q(t \mid c, x)$ for $0 < t < c \le \tau_x$. Conversely, any event-time sub-distribution with a density on $(0, \tau_x]$ that generates the same q is of this form on $(0, \tau_x]$.*

(c) *If $F_T(\tau_x \mid x)$ is identified by additional information, then $F_T(\cdot \mid x)$ is identified on $(0, \tau_x]$. If, in addition, $G_C(\cdot \mid x)$ is known or estimable from the study design, then $\pi(x)$ is identified.*

Thus absolute event-time probabilities require additional information on the terminal level $F_T(\tau_x \mid x)$.

4 Estimation and inference

4.1 Parametric conditional likelihood

By Theorem 1, parametric estimation can proceed without modelling the positive-labelling probability. Let $\{F_\theta(t \mid x) : \theta \in \Theta\}$ be a parametric model for $T \mid X = x$, with density $f_\theta(t \mid x)$. The conditional log-likelihood for labelled positives is

$$\ell_n(\theta) = \frac{1}{n_1} \sum_{i:S_i=1} \{\log f_\theta(T_i \mid X_i) - \log F_\theta(C_i \mid X_i)\}, \tag{6}$$

where $n_1 = \sum_{i=1}^n S_i$. The maximum likelihood estimator is defined as $\hat{\theta} = \arg\max_{\theta\in\Theta} \ell_n(\theta)$.

Let $\psi_i(\theta) = \partial_\theta \{\log f_\theta(T_i \mid X_i) - \log F_\theta(C_i \mid X_i)\}$ for $S_i = 1$. We impose the standard regularity conditions for M-estimation (see (P1)–(P4) in Section S2.1 of the Online Appendix). The asymptotic properties of $\hat{\theta}$ are summarized in the following theorem.

Theorem 3 (Asymptotic normality of parametric MLE)**.** *Suppose Assumptions 1–2 and Conditions (P1)–(P4) hold and the parametric model is correctly specified with true value θ_0. Then*

$$\sqrt{n_1}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A^{-1}), \tag{7}$$

where $A = \mathbb{E}\Big[-\frac{\partial}{\partial\theta'}\psi_i(\theta_0)\,\Big|\, S_i = 1\Big]$.

Under misspecification, $\hat{\theta}$ converges to the pseudo-true parameter and the usual sandwich variance applies. Standard sample analogues provide consistent variance estimators.

4.2 Estimation of event-before-censoring probabilities

The estimator $\hat{\theta}$ does not depend on whether G_C is known or estimated, whereas estimation of

$$\pi(x) = \mathbb{P}(T \le C \mid X = x) = \int F_{\theta_0}(c \mid x)\, dG_C(c \mid x)$$

requires the censoring-time distribution. For the single-cohort frame, consider the marginal probability $\pi = \mathbb{E}[F_{\theta_0}(C \mid X)]$. When $G_C(\cdot \mid x)$ is known by design, we use $\hat{\pi}^{\text{known}} = \mathbb{P}_n \pi_{\hat{\theta}}(X)$, where $\pi_\theta(x) = \int F_\theta(c \mid x)\, dG_C(c \mid x)$. When G_C is estimated from the cohort, we use $\hat{\pi}^{\text{est}} = \mathbb{P}_n F_{\hat{\theta}}(C \mid X)$. Proposition S2 in Section S3 of the Online Appendix gives the asymptotic distributions of these estimators. Estimating G_C adds the within-covariate component $\mathbb{E}[\text{Var}\{F_{\theta_0}(C \mid X) \mid X\}]$ to the asymptotic variance.

For the case-population frame, $\hat{\theta}$ is estimated from the labelled-positive sample and G_C from the independent population sample. Proposition S2(c) establishes consistency of the corresponding plug-in estimator; the variance comparison above applies only to the single-cohort frame. The extension to missing censoring times for some unlabelled subjects is given in Section S7.1 of the Online Appendix.

Remark 2 (Limited follow-up and additional information)**.** *The conditional likelihood* (6) *remains valid under limited follow-up, but information about scale parameters may become weak. For example, when $F_{\theta_0}(C \mid x)$ is small throughout the support of C, the Weibull conditional density approaches the scale-free form $\frac{at^{a-1}}{c^a}$, so the shape remains informative while information about the scale vanishes. This phenomenon is studied formally in Proposition 4 below.*

Additional information can improve estimation of the scale. Such information may come from an external estimate $\pi_{\text{ext}}(x)$ of the event-before-censoring probability, a validation sample in which Y is observed, or restrictions on the labelling probability $\rho(c,x)$. In the absence of such information, profile likelihood-ratio intervals can be used for weakly identified scale parameters, as illustrated in Section 5.

4.3 Efficiency

Consider the model for labelled-positive observations, in which the density of $(T, C, X) \mid S = 1$ can be written as $q_\theta(t \mid c, x)m(c, x)$, where m is the unrestricted density of $(C, X) \mid S = 1$. The score for θ is orthogonal to the nuisance tangent space for m, so $\hat{\theta}$ is semiparametrically efficient in this model; see Theorem S1 in Section S2.2 of the Online Appendix. The following theorem gives the efficiency bound when both labelled and unlabelled observations are used.

Theorem 4 (Efficiency under locally unrestricted labelling)**.** *In the single-cohort frame, suppose q_θ is differentiable in quadratic mean at θ_0 with nonsingular information I_q, the marginal distribution of (C, X) is unrestricted, and Conditions* (E1)–(E3) *in Section S2.3 of the Online Appendix hold. Then the efficient score for θ based on all observations is*

$$\widetilde{\ell}_\theta(O) = S\dot{\ell}_q(T, C, X),$$

and the efficient information is $I_{\text{full}} = p_S I_q$, where $p_S = \mathbb{P}(S = 1)$. Under Conditions (P1)–(P4) *and correct specification, $\hat{\theta}$ attains this bound.*

The result relies on the labelling mechanism being locally unrestricted. If $\rho(c, x)$ is known or restricted, the labelling indicator can provide additional information about θ; the corresponding efficiency bound is given in Section S2.3 of the Online Appendix. The complete-labelling case $\rho \equiv 1$ is outside the locally unrestricted model and reduces to ordinary right-censored data.

The intuition is that the observed labelling probability depends on θ and ρ through

$$\mathbb{P}(S = 1 \mid C = c, X = x) = \rho(c, x)F_\theta(c \mid x).$$

When ρ is unrestricted, variation in $F_\theta(c \mid x)$ can be absorbed by $\rho(c, x)$, so the labelling indicator provides no additional information about θ. Restricting ρ can therefore make the unlabelled observations informative.

4.4 Semiparametric proportional-hazards estimation

We consider a proportional-hazards (PH) model with an unspecified baseline hazard

$$H_T(t \mid x) = e^{x'\beta} H_0(t), \qquad F_T(t \mid x) = 1 - \exp\{-e^{x'\beta} H_0(t)\},$$

where β is finite dimensional and H_0 is an unknown baseline cumulative hazard. We parameterize the baseline through the log-hazard $\eta_0 = \log h_0$ and define $(\hat\beta, \hat\eta)$ as the maximizer of the conditional likelihood over β and a B-spline sieve for η. The sieve construction and regularity conditions are given in Section S4.2 of the Online Appendix.

This PH structure can resolve the level indeterminacy in Theorem 2 when the covariate effect is nondegenerate.

Proposition 2 (Identification under PH)**.** *Suppose Assumptions 1, 2, and 3(a)–(b) hold, the PH distribution is proper (i.e., $F_T(\infty \mid x) = 1$ for all x), $x \mapsto \tau_x$ is lower semicontinuous on the support of $X \mid S = 1$, $\mathrm{Var}(X \mid S = 1)$ is nonsingular, and $x \mapsto x'\beta_0$ is nonconstant on the support of $X \mid S = 1$. Then the PH model is identified, i.e., if two PH distributions with parameters (β, H) and (β_0, H_0) generate the same $q(t \mid c, x)$, then $\beta = \beta_0$ and $H = H_0$ on $\mathcal{T}_x$ for every x in the support of $X \mid S = 1$.*

The nonconstant covariate effect is essential for this result. When $x'\beta_0$ is constant, the baseline level is not identified as shown below.

Proposition 3 (Loss of level identification)**.** *Suppose $x'\beta_0$ is constant on the support of X given $S = 1$, absorb the constant into the baseline, and let $F_0(t) = 1 - e^{-H_0(t)}$ with $F_0(\tau) < 1$. Then for every $a \in (0, 1/F_0(\tau))$, the PH model with baseline $\widetilde{H}_a(t) = -\log\{1 - aF_0(t)\}$ generates the same $q(t \mid c, x)$ on $\{0 < t < c \le \tau\}$.*

Thus absolute levels are not identified when the covariate effect is constant, although relative event-time probabilities remain identified. The corresponding local flatness result is given in Section S4.3.4 of the Online Appendix.

We next give the asymptotic properties of the sieve estimator. We work on a fixed horizon $[0, \tau]$ and use the capped sample defined in Section S4.1 of the Online Appendix. Let $D = C \wedge \tau$ be the capped censoring time, c_{L} the lower trimming point, S^τ the capped label, n_1 the number of labelled-positive observations in this sample, and t_0 the lower endpoint specified in Condition (S3).

Theorem 5 (Consistency and convergence rates). *Under Conditions* (S1)–(S4)*, with* $J_n \asymp n_1^{1/(2p+1)}$*, we have*

$$\begin{aligned} &\|\hat{\beta} - \beta_0\| + \|\hat{\eta} - \eta_0\|_{L_2([0,\tau])} = O_p\Big(n_1^{-p/(2p+1)}\Big), \\ &\|\hat{\eta} - \eta_0\|_\infty = O_p\Big(n_1^{-(p-1/2)/(2p+1)}\Big), \\ &\sup_{t\in[t_0,\tau],\, x\in \mathrm{supp}(X|S^\tau=1)} |\hat{F}_T(t \mid x) - F_T(t \mid x)| = O_p\Big(n_1^{-p/(2p+1)}\Big). \end{aligned}$$

Theorem 6 (Asymptotic normality and efficiency). *Under the conditions of Theorem 5 and Conditions* (S5)–(S6)*,*

$$\sqrt{n_1}(\hat{\beta} - \beta_0) \xrightarrow{d} N\Big(0, \mathcal{I}(\beta_0)^{-1}\Big),$$

where $\mathcal{I}(\beta_0)$ *is the semiparametric efficient information for* β *based on the labelled-positive observations.*

Under undersmoothing, pointwise root-n inference is also available for smooth functionals of the baseline distribution, including $H_0(t)$, $F_T(t \mid x)$, and the event-before-censoring probability over the capped horizon. The formal results and variance estimation are given in Theorem S2 and Proposition S3 of the Online Appendix.

Limited follow-up can make the baseline scale weakly identified even when the PH model is identified. The following proposition quantifies the resulting loss of information.

Proposition 4 (Vanishing scale information). *Along a sequence of capped experiments with* $\sup_{d\in[c_{\mathrm{L}},\tau],x} A(d, x) \to 0$*, where* $A(d, x) = e^{x'\beta_0} H_0(d)$*,*

$$\mathcal{I}_{\mathrm{sc}} = \frac{1}{12}\mathbb{E}[F_T(D \mid X)^2 \mid S^\tau = 1](1 + o(1)),$$

where $\mathcal{I}_{\mathrm{sc}}$ *is the information for the one-dimensional baseline-scale submodel.*

Thus information about the baseline scale vanishes quadratically as follow-up becomes limited. The corresponding local asymptotic minimax bound is given in Theorem S3 of the Online Appendix.

5 Simulation

We examine the finite-sample properties of the proposed method in simulations. The baseline data-generating process is a Weibull PH model with $H_T(t \mid x) = e^{x'\beta_0}(t/\lambda_0)^{a_0}$, where $a_0 = 1.3$, $\lambda_0 = 30$, and $\beta_0 = (0.5, -0.4)$. The covariates are $X = (X_1, X_2)$, with $X_1 \sim N(0, 1)$ and $X_2 \sim \mathrm{Bernoulli}(0.5)$, and $T \perp C \mid X$. The labelling probability among event subjects is

$$\rho(C, X) = \mathrm{expit}(-0.2 + 0.9\,\tilde{C} + 0.5X_1 + 0.4X_2),$$

Table 1: Main comparison at $n = 3{,}000$: bias (RMSE) by target

Estimator	$\log\lambda$	β_1	β_2	$F_T(20 \mid x_0)$	$\pi(x_0)$
Full-data MLE	0.000 (0.024)	0.001 (0.024)	0.000 (0.046)	0.000 (0.011)	0.000 (0.008)
Conditional likelihood	0.000 (0.062)	0.001 (0.058)	−0.004 (0.102)	0.000 (0.025)	0.000 (0.021)
Positive-only estimator	−0.277 (0.278)	−0.192 (0.194)	0.186 (0.194)	0.119 (0.120)	0.087 (0.088)
S-as-event estimator	0.914 (0.915)	0.051 (0.061)	0.368 (0.374)	−0.245 (0.246)	−0.311 (0.311)

Note. 5,000 replications; truth $(\log\lambda, \beta_1, \beta_2, F_T, \pi) = (3.401, 0.5, -0.4, 0.446, 0.701)$. The Monte Carlo standard error of each bias is at most 0.002.

where $\tilde{C} = (C - \bar{C})/\mathrm{sd}(C)$. Thus SCAR fails while Assumption 2 holds.

We compare the full-data MLE, the proposed conditional likelihood (6), the positive-only estimator, and an estimator that treats S as the event indicator and unlabelled subjects as right-censored at C. The targets are $\log\lambda$, β_1, β_2, $F_T(20 \mid x_0)$, and $\pi(x_0)$ at $x_0 = (0, 0)$.

Table 1 reports the main results for $C \sim \mathrm{Uniform}[5, 80]$ and $n = 3{,}000$, based on 5,000 replications. The proposed estimator is essentially unbiased for every target, with root-mean-squared error (RMSE) between 2.2 and 2.6 times that of the full-data MLE. The positive-only estimator is biased for every target. Treating S as the event indicator also produces substantial bias because unlabelled subjects who experienced the event are incorrectly treated as censored. Additional results in Section S5.2 of the Online Appendix show that bias and RMSE decrease with sample size and that coverage is close to nominal when follow-up is sufficiently long.

We next vary the upper limit of the censoring distribution to examine the weak-identification phenomenon in Proposition 4. As the follow-up period increases, the bias in $\log\lambda$ falls toward zero and its standard error decreases (Figure 1). At the two designs with the most limited follow-up, bias remains large and some observed Hessians are noninvertible. The profile likelihood-ratio intervals can then be unbounded, reflecting weak identification of the scale parameter. They become bounded and their coverage approaches the nominal level as follow-up increases.

In the label-selection experiment, the absolute bias of the $\gamma = 0$ estimator increases with $|\gamma|$, reaching 0.09 for F_T and 0.05 for π at $\gamma = -0.25$ (Figure S1 of the Online Appendix). In the application, we report estimates over analyst-specified ranges of γ.

6 Empirical application: public disclosure of clinical-trial results

The publicly available Aggregate Analysis of ClinicalTrials.gov (AACT) database (CTTI, 2026) provides administrative follow-up for all trials included in the analysis, while results disclosure is observed only through registry postings and linked journal publications. We analyse the time from the Primary Completion Date to the first results disclosure recorded in AACT through either source. We refer to this endpoint as the recorded disclosure time.

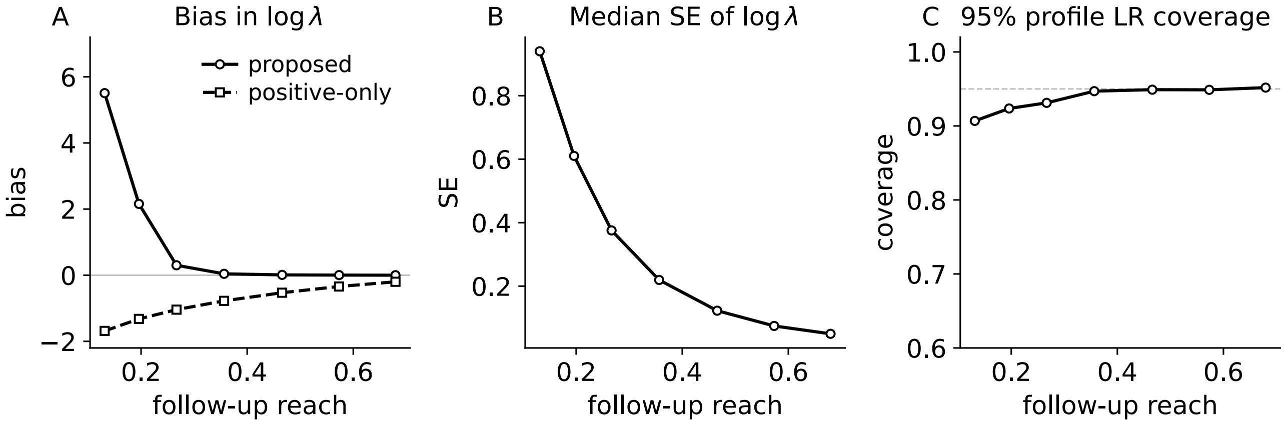

Figure 1: Follow-up and identification. A. Bias in $\log \lambda$ for the proposed and positive-only estimators. B. Median standard error for the proposed estimator. C. Coverage of the 95% profile likelihood-ratio interval for $\log \lambda$.

Alt text: Three side-by-side line charts with follow-up reach $\mathbb{E}[F_T(C \mid X)]$ on the horizontal axis. Panel A has bias on the vertical axis and shows the proposed and positive-only estimators. Panel B has standard error on the vertical axis and shows the median standard error of the proposed estimator. Panel C has coverage on the vertical axis and shows coverage of the 95% profile likelihood-ratio interval with a horizontal 95% reference.

6.1 Data and PU structure

For trial i, we use the AACT `primary_completion_date` field as the time origin d_i. From the 1 May 2026 snapshot, we include only interventional studies with `primary_completion_date_type` equal to `ACTUAL` and a Primary Completion Date between 1 January 2008 and 30 April 2026. A missing or estimated Primary Completion Date is not replaced with the Study Completion Date. Excluding 217 trials with a non-positive registry-posting delay leaves 264,889 trials. Their administrative follow-up period is $C_i = \tau_{\text{snap}} - d_i$, where τ_{snap} is the snapshot date.

The `results_first_posted_date` is the registry-posting date, and candidate journal publications come from AACT `DERIVED` and `RESULT` records. A fixed non-results PubMed Publication Type or citation phrase excludes a record; otherwise a `RESULT` record is retained and a `DERIVED` record is retained only with a clinical-trial Publication Type. The earliest retained date is used. Both the registry-posting and linked-publication dates must be later than d_i and no later than the snapshot date. Section S6.1 of the Online Appendix gives the complete field mappings and date and classification rules.

Let T_i be the time from d_i to the first qualifying results disclosure, whether through a registry posting or a qualifying journal publication, and define $Y_i = \mathbf{1}\{T_i \le C_i\}$. We set $S_i = 1$ if the AACT snapshot records a registry posting or contains a retained, dateable link to a qualifying journal publication. We assume that all registry postings and linked publications included in the analysis are genuine. Under this assumption, $S_i = 1$ implies $Y_i = 1$.

For labelled positives, we observe a recorded disclosure time $\widetilde{T}_i$. We treat $\widetilde{T}_i$ as T_i, which requires the first qualifying results disclosure and its date to be correctly recorded in the data used for the analysis. If an earlier qualifying publication is not linked in AACT, then $\widetilde{T}_i > T_i$.

For an unlabelled trial, $S_i = 0$, the registry-posting indicator is observed directly, so no registry posting had occurred by C_i. A qualifying journal publication may still have appeared by C_i without being included in the retained, dateable AACT publication links; in this case $Y_i = 1$ and the publication date and T_i are unavailable. Otherwise $Y_i = 0$. Thus, among trials without a recorded disclosure, those with and without a qualifying unlinked publication cannot be distinguished. This gives the PU structure described in Section 2.1.

The cohort contains 70,743 trials with registry postings and 62,435 with retained publication links, including 23,158 with both. The 110,020 trials in their union are the labelled positives. Under the assumption that $S_i = 1$ implies $Y_i = 1$, the observed fraction of 0.415 is a lower bound for the proportion of trials whose results were disclosed by the snapshot date.

The covariate vector x_i includes the trial phase and the explicit lead-sponsor class, and $z_i = \text{completion year}_i - 2016$ is the completion year centred at 2016. The government group consists of `NIH`, `FED`, and `OTHER_GOV`, `INDUSTRY` is a separate class, and the remaining sponsor classes form the reference group. Completion year enters the model as a numeric nuisance adjustment allowing a constant annual change on the log cumulative-hazard scale.

6.2 Parametric disclosure-time analysis

We compare the empirical distribution of the 110,020 recorded disclosure times with pooled Weibull fits in Figure 2. Fitting the pooled Weibull model by the conditional likelihood (6) gives a shape of 1.344 and a scale of 2.84 years, compared with a scale of 2.47 years for the positive-only fit. The conditional-likelihood fit therefore implies longer recorded disclosure times than the positive-only fit. Both disclosure timing and administrative follow-up vary with calendar time, so completion year is included in the covariate analysis below.

Some trials may never disclose results, so we allow the event-time distribution to be defective. We model the timing distribution conditional on eventual disclosure,

$$G(t \mid x, z) = \mathbb{P}(T \le t \mid T < \infty, X = x, Z = z).$$

The conditional likelihood identifies G, but not $\mathbb{P}(T < \infty \mid X = x, Z = z)$. The time ratios below therefore describe the timing of disclosure conditional on eventual disclosure, not the probability of disclosure.

For the covariate regression, we exclude trials completed after 2023 because their administrative follow-up is limited. Using the remaining 103,015 labelled positives, we fit

$$G(t \mid x, z) = 1 - \exp\left[-\exp\{x'\beta + \beta_{\mathrm{yr}} z\} \left(\frac{t}{\lambda}\right)^{a}\right], \qquad z = \text{completion year} - 2016,$$

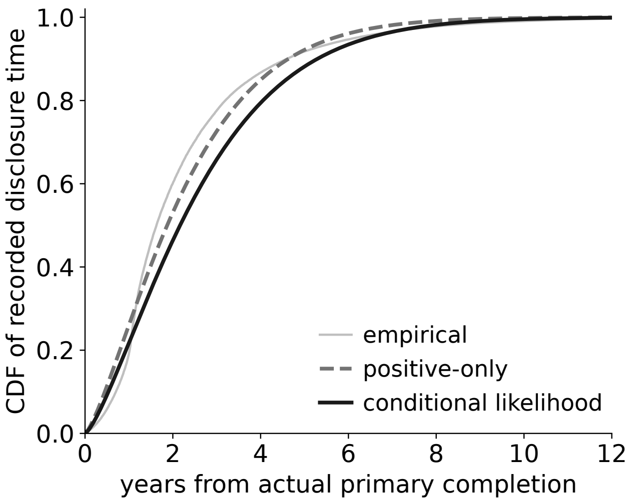

Figure 2: Clinical-trial results disclosure. Empirical distribution of times to first results disclosure recorded in AACT and Weibull timing distributions from the positive-only and conditional likelihood fits. Dashed curve: the positive-only fit; solid curve: the conditional likelihood fit.

Alt text: Line chart for the AACT analysis with years since actual Primary Completion Date on the horizontal axis and the empirical cumulative distribution of times to first results disclosure recorded in AACT on the vertical axis. It shows the empirical distribution among labelled positives, a dashed Weibull curve fitted to labelled positives only, and a solid conditional likelihood curve that lies to the right.

by the conditional likelihood (6). The common shape is 1.447, and the scale is 2.668 years at the 2016 reference setting, a Phase 2 trial in the reference sponsor group. Table 2 reports $\exp(-\beta_k/a)$ for each phase and sponsor coefficient, together with sandwich intervals based on the misspecification-robust covariance estimator in Section 4.1.

Table 2 shows differences in recorded disclosure times across trial phases and sponsor classes. The Phase 1/2 and industry time ratios are closest to one. These estimates describe associations with the timing of recorded disclosure in AACT; they do not identify effects on the probability of eventual disclosure or on the labelling probability. Replacing the linear completion-year term with a restricted cubic spline changes every phase and sponsor time ratio by less than 1%. In the five- and three-year common follow-up windows, the Phase 1/2, Phase 2/3, and industry ratios are close to one in at least one window, with intervals containing one, while the remaining contrasts retain their directions (Table S7 in Section S6.2 of the Online Appendix).

We examine violations of Assumption 2 using the exponential tilt in Remark 1. This analysis maintains the assumption that the recorded disclosure time equals the first qualifying disclosure time. For the analyst-specified ranges $\gamma \in [-0.15, 0]$, $[-0.20, 0]$, and $[-0.30, 0]$, the pooled Weibull scale increases from 2.84 years at $\gamma = 0$ to 3.21, 3.33, and 3.56 years, respectively. Detailed results are reported in Section S6.4 of the Online Appendix.

For registry posting, the by-snapshot event indicator is fully observed. The conditional-likelihood estimate of the Weibull scale is 3.41 years, while a full censored-data Weibull cure model gives a conditional scale of 3.31 years and an eventual-posting probability of 0.328. Thus

Table 2: Time ratios for first results disclosure recorded in AACT

Contrast	Time ratio	95% sandwich interval
Phase 1 vs Phase 2	1.156	$[1.132, 1.180]$
Phase 1/2 vs Phase 2	1.033	$[1.004, 1.064]$
Phase 2/3 vs Phase 2	0.934	$[0.898, 0.972]$
Phase 3 vs Phase 2	0.824	$[0.809, 0.839]$
Phase 4 vs Phase 2	0.874	$[0.856, 0.892]$
Phase not applicable vs Phase 2	1.046	$[1.031, 1.062]$
Industry vs other lead sponsor	0.986	$[0.974, 0.999]$
Government vs other lead sponsor	0.865	$[0.844, 0.885]$

Note. The endpoint is the time to first results disclosure recorded in AACT. Completion year, centred at 2016, is included as a nuisance covariate but is not displayed. A time ratio above one indicates a longer time.

the conditional-likelihood estimate of the timing distribution is close to that obtained when the event indicator is fully observed. Additional comparisons are reported in Sections S6.3–S6.6 of the Online Appendix.

7 Discussion

This paper studies survival analysis when event times are observed only for a subset of subjects known to have experienced the event. Under conditionally independent censoring and the labelling assumption in Assumption 2, conditioning on the censoring time removes the unknown labelling probability. The resulting conditional density identifies the reverse hazard and the relative event-time distribution over the observed follow-up range. Absolute event probabilities require additional information or restrictions.

The conditional likelihood provides a basis for parametric and semiparametric estimation without modelling the labelling mechanism. In the parametric model, it is efficient when the labelling mechanism is locally unrestricted, while restrictions on that mechanism may provide additional information. Limited follow-up can nevertheless lead to weak identification of scale parameters, as illustrated by the theory and simulations.

Our analysis assumes conditionally independent censoring and PU outcomes. Relaxing the censoring assumption, extending the framework to positive–negative–unlabelled (PNU) survival data, and combining the reverse-hazard estimating equations with conventional survival estimating equations are natural directions for future work.

Data availability

The application uses the 1 May 2026 monthly flat-file archive of the publicly available Aggregate Analysis of ClinicalTrials.gov (AACT) database (CTTI, 2026) and PubMed Publication Type metadata retrieved on 25 August 2026. AACT is provided by the Clinical Trials Transformation Initiative at `https://aact.ctti-clinicaltrials.org`, where snapshot downloads currently require an AACT account. The simulation studies require no external data.

Acknowledgements

This paper substantially revises and extends an earlier preprint, T. Toyabe, Y. Hasegawa, and T. Hoshino, "Positive-Unlabelled Survival Data Analysis" (arXiv:2011.13161, 2020): the present version develops the censoring-time conditional likelihood, the identification and efficiency theory, and the empirical analyses. We thank Y. Hasegawa for developing code for the earlier preprint.

Artificial-intelligence-assisted tools were used during manuscript preparation for language editing, code review, and consistency checks. All mathematical arguments, data definitions, statistical analyses, numerical results, and final wording were reviewed and approved by the authors, who take full responsibility for the manuscript.

Funding

This work was supported by JSPS KAKENHI Grant Number JP25K21651 and JST CREST, Japan, Grant Number JPMJCR2665 (Takahiro Hoshino), and JSPS KAKENHI Grant-in-Aid for Research Activity Start-up, Grant Number JP23K18788 (Tomoki Toyabe).

Conflict of interest

The authors declare no conflicts of interest.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer.

Bekker, J. and Davis, J. (2020), "Learning from Positive and Unlabeled Data: A Survey," *Machine Learning*, 109, 719–760.

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society: Series B*, 34, 187–220.

Clinical Trials Transformation Initiative (2026), "Aggregate Analysis of ClinicalTrials.gov (AACT) Database," `https://aact.ctti-clinicaltrials.org`. Accessed 24 September 2026.

Elkan, C. and Noto, K. (2008), "Learning Classifiers from Only Positive and Unlabeled Data," *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220.

Kalbfleisch, J. D. and Lawless, J. F. (1989), "Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS," *Journal of the American Statistical Association*, 84, 360–372.

Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481.

Lynden-Bell, D. (1971), "A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars," *Monthly Notices of the Royal Astronomical Society*, 155, 95–118.

Toyabe, T., Hasegawa, Y., and Hoshino, T. (2020), "Positive-Unlabelled Survival Data Analysis," arXiv preprint arXiv:2011.13161.

Vardi, Y. (1989), "Multiplicative Censoring, Renewal Processes, Deconvolution and Decreasing Density: Nonparametric Estimation," *Biometrika*, 76, 751–761.

Woodroofe, M. (1985), "Estimating a Distribution Function with Truncated Data," *The Annals of Statistics*, 13, 163–177.

Zeng, D. and Lin, D. Y. (2007), "Efficient Estimation for the Accelerated Failure Time Model," *Journal of the American Statistical Association*, 102, 1387–1396.

Online Appendix

S1 Proofs for identification results

S1.1 Proof of Proposition 1

Fix x and suppress the dependence on it in the notation. Let f, g be the densities of T, C on $[0, \tau]$, and write $h(c) = \rho(c)g(c)$ and $H(t) = \int_t^\tau h(c)\, dc$. When C is missing for labelled positives, the single-cohort law of $O^\dagger = (S, T\mathbf{1}\{S = 1\}, C\mathbf{1}\{S = 0\})$ is determined by the two subdensities

$$a(t) = \frac{d}{dt}\mathbb{P}(T \le t, S = 1) = f(t)\, H(t),$$

$$b(c) = \frac{d}{dc}\mathbb{P}(C \le c, S = 0) = g(c)\{1 - \rho(c)F(c)\} = g(c) - h(c)F(c),$$

together with the label prevalence $\mathbb{P}(S = 1) = \int a$. Reproducing the observed law therefore requires matching not only the labelled-positive law $f_{T|S=1} \propto a$ but also the unlabelled subdensity b and the prevalence $\int a$; matching $f_{T|S=1}$ alone does not suffice. We exhibit two distinct event-time laws with the same pair (a, b).

A smooth counterexample. Take $\tau = 1$. Under Model A,

$$f_A \equiv 1, \qquad g_A \equiv 1, \qquad \rho_A \equiv \tfrac{1}{2}, \qquad \text{so } F_A(t) = t,$$

whence $a_A(t) = \frac{1}{2}(1 - t)$, $b_A(c) = 1 - \frac{c}{2}$, and $\mathbb{P}_A(S = 1) = \int_0^1 a_A = \frac{1}{4}$. Under Model B take the distinct proper law

$$f_B(t) = \tfrac{3}{2} - t, \qquad F_B(t) = \tfrac{3}{2}t - \tfrac{1}{2}t^2 \quad \big(F_B(1) = 1,\ f_B \ge \tfrac{1}{2} > 0\big),$$

and define the censoring and labelling mechanism through

$$H_B(t) = \frac{a_A(t)}{f_B(t)} = \frac{1 - t}{3 - 2t}, \qquad h_B(c) = -H_B'(c) = \frac{1}{(3 - 2c)^2},$$

$$g_B = b_A + h_B F_B, \qquad \rho_B = \frac{h_B}{g_B}.$$

Then g_B is a density: by Fubini $\int_0^1 h_B F_B = \int_0^1 f_B H_B = \int_0^1 a_A = \frac{1}{4}$, so $\int_0^1 g_B = \int_0^1 b_A + \frac{1}{4} = \frac{3}{4} + \frac{1}{4} = 1$. The labelling probability is strictly interior: $g_B \ge b_A \ge \frac{1}{2} > 0$, and

$$g_B(c) - h_B(c) = b_A(c) - h_B(c)\{1 - F_B(c)\} = \frac{2 - c}{2}\left\{1 - \frac{1 - c}{(3 - 2c)^2}\right\} > 0$$

because $(3-2c)^2-(1-c)=4c^2-11c+8>0$ on $[0,1]$; hence $0<\rho_B<1$. Since $\rho_B g_B=h_B$, the observed subdensities coincide,

$$a_B(t)=f_B(t)\int_t^1 \rho_B g_B=f_B(t)H_B(t)=a_A(t), \qquad b_B(c)=g_B(c)-h_B(c)F_B(c)=b_A(c).$$

Models A and B therefore induce the identical single-cohort law of $O^\dagger$ (the same labelled-positive law, the same unlabelled-C law, and the same prevalence $\mathbb{P}(S=1)=\frac{1}{4}$), while $F_A(t)=t\neq F_B(t)=\frac{3}{2}t-\frac{1}{2}t^2$ (for instance $F_A(\frac{1}{2})=\frac{1}{2}\neq\frac{5}{8}=F_B(\frac{1}{2})$), so F_T is not identified.

Case-population frame. Here a direct counterexample keeps the population censoring distribution fixed. Let $g_A=g_B\equiv 1$ on $[0,1]$. Model A takes $f_A\equiv 1$ and $\rho_A\equiv 1/2$, giving $a_A(t)=\frac{1}{2}(1-t)$. Model B takes $f_B(t)=\frac{3}{2}-t$ and, for any $k\in(0,1/2)$, sets

$$H_B(t)=\int_t^1 \rho_B(c)\,dc=\frac{k(1-t)}{3/2-t}, \qquad \rho_B(c)=-H_B'(c)=\frac{k}{2(3/2-c)^2}.$$

Then $0<\rho_B(c)<1$ on $[0,1]$ and $a_B(t)=f_B(t)H_B(t)=k(1-t)$, so both models produce the same normalized positive-sample density $f_{T|S=1}(t)=2(1-t)$ and the same population law G_C, although $F_A\neq F_B$. The positive-sample size is fixed by design, so the different prevalences $\int a_A$ and $\int a_B$ are not observed. Thus the two models are observationally equivalent in the case-population frame. $\square$

S1.2 Proof of Theorem 1

For the relevant almost-everywhere versions and $t<c$, $Y=1$ is equivalent to $T\leq C$. By independent censoring,

$$f_{T,C|X,Y=1}(t,c\mid x)=\frac{f_T(t\mid x)g_C(c\mid x)}{\mathbb{P}(Y=1\mid X=x)}, \qquad t<c.$$

Marginalizing over T gives

$$g_{C|X,Y=1}(c\mid x)=\frac{g_C(c\mid x)F_T(c\mid x)}{\mathbb{P}(Y=1\mid X=x)}.$$

Therefore,

$$f_{T|C,X,Y=1}(t\mid c,x)=\frac{f_{T,C|X,Y=1}(t,c\mid x)}{g_{C|X,Y=1}(c\mid x)}=\frac{f_T(t\mid x)}{F_T(c\mid x)}, \qquad t<c.$$

By Assumption 2,

$$f_{T|C,X,S=1}(t\mid c,x)=f_{T|C,X,Y=1}(t\mid c,x),$$

which proves (1) for $P_{C,X|S=1}$-almost every (c,x) and Lebesgue-almost every $t<c$. If $t>c$, then $Y=1$ is impossible, and because $S=1\Rightarrow Y=1$, the conditional density is zero almost everywhere. □

S1.3 Proof of Corollary 2(a) (reverse-hazard identification)

The assumptions and almost-everywhere convention of Theorem 1 remain in force for part (a). In this proof, the limit $c \downarrow t$ and continuity in c are relative to the support of the conditional distribution of C given $X = x$ and $S = 1$. A value t is covered by that condition only if $\mathbb{P}(t < C < t + \delta \mid X = x, S = 1) > 0$ for every $\delta > 0$. Hence, under the corollary's almost-everywhere continuity hypothesis, the set of values $t \in \mathcal{T}_x$ for which no such approach is available has Lebesgue measure zero.

By Fubini's theorem, for $P_{X|S=1}$-almost every x and Lebesgue-almost every such t, the identity holds for $P_{C|X=x,S=1}$-almost every supported $c > t$:

$$f_{T|C,X,S=1}(t \mid c, x) = \frac{f_T(t \mid x)}{F_T(c \mid x)}.$$

Fix such an x and a covered t. The conditional density of C given $X = x$ and $S = 1$ is $g_{C|X,S=1}(c \mid x) = \rho(c, x) F_T(c \mid x) g_C(c \mid x)/\mathbb{P}(S = 1 \mid X = x)$, so the set on which $F_T(c \mid x) = 0$ has conditional probability zero. Every interval immediately to the right of t has positive conditional probability, whereas the identity above fails only on a conditional null set. We may therefore choose $c_m \downarrow t$ from the equality set with $F_T(c_m \mid x) > 0$. If $F_T(t \mid x) = 0$, monotonicity and continuity imply that t is the supremum of the set on which $F_T(\cdot \mid x)$ is zero; for each x there is at most one such value. Excluding this point and the preceding null sets, the assumed continuity of the conditional-density version in c from the right and continuity of $F_T(\cdot \mid x)$ along c_m give

$$\lim_{c \downarrow t} f_{T|C,X,S=1}(t \mid c, x) = \lim_{m \to \infty} \frac{f_T(t \mid x)}{F_T(c_m \mid x)} = \frac{f_T(t \mid x)}{F_T(t \mid x)}.$$

On every compact subinterval on which $F_T(\cdot \mid x) > 0$, $\log F_T(\cdot \mid x)$ is absolutely continuous, with derivative $f_T(\cdot \mid x)/F_T(\cdot \mid x)$ almost everywhere. This proves (3) and (4) on the asserted set. □

S1.4 Proof of Corollary 2(b) (distribution recovery)

Under Assumption 3(a)–(b), fix $t \in \mathcal{T}_x = (0, \tau_x]$. Then $F_T(\cdot \mid x)$ is positive on $[t, \tau_x]$, and $\log F_T(\cdot \mid x)$ is absolutely continuous on that interval. Integrating (4) gives

$$\log F_T(\tau_x \mid x) - \log F_T(t \mid x) = \int_t^{\tau_x} r_T(u \mid x)\, du.$$

Rearranging gives (5). □

S1.5 Proof of Theorem 2

(a) First take a supported c at which the conditional-density identity of Theorem 1 holds. The denominator $F_T(c \mid x) = \mathbb{P}(T \leq c \mid X = x) > 0$ is constant in t. For $0 < u < c$, integration over $t \in (0, u]$ gives

$$Q(u \mid c, x) = \int_0^u q(t \mid c, x)\, dt = \frac{1}{F_T(c \mid x)} \int_0^u f_T(t \mid x)\, dt = \frac{F_T(u \mid x)}{F_T(c \mid x)},$$

the equality holding for every $u < c$ because densities that agree almost everywhere induce the same CDF. The right side is continuous in (u, c) on the supported region. It therefore supplies the unique continuous version of Q, extending identification from almost every supported c to every supported pair $0 < u < c \leq \tau_x$.

Set $c = \tau_x$: $Q(u \mid \tau_x, x) = F_T(u \mid x)/F_T(\tau_x \mid x)$ is identified for every $u \in (0, \tau_x]$, which is the relative distribution. When C is continuously distributed, $\{C = \tau_x\}$ is a null event and $Q(u \mid \tau_x, x)$ is recovered as the limit of $Q(u \mid c, x)$ as $c \uparrow \tau_x$ through the support of C; the limit exists and equals $F_T(u \mid x)/F_T(\tau_x \mid x)$ by continuity of $F_T(\cdot \mid x)$ at τ_x. (At $u = \tau_x$ the ratio is 1.)

(b) Let $a(x) > 0$ satisfy $a(x)F_T(\tau_x \mid x) \leq 1$, and define $\widetilde{F}_T(t \mid x) = a(x)F_T(t \mid x)$ with density $\widetilde{f}_T(t \mid x) = a(x)f_T(t \mid x)$ on $(0, \tau_x]$. Then $\widetilde{F}_T$ is a valid (possibly defective) sub-distribution: it is nondecreasing, right-continuous, and bounded by 1 on $(0, \tau_x]$. For almost every $0 < t < c \leq \tau_x$,

$$\widetilde{q}(t \mid c, x) = \frac{\widetilde{f}_T(t \mid x)}{\widetilde{F}_T(c \mid x)} = \frac{a(x)f_T(t \mid x)}{a(x)F_T(c \mid x)} = \frac{f_T(t \mid x)}{F_T(c \mid x)} = q(t \mid c, x),$$

so q is invariant to $a(x)$. Because two distinct values $a_1(x) \neq a_2(x)$ produce identical q but distinct terminal levels $a_j(x)F_T(\tau_x \mid x)$, the level is not identified from q alone. For the converse, suppose $\widetilde{F}_T(\cdot \mid x)$ has a density $\widetilde{f}_T(\cdot \mid x)$ on $(0, \tau_x]$ and generates the same q almost everywhere on $\{0 < t < c \leq \tau_x\}$. Applying step (a) to $\widetilde{F}_T$ gives $\widetilde{F}_T(u \mid x)/\widetilde{F}_T(c \mid x) = Q(u \mid c, x) = F_T(u \mid x)/F_T(c \mid x)$ for all $0 < u < c \leq \tau_x$. Setting $c = \tau_x$ (for continuously distributed C, taking the limit $c \uparrow \tau_x$ as in step (a), using continuity of F_T and $\widetilde{F}_T$ at τ_x) yields $\widetilde{F}_T(u \mid x) = a(x)F_T(u \mid x)$ for all $u \in (0, \tau_x]$ with $a(x) = \widetilde{F}_T(\tau_x \mid x)/F_T(\tau_x \mid x)$. Thus the rescaling class in (b) is exactly the class of distributions that generate the same q. □

S1.6 Statement and proof of Proposition S1

Let $G_{C,X|S=s}$ be the conditional distribution of (C, X) given $S = s$, for $s \in \{0, 1\}$.

Proposition S1 (Target distribution of censoring time)**.** *Under Assumptions 1 and 2,*

$$dG_{C,X|S=1}(c,x) \propto \rho(c,x)F_T(c \mid x)\, dG_C(c \mid x)\, dF_X(x).$$

Hence the distribution of C among labelled positives generally differs from the target censoring distribution $G_C(\cdot \mid x)$.

By Bayes' rule, the joint distribution of (C, X) among labelled positives is proportional to $\mathbb{P}(S = 1 \mid C = c, X = x)\, dG_C(c \mid x)\, dF_X(x)$. Since $S = 1$ requires $Y = 1$ and, given $Y = 1$, occurs with probability $\rho(c, x)$,

$$\mathbb{P}(S = 1 \mid C = c, X = x) = \rho(c,x)\mathbb{P}(T \leq c \mid X = x) = \rho(c,x)F_T(c \mid x).$$

This gives the displayed result. Thus the distribution of (C, X) among labelled positives generally differs from its population distribution, and averaging $F_{\hat{\theta}}(C \mid X)$ over the positive sample does not in general recover

$$\pi(x) = \int F_T(c \mid x)\, dG_C(c \mid x).$$

In a single-cohort frame, the $S = 0$ subset is also generally not a population C-sample, since

$$dG_{C,X|S=0}(c,x) \propto \{1 - \rho(c,x)F_T(c \mid x)\}dG_C(c \mid x)\, dF_X(x).$$

Thus G_C must be obtained from the frame-specific C-sample: all sampled cohort units in the single-cohort frame, or the operationally unlabelled population sample in the case-population frame. □

S2 Parametric conditional likelihood and efficiency

S2.1 Conditions and proof of Theorem 3

The M-estimation conditions referenced in Theorem 3 are as follows.

(P1) $\theta_0 \in \mathrm{int}(\Theta)$ and Θ is compact and convex;

(P2) $\theta \mapsto \log q_\theta$ is twice continuously differentiable on an open convex neighbourhood of Θ, with $P|\log q_{\theta_0}| < \infty$, $P \sup_{\theta\in\Theta} \|\partial_\theta \log q_\theta\| < \infty$ and a dominated Hessian;

(P3) identification holds, $P \log q_\theta < P \log q_{\theta_0}$ for $\theta \neq \theta_0$;

(P4) $A = P[-\partial_{\theta\theta'} \log q_{\theta_0}]$ and $B = P[\psi\psi']$ exist and A is nonsingular, where P is throughout the distribution of (T, C, X) given $S = 1$.

By Theorem 1, the conditional density of T given $(C, X, S = 1)$ is $q_{\theta_0}(t \mid c, x) = f_{\theta_0}(t \mid x)/F_{\theta_0}(c \mid x)$. Thus $\hat\theta$ is the M-estimator maximizing $\mathbb{P}_{n_1} \log q_\theta$. The derivative envelope in (P2) makes $\{\log q_\theta : \theta \in \Theta\}$ Glivenko–Cantelli, and (P1)–(P3) give a well-separated maximizer. Theorem 5.7 of van der Vaart (1998) therefore gives $\hat\theta \to_p \theta_0$. Differentiating $\int q_\theta(t \mid c, x)\, dt = 1$ gives $\mathbb{E}\{\psi(\theta_0) \mid C, X, S = 1\} = 0$; hence (P4) and the standard expansion in Theorem 5.23 of the same reference yield (7). Under correct specification $A = B$; the sandwich covariance formula uses empirical versions of A and B. Finally, $n_1/n \to_p p_S > 0$ and the full-sample representation

$$n_1^{-1/2} \sum_{i:S_i=1} \psi_i = (n/n_1)^{1/2} n^{-1/2} \sum_{i=1}^{n} S_i \psi_i$$

extend the expansion to the labelled-positive count by Slutsky's theorem. $\square$

S2.2 Statement and proof of Theorem S1

Theorem S1 (Semiparametric efficiency)**.** *Consider the semiparametric model for a labelled positive observation*

$$\mathcal{M} = \Big\{ p_{\theta,m}(t, c, x \mid S = 1) = q_\theta(t \mid c, x)\, m(c, x) : \theta \in \Theta,\ m \in \mathcal{M}_{CX} \Big\},$$

where m ranges over all conditional densities of (C, X) given $S = 1$, thereby treating the induced distribution of $(C, X) \mid S = 1$ as an unrestricted nuisance. Suppose q_θ is differentiable in quadratic mean at θ_0 with score $\dot\ell_q(T, C, X) = \partial_\theta \log q_{\theta_0}(T \mid C, X)$ and nonsingular information $I_q = \mathbb{E}[\dot\ell_q \dot\ell_q' \mid S = 1]$. Then the nuisance tangent space is $\mathcal{T}_m = \{a(C, X) : \mathbb{E}[a(C, X) \mid S = 1] = 0\}$, the score $\partial_\theta \log q_\theta(T \mid C, X)$ is orthogonal to $\mathcal{T}_m$, and hence the efficient score for θ is $\tilde\ell_\theta = \frac{\partial}{\partial\theta} \log q_\theta(T \mid C, X)$. If, in addition, Conditions (P1)–(P4) *hold and the parametric model* *is correctly specified with true value θ_0, $\hat{\theta}$ attains the semiparametric efficiency bound for θ in $\mathcal{M}$.*

A path $m_s = m(1+sa)$ has score $a(C,X)$, so the nuisance tangent space is $\mathcal{T}_m = \{a(C,X) : \mathbb{E}[a(C,X) \mid S=1] = 0\}$. Differentiability in quadratic mean gives

$$\mathbb{E}\{\dot{\ell}_q(T,C,X) \mid C,X,S=1\} = 0,$$

and hence $\dot{\ell}_q$ is orthogonal to $\mathcal{T}_m$. Its projection onto $\mathcal{T}_m$ is zero. Under Conditions (P1)–(P4) and correct specification, Theorem 3 shows that the influence function is $I_q^{-1}\dot{\ell}_q$, so $\hat{\theta}$ attains the bound. $\square$

S2.3 Conditions and proof of Theorem 4

In the single-cohort frame, let $T^S := ST$ be the recorded event-time variable, with $T^S = 0$ when $S=0$, and reparametrize the observable distribution of $O = (S, T^S, C, X)$ as

$$\begin{aligned} p_{\theta,\mu,r}(o) &= r(c,x)\left[\mu(c,x)\, q_\theta(t \mid c,x)\right]^s \left[1-\mu(c,x)\right]^{1-s}, \\ \mu(c,x) &= \mathbb{P}(S=1 \mid C=c, X=x) = \rho(c,x)\, F_\theta(c \mid x), \end{aligned}$$

where $r(c,x)$ is the marginal density of (C,X). The conditions of Theorem 4 are as follows.

(E1) q_θ is differentiable in quadratic mean at θ_0 with score $\dot{\ell}_q(T,C,X) = \partial_\theta \log q_{\theta_0}(T \mid C,X)$ and nonsingular information $I_q = \mathbb{E}[\dot{\ell}_q \dot{\ell}_q' \mid S=1]$;

(E2) the labelling mechanism is locally free with uniform relative slack: for almost all (c,x) under $P_{C,X}$,

$$0 < \mu_0(c,x) < 1 \qquad \text{and} \qquad \rho_0(c,x) = \mu_0(c,x)/F_{\theta_0}(c \mid x) \le 1-\delta$$

for some $\delta > 0$ not depending on (c,x) (equivalently, $\mu_0 \le (1-\delta)F_{\theta_0}$ almost everywhere); moreover $\theta \mapsto F_\theta(c \mid x)$ admits a local Lipschitz envelope at θ_0: $|F_\theta(c \mid x) - F_{\theta_0}(c \mid x)| \le \|\theta - \theta_0\|\, \dot{F}(c,x)$ for θ near θ_0, with $\mathbb{E}[\dot{F}(C,X)^2/\{\mu_0(1-\mu_0)\}] < \infty$. For $\varepsilon > 0$ write

$$\mathcal{C}_\varepsilon = \{(c,x) : \varepsilon \le \mu_0(c,x) \le 1-\varepsilon,\ \dot{F}(c,x) \le 1/\varepsilon\};$$

the feasible nuisance directions are the bounded b supported on some $\mathcal{C}_\varepsilon$; with θ fixed at θ_0, the path $\mu_\eta = \mu_0 + \eta b$ remains a valid conditional labelling probability for sufficiently small $|\eta|$. Since $0 < \mu_0 < 1$ and $\dot{F} < \infty$ almost everywhere, $\mathbb{P}\left(\bigcup_{\varepsilon>0} \mathcal{C}_\varepsilon\right) = 1$ and these directions are dense in $L_2(P_{C,X})$. Complete labelling, $\rho_0 \equiv 1$, violates the relative-slack requirement;

(E3) $r(c,x)$ is unrestricted.

We use the tangent-space projection argument of van der Vaart (1991, Section 6, pp. 188–189). Under the (θ,μ,r) parameterization in Section S2.3, the score for θ is $\dot\ell_\theta = S\dot\ell_q$. Differentiating the Bernoulli contribution along $\mu_\eta = \mu_0 + \eta b$ gives

$$\dot\ell_\mu[b] = \frac{\{S-\mu_0(C,X)\}b(C,X)}{\mu_0(C,X)\{1-\mu_0(C,X)\}}.$$

On each $\mathcal{C}_\varepsilon$ the denominator is bounded away from zero and its reciprocal can be absorbed into the direction. Condition (E2) makes these bounded, supported directions feasible and dense in $L_2(P_{C,X})$. Thus the nuisance tangent spaces are

$$\mathcal{T}_\mu = \mathrm{cl}\Big\{\{S-\mu_0(C,X)\}b(C,X) : b \in L_2(P_{C,X})\Big\}, \qquad \mathcal{T}_r = \{a(C,X) : \mathbb{E}a(C,X) = 0\}.$$

For a bounded b supported on $\mathcal{C}_\varepsilon$, Condition (E2) bounds μ_0 and $1-\mu_0$ away from zero, while $F_{\theta_0} - \mu_0 \geq \delta F_{\theta_0} \geq \delta\varepsilon$. Hence $\mu_0 + \eta b \in (0, F_{\theta_0})$ for sufficiently small $|\eta|$.

A path in θ need not preserve $\mu_0 < F_\theta$ uniformly. For a differentiable path θ_η through θ_0, set

$$\widetilde\mu_\eta = \min\{\mu_0, (1-\delta/2)F_{\theta_\eta}\}, \qquad A_\eta = \{(1-\delta/2)F_{\theta_\eta} < \mu_0\}.$$

This path is feasible and $\widetilde\mu_0 = \mu_0$. Moreover, $\mathbf{1}\{A_\eta\} \to 0$ almost surely and, on A_η, $|\widetilde\mu_\eta - \mu_0| \leq K|\eta|\dot F$ for some $K < \infty$. The integrability requirement in (E2) and dominated convergence therefore give

$$\int \frac{(\widetilde\mu_\eta - \mu_0)^2}{\mu_0(1-\mu_0)}\, dP_{C,X} \leq K^2\eta^2\mathbb{E}\left[\frac{\mathbf{1}\{A_\eta\}\dot F(C,X)^2}{\mu_0(1-\mu_0)}\right] = o(\eta^2).$$

The squared Hellinger distance between the corresponding Bernoulli laws is bounded by this integral, because $(\sqrt{a}-\sqrt{b})^2 \leq (a-b)^2/b$ for $a, b > 0$, applied to both outcomes. The Bernoulli contribution is thus differentiable in quadratic mean with zero score. Together with (E1), this yields a feasible observed-data path with score $S\dot\ell_q$.

Since $\mathbb{E}(\dot\ell_q \mid C,X,S=1) = 0$, for $a \in \mathcal{T}_r$ and any b defining a μ-direction,

$$\mathbb{E}(S\dot\ell_q a) = 0, \qquad \mathbb{E}[S\dot\ell_q\{S-\mu_0(C,X)\}b] = \mathbb{E}[\mu_0(1-\mu_0)b\mathbb{E}(\dot\ell_q \mid C,X,S=1)] = 0.$$

Thus $S\dot\ell_q$ is orthogonal to $\mathcal{T}_r \oplus \mathcal{T}_\mu$ and is the efficient score, with information

$$I_{\text{full}} = \mathbb{E}[S\dot\ell_q\dot\ell_q'] = p_S I_q.$$

Under the additional conditions in the theorem, Theorem 3 shows that the full-sample influence function is $I_{\text{full}}^{-1}S\dot\ell_q$. Le Cam's third lemma along the feasible differentiable submodels above makes this expansion regular, so $\hat\theta$ attains the bound; $n_1/n \to_p p_S$ gives the stated $\sqrt{n_1}$ limit.

Now suppose $\rho_\alpha(c,x)$ is regular, write $\mu_{\theta,\alpha} = \rho_\alpha F_\theta$, and let

$$s_\theta^S = \frac{S - \mu_{\theta,\alpha}}{\mu_{\theta,\alpha}(1-\mu_{\theta,\alpha})} \partial_\theta \mu_{\theta,\alpha}.$$

The same conditioning argument makes $S\dot{\ell}_q$ orthogonal to the labelling-indicator scores. If $\mathcal{T}_\alpha$ is the labelling-nuisance tangent space, the efficient score is

$$S\dot{\ell}_q + s_\theta^S - \Pi(s_\theta^S \mid \mathcal{T}_\alpha),$$

with information

$$p_S I_q + \mathrm{Var}\{s_\theta^S - \Pi(s_\theta^S \mid \mathcal{T}_\alpha)\}.$$

This gives the cases stated after the theorem: known labelling, a finite-dimensional labelling model, and complete labelling. □

S3 Statement and proof of Proposition S2

Proposition S2 (Estimation of event-before-censoring probability)**.** *For parts* (a) *and* (b)*, consider the single-cohort frame, in which all n units are i.i.d. from one cohort, $\mathbb{P}_n$ is the empirical average over them, and $n_1/n \to_p p_S \in (0,1)$; let ϕ_i be the influence function of $\sqrt{n_1}(\hat{\theta} - \theta_0)$ in Theorem 3. Suppose the conditions of Theorem 3 hold and that $\theta \mapsto F_\theta(c \mid x)$ is differentiable at θ_0 with $\Gamma = \mathbb{E}[\partial_\theta F_{\theta_0}(C \mid X)]$ finite and locally Lipschitz in θ with a square-integrable Lipschitz constant.*

(a) *If $G_C(\cdot \mid x)$ is known by design, the estimator obtained by analytic integration, $\hat{\pi}^{\text{known}} = \mathbb{P}_n \, \pi_{\hat{\theta}}(X)$ with $\pi_\theta(x) = \int F_\theta(c \mid x)\, dG_C(c \mid x)$, satisfies*

$$\sqrt{n}\,(\hat{\pi}^{\text{known}} - \pi) = \sqrt{n}\,(\mathbb{P}_n - \mathbb{P})\,\pi_{\theta_0}(X) + \Gamma' \sqrt{n}\,(\hat{\theta} - \theta_0) + o_p(1) \xrightarrow{d} N(0, V_{\text{known}}),$$

$$V_{\text{known}} = \mathrm{Var}\{\pi_{\theta_0}(X)\} + p_S^{-1}\, \Gamma' \,\mathrm{Var}(\phi \mid S = 1)\, \Gamma.$$

(b) *If G_C is unknown and the integral over censoring times is replaced by the empirical average over the observed pairs (C_i, X_i), the resulting estimator $\hat{\pi}^{\text{est}} = \mathbb{P}_n \, F_{\hat{\theta}}(C \mid X)$ satisfies*

$$\sqrt{n}\,(\hat{\pi}^{\text{est}} - \pi) = \sqrt{n}\,(\mathbb{P}_n - \mathbb{P})\, F_{\theta_0}(C \mid X) + \Gamma' \sqrt{n}\,(\hat{\theta} - \theta_0) + o_p(1) \xrightarrow{d} N(0, V_{\text{est}}),$$

$$V_{\text{est}} = V_{\text{known}} + \mathbb{E}[\mathrm{Var}\{F_{\theta_0}(C \mid X) \mid X\}] \geq V_{\text{known}}.$$

(c) *In the case-population frame, let P_0 be the distribution in the target population, $\mathbb{P}_{n_0}$ be the empirical average over an i.i.d. sample of size n_0 from P_0 that is independent of the* *labelled-positive sample of size n_1, and $\hat\theta$ be estimated from that labelled-positive sample. The corresponding estimators satisfy*

$$\hat\pi^{\text{known}} = \mathbb{P}_{n_0}\,\pi_{\hat\theta}(X) \ \to_p \ \pi, \qquad \hat\pi^{\text{est}} = \mathbb{P}_{n_0}\,F_{\hat\theta}(C \mid X) \ \to_p \ \pi$$

as $n_1, n_0 \to \infty$.

Write both estimators as $\hat\pi = \mathbb{P}_n m_{\hat\theta}$, with $m_\theta(C,X) = \pi_\theta(X)$ when G_C is known and $m_\theta(C,X) = F_\theta(C \mid X)$ when it is estimated. The local Lipschitz requirement gives stochastic equicontinuity and $L_2(\mathbb{P})$ continuity (van der Vaart, 1998, Example 19.7 and Lemma 19.24); in either case $\mathbb{P}m_\theta$ has derivative Γ' at θ_0. Hence

$$\sqrt{n}(\hat\pi - \pi) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})m_{\theta_0} + \Gamma'\sqrt{n}(\hat\theta - \theta_0) + o_p(1).$$

The M-estimation expansion gives $\phi = A^{-1}\psi(\theta_0)$ and hence $\mathbb{E}(\phi \mid C, X, S = 1) = 0$. The full-sample influence function for $\hat\theta$ is $p_S^{-1}S\phi$, and

$$\mathbb{E}[(m_{\theta_0} - \mathbb{P}m_{\theta_0})S\phi] = \mathbb{E}[(m_{\theta_0} - \mathbb{P}m_{\theta_0})\mu_0\mathbb{E}(\phi \mid C, X, S = 1)] = 0.$$

Thus the two terms in the expansion are uncorrelated and

$$V_{\text{known}} = \operatorname{Var}\{\pi_{\theta_0}(X)\} + p_S^{-1}\,\Gamma'\operatorname{Var}(\phi \mid S = 1)\,\Gamma, \quad V_{\text{est}} = V_{\text{known}} + \mathbb{E}[\operatorname{Var}\{F_{\theta_0}(C \mid X) \mid X\}],$$

where the second identity follows from the law of total variance.

For part (c), use the same notation m_θ as above, with $\mathbb{P}_{n_0}$ in place of $\mathbb{P}_n$. Both choices of m_θ take values in $[0,1]$, and $\theta \mapsto P_0 m_\theta$ is continuous at θ_0 by dominated convergence. Conditional on the labelled-positive sample, independence gives

$$\mathbb{E}\Big[\{\mathbb{P}_{n_0}m_{\hat\theta} - P_0 m_{\hat\theta}\}^2 \,\Big|\, \hat\theta\Big] \le \frac{1}{4n_0},$$

so $\mathbb{P}_{n_0}m_{\hat\theta} - P_0 m_{\hat\theta} = o_p(1)$. Theorem 3 gives $\hat\theta \to_p \theta_0$, and hence $P_0 m_{\hat\theta} \to_p P_0 m_{\theta_0} = \pi$. This proves the asserted consistency. □

S4 Semiparametric proportional-hazards theory

S4.1 Statement and proof of Lemma S1

Lemma S1 (Administrative capping). *Let τ_C be the upper endpoint of the support of C. Fix $c_{\mathrm{L}} < \tau < \tau_C$, set $D = C \wedge \tau$, and define $S^\tau = S\,\mathbf{1}\{T \le \tau\}\,\mathbf{1}\{C \ge c_{\mathrm{L}}\}$. Suppose that $\mathbb{P}(S^\tau = 1) > 0$. Under Assumptions 1–2, for $P_{D,X|S^\tau=1}$-almost every (d,x) and Lebesgue-almost* *every* $t \in (0, d)$,

$$f_{T|D,X,S^\tau=1}(t \mid d, x) = \frac{f_T(t \mid x)}{F_T(d \mid x)}, \qquad c_{\mathrm{L}} \le d \le \tau.$$

Proof. Because D is a function of C, Assumption 1 gives $T \perp D \mid X$ and hence $\mathbb{P}(T \le D \mid D, X) = F_T(D \mid X)$. Let $\mathcal{A} = \{F_T(D \mid X) = 0\}$. Since $S^\tau = 1$ implies $T \le D$,

$$\mathbb{P}(\mathcal{A}, S^\tau = 1) \le \mathbb{P}(\mathcal{A}, T \le D) = \mathbb{E}[\mathbf{1}\{\mathcal{A}\}\mathbb{P}(T \le D \mid D, X)] = \mathbb{E}[\mathbf{1}\{\mathcal{A}\}F_T(D \mid X)] = 0.$$

Thus $F_T(D \mid X) > 0$ almost surely conditional on $S^\tau = 1$.

For $c_{\mathrm{L}} \le d < \tau$, $D = d$ implies $C = d$. Because $S = 1$ implies $T \le C = d < \tau$, the additional restriction $T \le \tau$ is automatic. The restriction of $P_{D,X|S^\tau=1}$ to $\{D < \tau\}$ is therefore absolutely continuous with respect to $P_{C,X|S=1}$, so the asserted almost-everywhere identity follows from Theorem 1.

At the atom $d = \tau$, whenever it is present, for $P_{X|D=\tau,S^\tau=1}$-almost every x,

$$\mathbb{P}(T \in dt, D = \tau, S^\tau = 1 \mid X = x) = f_T(t \mid x)\mathbf{1}\{t \le \tau\}\, dt \int_{c \ge \tau} \rho(c, x)\, dG_C(c \mid x).$$

The factor involving ρ is free of t and cancels after normalization, giving $f_T(t \mid x)/F_T(\tau \mid x)$ for Lebesgue-almost every $t \in (0, \tau)$. □

S4.2 Proofs for semiparametric sieve estimator

S4.2.1 Construction and regularity conditions

Construction. For a candidate log-hazard η, write $H_\eta(s) = \int_0^s e^{\eta(u)}\, du$, so that $H_{\eta_0} = H_0$ at the truth. By Theorem 1, the conditional log-likelihood contribution of a labelled positive observation with $0 < t < c$ is

$$\ell(\beta, \eta; t, c, x) = x'\beta + \eta(t) - e^{x'\beta}H_\eta(t) - \log\Big[1 - \exp\{-e^{x'\beta}H_\eta(c)\}\Big]. \tag{S1}$$

We work in the capped experiment of Lemma S1, with $c_{\mathrm{L}} \in (0, \tau)$. The atom at $D = \tau$ supplies information near the right endpoint of the sieve domain, while the lower threshold avoids values at which the truncation denominator is close to zero; to lighten notation we write C, S, and n_1 for D, S^τ, and $n_{1,\tau} = \sum_i S_i^\tau$. Let $\{B_1, \ldots, B_{J_n}\}$ be a normalized B-spline basis of order r on a quasi-uniform mesh on $[0, \tau]$, and let $\mathcal{H}_n = \Big\{\eta = \sum_{k \le J_n} \xi_k B_k : \ \|\xi\|_\infty \le K\Big\}$ for $\xi = (\xi_1, \ldots, \xi_{J_n})$. Let $\mathcal{B} \subset \mathbb{R}^d$ be compact with $\beta_0 \in \mathrm{int}(\mathcal{B})$. The estimator is

$$(\hat\beta, \hat\eta) = \underset{\beta \in \mathcal{B},\ \eta \in \mathcal{H}_n}{\arg\max}\ \frac{1}{n_1} \sum_{i:S_i=1} \ell(\beta, \eta; T_i, C_i, X_i), \tag{S2}$$

a smooth maximization in $(\beta,\xi) \in \mathbb{R}^{d+J_n}$, since the parameterization ensures positivity of the candidate hazard e^{η}.

Throughout the sieve proofs, $P = \mathcal{L}\{(T,D,X) \mid S^{\tau} = 1\}$, with D, S^{τ} written as C, S, $\|f\|^2 = Pf^2$, and $\mathbb{G}_{n_1}$ the retained-sample empirical process. Conditional on the selection indicators the observations are independent with law P, and, given $n_1 = m$, the sieve design is deterministic. The bounds below are uniform for large m in the fixed capped experiment, but not over weak-identification arrays; $n_1/n \to_p p_{S^{\tau}} \in (0,1)$ transfers the conditional orders. Write $\eta_0 = \log h_0$ and $H_0(t) = \int_0^t e^{\eta_0}$. The coefficient bound makes $\mathcal{H}_n$ compact, while Remark S1 makes it asymptotically inactive and justifies the first-order conditions below.

We use one entropy calculation for both proofs. Put $\|\nu - \nu_0\|_{\mathcal{V}}^2 = \|\beta - \beta_0\|^2 + \|\eta - \eta_0\|_{L_2(du)}^2$ and, for fixed uniformly bounded spline directions a_n, $Z_{\nu,a_n} = \dot{\ell}_{\nu,\beta} - \dot{\ell}_{\nu,\eta}[a_n]$. Let $\mathcal{M}_{n,\delta}$ collect $m_{\nu} - m_{\nu_0}$ and $\mathcal{S}_{n,\delta}(a_n)$ collect the coordinate differences $e_j'\{Z_{\nu,a_n} - Z_{\nu_0,a_n}\}$, in both cases over $\nu \in \mathcal{B} \times \mathcal{H}_n$ with $\|\nu - \nu_0\|_{\mathcal{V}} \le \delta$. Bounded X, $C \ge c_{\mathrm{L}}$ and the coefficient box give uniform envelopes. Although point evaluation is not $L_2(du)$-bounded, (S3b) controls $\eta(T)$ in $L_2(P)$ and the two H_{η} terms are L_2-Lipschitz. Local support and normalized B-spline norm equivalence give, for either class $\mathcal{C}_{n,\delta}$,

$$
\begin{aligned}
\sup_{f \in \mathcal{C}_{n,\delta}} \|f\|_{P,2} &\lesssim \delta, \qquad \log N_{[\,]}(\epsilon, \mathcal{C}_{n,\delta}, L_2(P)) \lesssim J_n \log(eC\delta/\epsilon), \\
J_{[\,]}(C\delta, \mathcal{C}_{n,\delta}, L_2(P)) &\lesssim \int_0^{C\delta} \sqrt{1 + J_n \log(eC\delta/\epsilon)}\, d\epsilon \lesssim \delta\sqrt{J_n}, \\
\mathbb{E}^* \sup_{f \in \mathcal{C}_{n,\delta}} |\mathbb{G}_{n_1} f| &\lesssim \delta\sqrt{J_n} + J_n/\sqrt{n_1}, \qquad 0 < \epsilon \le C\delta.
\end{aligned}
$$

The first two lines are the model-specific conditional likelihood calculation, with localized coefficient entropy as in Shen and Wong (1994, Example 3, Case 3, pp. 596–597); equation (3.4.4) of van der Vaart and Wellner (2023, pp. 432–433) gives the last line. At fixed radius it also gives $\mathbb{E}^* \sup |(\mathbb{P}_{n_1} - P)m_{\nu}| \lesssim \sqrt{J_n/n_1} + J_n/n_1$.

Throughout, write $\psi = x'\beta$, $A = e^{\psi} H_{\eta}(c)$, $B_t = e^{\psi} H_{\eta}(t)$, and the truncation weight $w(A) = Ae^{-A}/(1 - e^{-A})$. At the truth $\nu_0 = (\beta_0, \eta_0)$ these reduce to $A = e^{x'\beta_0} H_0(c)$ and $B_t = e^{x'\beta_0} H_0(t)$, the evaluations used wherever the efficient score, the information, or a least-favourable direction is formed. Differentiating (S1) gives the scores

$$
\dot{\ell}_{\beta} = x\Big(1 - B_t - w(A)\Big), \quad \dot{\ell}_{\eta}[g] = g(t) - e^{\psi} G(t) - e^{\psi} G(c)\, \frac{e^{-A}}{1 - e^{-A}}, \tag{S3}
$$

for a perturbation direction g, where $G(s) = \int_0^s e^{\eta(u)} g(u)\, du$; the efficient score and information are defined in the usual way from the nuisance tangent space.

Spline approximation. For $q > 0$, let $m = \lceil q \rceil - 1$ and $\alpha = q - m \in (0,1]$. A Hölder ball of smoothness q on $[0,\tau]$ means that, for a fixed $L<\infty$, each member $g\in C^m[0,\tau]$ satisfies

$$\max_{0\le \ell\le m}\|D^\ell g\|_\infty+\sup_{u\neq v}\frac{|D^m g(u)-D^m g(v)|}{|u-v|^\alpha}\le L.$$

Thus, at an integer q, $D^{q-1}g$ is Lipschitz. We take each endpoint of the spline mesh to be repeated r times and write h_n for its maximum knot spacing; quasi-uniformity and fixed order give $h_n\asymp J_n^{-1}$. If $0<q\le r$ and $s_{n,g}$ is an L_∞-best spline approximant to g in $\operatorname{span}\{B_1,\dots,B_{J_n}\}$, then

$$\|g-s_{n,g}\|_\infty=\operatorname{dist}_{L_\infty}\Big(g,\operatorname{span}\{B_1,\dots,B_{J_n}\}\Big)\lesssim h_n^m\omega(D^m g;h_n)\lesssim J_n^{-q}$$

by de Boor (2001, Chapter XII, (6), pp. 148–149), where ω is the modulus of continuity. Hence $\|g-s_{n,g}\|_{L_2([0,\tau])}\lesssim J_n^{-q}$ and $\|s_{n,g}\|_\infty\le\|g\|_\infty+O(J_n^{-q})$. Finally, if $s=\sum_k\xi_kB_k$, then

$$D_r^{-1}\|\xi\|_\infty\le\|s\|_{L_\infty[0,\tau]}\le\|\xi\|_\infty,$$

where D_r depends only on the order: the first inequality is de Boor (2001, Chapter XI, (8), p. 133), and the second follows from nonnegativity and partition of unity. The first inequality does not require quasi-uniformity.

Conditions for a fixed capped experiment.

(S1) *Model and design.* The parameter space $\mathcal{B}$ is as defined above; X is bounded with $\mathbb{E}[XX'\mid S=1]$ nonsingular.

(S2) η_0 lies in a Hölder ball of smoothness $p\ge 1$ on $[0,\tau]$ and h_0 is bounded above and below away from 0 on $[0,\tau]$.

(S3a) *Identification in the capped experiment.* There is a measurable covariate subset $\mathcal{X}_\tau$ with $\mathbb{P}(X\in\mathcal{X}_\tau\mid S^\tau=1)>0$ on which $\operatorname{Var}(X\mid X\in\mathcal{X}_\tau,S^\tau=1)$ is nonsingular, $x'\beta_0$ is nonconstant, and $\mathbb{P}(D=\tau\mid X=x,S^\tau=1)>0$ for every $x\in\mathcal{X}_\tau$. A common event-time interval $[s_1,s_2]\subset(0,\tau)$ lies in the conditional support of T given $(D=\tau,X=x,S^\tau=1)$ for all $x\in\mathcal{X}_\tau$; write $t_0:=s_1$ for its left endpoint.

(S3b) The marginal density p_+ of $T\mid S^\tau=1$ satisfies $0<\underline{p}\le p_+(t)\le\overline{p}<\infty$ on $[0,\tau]$.

(S4) The B-spline order r, in the order convention of de Boor, 2001 (order = degree +1), satisfies $r\ge p$. Also, $J_n\to\infty$ and $J_n^2/n_1\to 0$. The additional choice of J_n is stated in each theorem that uses it.

(S5) The least-favourable direction a^* exists in the closure of the tangent set and $\mathcal{I}(\beta_0)$ is nonsingular.

(S6) Each coordinate of a^* belongs to a Hölder ball of smoothness $s_* > 1/2$ on $[0, \tau]$; for results using this condition, the order in (S4) also satisfies $r \geq s_*$.

Conditions (S5)–(S6) are used only for efficient inference on β. Under (S1)–(S2) and $C \geq c_{\mathrm{L}}$, Condition (S3b) holds, for example, if $\mathbb{P}(D = \tau \mid S^\tau = 1) \geq \kappa > 0$, because this atom contributes $\kappa \inf_x q(t \mid \tau, x) > 0$ to $p_+(t)$ uniformly on $[0, \tau)$.

Under (S6), let $a^*_{j,n,\infty}$ be an L_∞-best spline approximant to coordinate a^*_j and put $a^*_{n,\infty} = (a^*_{1,n,\infty}, \ldots, a^*_{d,n,\infty})'$. The spline approximation above gives

$$\max_j \|a^*_{j,n,\infty} - a^*_j\|_\infty = O(J_n^{-s_*}),$$
$$\max_j \|a^*_{j,n,\infty} - a^*_j\|_{L_2([0,\tau])} = O(J_n^{-s_*}),$$
$$\sup_n \max_j \|a^*_{j,n,\infty}\|_\infty < \infty.$$

These approximants provide the uniformly bounded spline directions used in Lemma S4 and the proof of Theorem 6.

Remark S1 (Coefficient bound is slack)**.** *The coefficient bound is not a smoothing penalty; it is needed only for the asymptotic theory. Put* $M_0 = \|\eta_0\|_{L_\infty[0,\tau]}$ *and fix* $K = D_r(M_0 + 1)$*. By* (S2), (S4)*, and the spline approximation above, an* L_∞*-best spline approximant* $\eta^*_n = \sum_k \xi^*_{k,n} B_k$ *satisfies*

$$\|\eta^*_n - \eta_0\|_\infty = O(J_n^{-p}), \qquad \|\xi^*_n\|_\infty \leq D_r \|\eta^*_n\|_\infty \leq D_r\{M_0 + O(J_n^{-p})\} < K$$

for all sufficiently large n*, and hence* $\eta^*_n \in \mathcal{H}_n$*. The uniform consistency in Theorem 5 and the same coefficient stability give*

$$\|\hat{\xi}\|_\infty \leq D_r \|\hat{\eta}\|_\infty \leq D_r\{M_0 + o_p(1)\} < K$$

with probability tending to one. Thus the coefficient constraint is inactive at $(\hat{\beta}, \hat{\eta})$ *with probability tending to one, and the interior first-order conditions used in Theorems 6 and S2 hold without a Lagrange multiplier. The bound serves only to make* $\mathcal{H}_n$ *a compact sieve.*

S4.2.2 Main statements

Theorems 5 and 6 are stated in Section 4.4. The additional results are as follows.

Theorem S2 (Root-n inference for regular functionals)**.** *Let* $\nu = (\beta, \eta)$ *and* $F_{\beta,\eta}(s \mid x) = 1 - \exp\{-e^{x'\beta} H_\eta(s)\}$*. Suppose that* $p > 1$ *and* (S1), (S2), (S3a)–(S3b)*, and* (S4)–(S6) *hold, with* $\sqrt{n_1} J_n^{-p} \to 0$*. Fix* $t \in (0, \tau]$ *and* $x \in \operatorname{supp}(X \mid S^\tau = 1)$*. We consider the functionals* $H_\eta(t)$*,* $F_{\beta,\eta}(t \mid x)$*,* $\int_0^\tau w(s)\, dF_{\beta,\eta}(s \mid x)$ *for* $w \in L_\infty([0, \tau])$*, and* $\int F_{\beta,\eta}(c \wedge \tau \mid x)\, dG_C(c \mid x)$*. At*

ν_0 these are $H_0(t)$, $F_T(t \mid x)$, $\int_0^\tau w(s)\,dF_T(s \mid x)$, and $\pi_\tau(x)$, respectively; for the last functional, $G_C(\cdot \mid x)$ is known or fixed by design.

For a real-valued functional χ, write $\langle a, b\rangle_{\mathcal{I}} = P\{\dot\ell_{\nu_0}[a]\dot\ell_{\nu_0}[b]\}$, $\|a\|_{\mathcal{I}}^2 = \langle a, a\rangle_{\mathcal{I}}$, and $\mathcal{V}_n = \mathbb{R}^d \times \operatorname{span}\{B_1, \dots, B_{J_n}\}$. Assume the following conditions.

(R1) *χ is pathwise differentiable at ν_0, with derivative bounded with respect to $\|\cdot\|_{\mathcal{I}}$, and hence has a joint Riesz representer a^χ satisfying $\dot\chi_{\nu_0}[a] = \langle a^\chi, a\rangle_{\mathcal{I}}$; put $\widetilde{\ell}_\chi = \dot\ell_{\nu_0}[a^\chi]$.*

(R2) *Put $\delta_J = J_n^{-p} + \sqrt{J_n/n_1}$. There are r_n and ϵ_n such that $\delta_J = o(r_n)$, $\sqrt{n_1} r_n^2 \to 0$, and $\epsilon_n \to 0$. For $\mathcal{N}_n = \{\nu \in \mathcal{B}\times\mathcal{H}_n : \|\nu-\nu_0\|_{\mathcal{I}} \le r_n,\ \|\eta-\eta_0\|_\infty \le \epsilon_n\}$, one has $\mathbb{P}(\hat\nu \in \mathcal{N}_n) \to 1$ and, uniformly for $\nu, \bar\nu \in \mathcal{N}_n$,*

$$\Big|\chi(\nu) - \chi(\bar\nu) - \dot\chi_{\bar\nu}[\nu - \bar\nu]\Big| + \Big\|\ell_\nu - \ell_{\bar\nu} - \dot\ell_{\bar\nu}[\nu - \bar\nu]\Big\|_{L_2(P)} \lesssim \|\nu - \bar\nu\|_{\mathcal{I}}^2.$$

(R3) *Let $\nu_{0,n}$ maximize $P\ell_\nu$ on $\mathcal{B} \times \mathcal{H}_n$ and let $a_n^\chi \in \mathcal{V}_n$ satisfy*

$$\mathbb{A}_{0,n}(a, b) = -P\ddot\ell_{\nu_{0,n}}[a, b], \qquad \mathbb{A}_{0,n}(a_n^\chi, b_n) = \dot\chi_{\nu_{0,n}}[b_n], \qquad b_n \in \mathcal{V}_n.$$

Require

$$\begin{aligned}
&\sqrt{n_1}|\chi(\nu_{0,n}) - \chi(\nu_0)| \to 0, \qquad\qquad \|\dot\ell_{\nu_0}[a_n^\chi - a^\chi]\|_{L_2(P)} \to 0,\\
&\sup_{\nu\in\mathcal{N}_n} \sqrt{n_1}\Big|P\{\dot\ell_\nu[a_n^\chi] - \dot\ell_{\nu_{0,n}}[a_n^\chi]\} + \mathbb{A}_{0,n}(a_n^\chi, \nu - \nu_{0,n})\Big| = o(1),\\
&\sup_{\nu\in\mathcal{N}_n} \Big|\sqrt{n_1}(\mathbb{P}_{n_1} - P)\{\dot\ell_\nu[a_n^\chi] - \dot\ell_{\nu_0}[a_n^\chi]\}\Big| = o_p(1).
\end{aligned}$$

Then

$$\sqrt{n_1}\{\chi(\hat\nu) - \chi(\nu_0)\} = \frac{1}{\sqrt{n_1}} \sum_{i:S_i=1} \widetilde{\ell}_\chi(T_i, C_i, X_i) + o_p(1) \xrightarrow{d} N\Big(0, P\widetilde{\ell}_\chi^2\Big).$$

For the four displayed functionals, (R1)–(R3) *follow from the stated conditions and undersmoothing. The conclusion is pointwise in (t, x) and does not assert weak convergence of the entire estimated curve. Nonparametric estimation of $G_C(\cdot \mid x)$ at a fixed continuous x and point evaluation of h_0 are not covered.*

Proposition S3 (Variance estimation for β)**.** *Let $\hat{\mathbb{I}}$ be the negative per-observation Hessian of* (S2) *in (β, ξ) at the optimum. Under Theorem 6 and Lemma S4,*

$$[\hat{\mathbb{I}}^{-1}]_{\beta\beta} \xrightarrow{p} \mathcal{I}(\beta_0)^{-1}.$$

Thus $n_1^{-1}[\hat{\mathbb{I}}^{-1}]_{\beta\beta}$ provides the covariance used for Wald inference on $\hat\beta$.

S4.2.3 Identification lemmas and proofs of main statements

Lemma S2 (Coercivity, curvature, and identification on fixed analysis horizon)**.** *Let P_τ be the law of $(T, D, X) \mid S^\tau = 1$ with $D = C \wedge \tau \in [c_{\mathrm{L}}, \tau]$. Under* (S1)*,* (S2)*,* (S3a)*–*(S3b)*: (a) the score operator $\mathcal{A}(b, g) = b'\dot{\ell}_\beta + \dot{\ell}_\eta[g]$ extends to an injective bounded linear map $\mathbb{R}^d \times L_2([0, \tau], du) \to L_2(P_\tau)$, and the information form is coercive,*

$$\mathbb{E}\Big\{\mathcal{A}(b, g)^2 \mid S^\tau = 1\Big\} \;\geq\; c_0 \left(\|b\|^2 + \|g\|^2_{L_2(du)}\right) \qquad \textit{for all } b \in \mathbb{R}^d \textit{ and } g \textit{ in the tangent set,}$$

for some $c_0 > 0$; (b) let K be the fixed coefficient bound in the sieve construction, taken so that $\|\eta_0\|_\infty < K$ as in Remark S1. There are constants $\epsilon_0, c > 0$, independent of the sieve dimension, such that the population criterion $M(\beta, \eta) = P_\tau m_{\beta,\eta}$ satisfies

$$M(\beta_0, \eta_0) - M(\beta, \eta) \;\geq\; c\left\{\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(du)}\right\}$$

for every $\beta \in \mathcal{B}$ and every measurable η with $\|\eta\|_\infty \leq K$ and $\|\beta - \beta_0\|^2 + \|\eta - \eta_0\|^2_{L_2(du)} \leq \epsilon_0^2$. In particular, the same local quadratic margin holds uniformly over every coefficient-bounded sieve $\mathcal{B} \times \mathcal{H}_n$; the rate step does not require the sieve estimator itself to remain in the fixed Hölder ball of (S2)*.*

Proof. Step 1: injectivity of the score operator. Suppose $b'\dot{\ell}_\beta + \dot{\ell}_\eta[g] = 0$ a.s. Define, for $(b, g) \in \mathbb{R}^d \times L_2([0, \tau], du)$, the linear perturbations $\delta F_T(t \mid x) = \{1 - F_T(t \mid x)\}\, e^{x'\beta_0}\{x'b\, H_0(t) + G(t)\}$ with $G(s) = \int_0^s e^{\eta_0} g$, finite and continuous by the Cauchy–Schwarz inequality, $\delta f_T = \partial_t\, \delta F_T$, $\delta \log f_T = \delta f_T / f_T$, and $\delta \log F_T = \delta F_T / F_T$. Direct computation from (S3) gives the algebraic identity $b'\dot{\ell}_\beta + \dot{\ell}_\eta[g] = \delta \log f_T(t \mid x) - \delta \log F_T(c \mid x)$ at every (t, c, x) with $0 < t < c$, for every such (b, g); no parameter path is invoked. The hypothesis therefore forces $\delta \log f_T(t \mid x) = \delta \log F_T(c \mid x)$ for a.e. $t < c$. Restrict the identity to the atom $\{D = \tau\}$: by (S3a) the atom has positive conditional probability for every $x \in \mathcal{X}_\tau$, and by (S2) and Lemma S1, the conditional density of T given $(D = \tau, X = x, S^\tau = 1)$ is positive on $(0, \tau)$, so for P_X-almost every $x \in \mathcal{X}_\tau$, $\delta \log f_T(t \mid x) = \delta \log F_T(\tau \mid x) =: \epsilon(x)$ for Lebesgue-almost every $t \in (0, \tau)$; integrating $\delta f_T(t \mid x) = \epsilon(x) f_T(t \mid x)$ gives $\delta F_T(t \mid x) = \epsilon(x) F_T(t \mid x)$ for $0 < t < \tau$. Writing $v = H_0(t) \in (0, H_0(\tau)]$, $\lambda = e^{x'\beta_0}$, and $\Psi(v) = G(H_0^{-1}(v))$,

$$x'b\, v + \Psi(v) = \epsilon(x)\, e^{-x'\beta_0}\left(e^{\lambda v} - 1\right) \quad \text{for all } v \in (0, H_0(\tau)),$$

for P_X-almost every $x \in \mathcal{X}_\tau$. The identity holds for almost every v by the almost-sure hypothesis, and both sides are continuous in v, so it holds for all v in the stated range at each such x. Let $\mathcal{X}_\tau^0 \subseteq \mathcal{X}_\tau$ be a full-measure subset of covariate values at which it holds; discarding a null set changes neither the nonsingular conditional variance nor the nonconstancy of $x'\beta_0$ in (S3a), so we may fix $x^-, x^+ \in \mathcal{X}_\tau^0$ with $\lambda^- = e^{x^{-\prime}\beta_0} \neq \lambda^+ = e^{x^{+\prime}\beta_0}$. For any pair x_1, x_2 with $\lambda_1 \neq \lambda_2$, differencing eliminates Ψ:

$$(x_1 - x_2)'b\,v = c_1\Big(e^{\lambda_1 v} - 1\Big) - c_2\Big(e^{\lambda_2 v} - 1\Big), \qquad c_i = \epsilon(x_i)e^{-x_i'\beta_0},$$

and the functions v, $e^{\lambda_1 v} - 1$, $e^{\lambda_2 v} - 1$ are linearly independent on any interval (distinct positive exponents), so $c_1 = c_2 = 0$ and $(x_1 - x_2)'b = 0$. Applying this to the pair (x^-, x^+), and then for each $x \in \mathcal{X}_\tau^0$ to (x, x^-) when $e^{x'\beta_0} \neq \lambda^-$ and to (x, x^+) otherwise, gives $\epsilon = 0$ and a common value of $x'b$ on $\mathcal{X}_\tau^0$; nonsingular $\mathrm{Var}(X \mid X \in \mathcal{X}_\tau, S^\tau = 1)$ (S3a) then forces $b = 0$. Substitution at either comparison point then gives $\Psi(v) = 0$ on all of $(0, H_0(\tau))$, hence $g = 0$ almost everywhere on $(0, \tau)$ because $e^{\eta_0} > 0$. The nonconstancy of $x'\beta_0$ on $\mathcal{X}_\tau$ required by (S3a) is what supplies the two comparison points; when $x'\beta_0$ is constant on the whole support, injectivity fails at $(0, g^*)$, the flat direction of Proposition 3 (see the end of Step 3).

Step 2: coercivity. Work under P_τ, the law of $(T, D, X) \mid S^\tau = 1$, on the sieve domain $[0, \tau]$. Write $\mathcal{A}(b, g) = b'\dot{\ell}_\beta + \dot{\ell}_\eta[g]$ as a bounded linear map $\mathbb{R}^d \times L_2([0, \tau], du) \to L_2(P_\tau)$; by (S3), $\dot{\ell}_\eta[g] = (Jg)(T) - (Kg)(T, D, X)$, where $Jg = g(T)$ and $Kg = e^\psi G(T) + w(A)\,\overline{G}(D)$ with $\overline{G}(d) = G(d)/H_0(d)$ and $G(s) = \int_0^s e^{\eta_0} g$. The direct part is bounded below: by (S3b),

$$\|Jg\|^2_{L_2(P_\tau)} = \mathbb{E}\{g(T)^2 \mid S^\tau = 1\} = \int_0^\tau g(t)^2\, p_+(t)\, dt \;\geq\; \underline{p}\, \|g\|^2_{L_2(du)},$$

the lower bound $\underline{p} > 0$ being the content of (S3b) itself, with the atom at $D = \tau$ noted after the condition as one sufficient primitive condition. The kernels of K are bounded on the compact rectangle $[0, \tau]^2 \times \mathcal{X}$ (h_0 two-sided bounded, $w(A) \leq 1$, and $H_0(D) \geq H_0(c_\mathrm{L}) > 0$ since $D \geq c_\mathrm{L}$), so K is Hilbert–Schmidt, hence compact. If coercivity fails, take (b_n, g_n) with $\|b_n\|^2 + \|g_n\|^2_{L_2(du)} = 1$ and $\|\mathcal{A}(b_n, g_n)\|_{L_2(P_\tau)} \to 0$; along a subsequence $b_n \to b$ and $g_n \rightharpoonup g$ in $L_2([0, \tau])$, so $Kg_n \to Kg$ and $b_n'\dot{\ell}_\beta \to b'\dot{\ell}_\beta$ strongly, whence $Jg_n = \mathcal{A}(b_n, g_n) - b_n'\dot{\ell}_\beta + Kg_n$ converges strongly. As J is bounded below, $g_n \to g$ strongly, giving $\|b\|^2 + \|g\|^2_{L_2(du)} = 1$ and $\mathcal{A}(b, g) = 0$, contradicting Step 1; the constant $c_0 > 0$ exists for each fixed design under (S1), (S2) and (S3a)–(S3b), although no common positive lower bound need hold along a sequence with increasingly limited follow-up.

Step 3: a uniform local margin on the coefficient-bounded class. Write $q_{\beta,\eta}$ for the conditional density in (S1). Boundedness of X and $\mathcal{B}$, the bound $\|\eta\|_\infty \leq K$, and $D \geq c_\mathrm{L} > 0$ give constants $0 < \underline{q} < \overline{q} < \infty$ such that

$$\underline{q} \leq q_{\beta,\eta}(t \mid d, x) \leq \overline{q}, \qquad 0 < t < d,$$

uniformly over the coefficient-bounded class. Let $z = q_{\beta,\eta}/q_{\beta_0,\eta_0}$ and $r_{\beta,\eta} = \log z$. Conditional normalization gives $\int q_{\beta_0,\eta_0}(z - 1) = 0$. Since z ranges over a fixed compact subset of $(0, \infty)$,

$z - 1 - \log z \geq c_1(\log z)^2$ there for some $c_1 > 0$. Consequently,

$$M(\beta_0, \eta_0) - M(\beta, \eta) \geq c_1 \, \|r_{\beta,\eta}\|^2_{L_2(P_\tau)}. \tag{S4}$$

Suppose the local quadratic margin in part (b) were false. There would then be a sequence (β_j, η_j) in the coefficient-bounded class such that

$$d_j = \left\{\|\beta_j - \beta_0\|^2 + \|\eta_j - \eta_0\|^2_{L_2(du)}\right\}^{1/2} \downarrow 0, \qquad \frac{\|r_{\beta_j,\eta_j}\|_{L_2(P_\tau)}}{d_j} \longrightarrow 0. \tag{S5}$$

Set $b_j = (\beta_j - \beta_0)/d_j$, $v_j = (\eta_j - \eta_0)/d_j$, and $g_j = \eta_j - \eta_0$. Along a subsequence, $b_j \to b$ and $v_j \rightharpoonup v$ in $L_2([0,\tau], du)$, while $\|b_j\|^2 + \|v_j\|^2_{L_2(du)} = 1$. Define $R(s) = (e^s - 1)/s$ for $s \neq 0$ and $R(0) = 1$. The functions g_j are uniformly bounded and converge to zero in L_2, so $R(g_j)$ is uniformly bounded and converges to one in measure. For every fixed $\varphi \in L_2$, $(R(g_j) - 1)\varphi \to 0$ in L_2; hence $R(g_j)v_j \rightharpoonup v$.

Let $H_j(t) = \int_0^t e^{\eta_j(u)}\, du$ and let $Vw(t) = \int_0^t e^{\eta_0(u)} w(u)\, du$. The Volterra operator $V : L_2([0,\tau]) \to C([0,\tau])$ is compact, and

$$\frac{H_j - H_0}{d_j} = V\{R(g_j)v_j\} \;\longrightarrow\; Vv \quad \text{uniformly on } [0,\tau]. \tag{S6}$$

With $\lambda_j(x) = e^{x'\beta_j}$ and $\lambda_0(x) = e^{x'\beta_0}$, boundedness of X also gives, uniformly in (t,x),

$$\frac{\lambda_j(x)H_j(t) - \lambda_0(x)H_0(t)}{d_j} \longrightarrow \lambda_0(x)\{x'b\, H_0(t) + (Vv)(t)\}. \tag{S7}$$

The normalizing argument $\lambda_j(x)H_j(d)$ is bounded away from zero because $d \geq c_{\mathrm{L}}$. Thus $a \mapsto \log(1 - e^{-a})$ is continuously differentiable on a common compact interval containing all these arguments. It follows from (S6)–(S7) that, after division by d_j, every term other than $v_j(t)$ in the exact identity

$$\begin{aligned} r_{\beta_j,\eta_j}(t,d,x) &= g_j(t) + x'(\beta_j - \beta_0) - \{\lambda_j(x)H_j(t) - \lambda_0(x)H_0(t)\} \\ &\quad - \Big[\log\{1 - e^{-\lambda_j(x)H_j(d)}\} - \log\{1 - e^{-\lambda_0(x)H_0(d)}\}\Big]. \end{aligned}$$

converges strongly in $L_2(P_\tau)$ to its pathwise derivative at (β_0, η_0).

Let $Jv = v(T)$. Equations (S5)–(S7) therefore force Jv_j to converge strongly in $L_2(P_\tau)$. Its weak limit is Jv, and the lower bound $\|Jw\|^2_{L_2(P_\tau)} \geq \underline{p}\|w\|^2_{L_2(du)}$ from Step 2 then gives $v_j \to v$ strongly in $L_2(du)$. Hence $\|b\|^2 + \|v\|^2_{L_2(du)} = 1$. Taking limits in the normalized log-density identity gives $\mathcal{A}(b,v) = 0$ in $L_2(P_\tau)$, contradicting the injectivity in Step 1. Combining this uniform local inverse bound with (S4) proves part (b). Finally, as the spread of $x'\beta_0$ shrinks, the information along the direction approaching the flat direction g^* of Proposition 3 tends to zero by continuity of the information form in β_0, so c_0 and the local margin constant tend to zero; at the degenerate point Step 1 fails for $(0, g^*)$. $\square$

Lemma S2 supplies the curvature and joint coercivity used below. Lemma S3 establishes separation of the maximum over the growing sieves.

Lemma S3 (Separation over the coefficient-bounded sieves)**.** *Let $M(\beta, \eta) = P_\tau m_{\beta,\eta}$ be the population capped criterion of Lemma S2, with m the integrand of* (S1) *and P_τ written P in the lightened notation of the construction above. Under* (S1), (S2), (S3a) *and the spline sieve of* (S4)*, for every $\varepsilon > 0$ there are a $\delta > 0$ and an n_ε such that*

$$\sup\Big\{M(\beta, \eta) - M(\beta_0, \eta_0) : \ \beta \in \mathcal{B}, \ \eta \in \mathcal{H}_n, \ \|\beta - \beta_0\| + \|\eta - \eta_0\|_{L_2([0,\tau])} \ge \varepsilon\Big\} \ \le \ -\delta$$

for every $n \ge n_\varepsilon$. The maximum of M is therefore well separated along the growing sieves and not only over the fixed Hölder ball of (S2)*.*

Proof. Since m is a conditional log-likelihood, $M(\beta_0, \eta_0) - M(\beta, \eta) = \mathbb{E}[\mathrm{KL}\{q_{\beta_0,\eta_0}(\cdot \mid D, X), q_{\beta,\eta}(\cdot \mid D, X)\} \mid S^\tau = 1] \ge 0$, so the displayed supremum is nonpositive for every n; what has to be excluded is that it tends to zero. Suppose it does. Then there are an $\varepsilon > 0$, a subsequence of n, and points $(\beta_n, \eta_n) \in \mathcal{B} \times \mathcal{H}_n$ with $\|\beta_n - \beta_0\| + \|\eta_n - \eta_0\|_{L_2([0,\tau])} \ge \varepsilon$ and $M(\beta_n, \eta_n) \to M(\beta_0, \eta_0)$. By Pinsker's inequality and the Cauchy–Schwarz inequality the conditional total variation distances converge in mean, so along a further subsequence $q_{\beta_n,\eta_n}(\cdot \mid c, x) \to q_{\beta_0,\eta_0}(\cdot \mid c, x)$ in $L_1(dt)$ for P_τ-almost every (c, x).

Step 1: compactness of the sieve sequence. Since $\mathcal{B}$ is compact, pass to a subsequence along which $\beta_n \to \beta^\star$. The coefficient box of $\mathcal{H}_n$ and the partition-of-unity property give $\|\eta_n\|_\infty \le K$, so $e^{-K} \le e^{\eta_n} \le e^K$ and the baselines $H_n(t) = \int_0^t e^{\eta_n(u)} du$ are uniformly bounded and equi-Lipschitz on $[0, \tau]$. By the Arzelà–Ascoli theorem, along a further subsequence $H_n \to H^\star$ uniformly on $[0, \tau]$, where $H^\star(0) = 0$ and $e^{-K}(t - s) \le H^\star(t) - H^\star(s) \le e^K(t - s)$ for $s \le t$; in particular $H^\star$ is continuous and strictly increasing, and extended beyond τ it is a proper PH baseline. Only the coefficient bound is used here, so the argument covers every sieve dimension permitted by (S4).

Step 2: identification of the limit. For $0 < t < c$ write $q_{\beta,\eta}(t \mid c, x) = e^{\eta(t)} A_{\beta,H}(t, c, x)$ with $A_{\beta,H}(t, c, x) = e^{x'\beta} \exp\{-e^{x'\beta} H(t)\}/[1 - \exp\{-e^{x'\beta} H(c)\}]$. On the capped domain $\{0 < t < c, \ c \in [c_{\mathrm{L}}, \tau]\}$ the denominators are bounded away from zero, since $H_n(c) \ge e^{-K} c_{\mathrm{L}} > 0$ and X and $\mathcal{B}$ are bounded, so $A_{\beta_n,H_n} \to A_{\beta^\star,H^\star}$ uniformly there and both are bounded above and below away from zero. The $L_1(dt)$ convergence above is convergence of conditional densities on the support $\{0 < t < c\}$, so the ratio identity is used at the atom, where the support is all of $(0, \tau)$: since $\mathbb{P}(D = \tau \mid X = x, S^\tau = 1) > 0$ on $\mathcal{X}_\tau$ by (S3a), the P_τ-full set of convergence points has a $c = \tau$ section carrying P_X-almost every $x \in \mathcal{X}_\tau$, and for every such x, $e^{\eta_n} = q_{\beta_n,\eta_n}(\cdot \mid \tau, x)/A_{\beta_n,H_n}(\cdot, \tau, x)$ converges in $L_1([0, \tau], du)$ to $q_{\beta_0,\eta_0}(\cdot \mid \tau, x)/A_{\beta^\star,H^\star}(\cdot, \tau, x)$.

The left side does not depend on x, so the limit is a function $h^\star$ of t alone with $e^{-K} \le h^\star \le e^K$, and $q_{\beta_0,\eta_0}(t \mid \tau, x) = h^\star(t) A_{\beta^\star,H^\star}(t,\tau,x)$ for almost every $t \in (0,\tau)$, both sides being strictly positive there because h_0 is bounded below by (S2). Passing to the limit in $H_n(t) = \int_0^t e^{\eta_n}$ gives $H^\star(t) = \int_0^t h^\star$, so with $\eta^\star = \log h^\star$ the proper PH pair $(\beta^\star, H^\star)$ generates the same conditional density as (β_0, H_0) at the truncation point τ, for P_X-almost every $x \in \mathcal{X}_\tau$; equality over the rest of the capped support is not asserted at this stage and follows once Step 3 identifies the parameters.

Step 3: contradiction. By (S3a), $\mathrm{Var}(X \mid X \in \mathcal{X}_\tau, S^\tau = 1)$ is nonsingular and $x'\beta_0$ is nonconstant, so it takes two distinct values outside any P_τ-null set, while the atom at $D = \tau$ supplies the common event-time interval of (S3a). Equality of the conditional densities at the truncation point τ (Step 2) implies, after integration in t, $F_T^{\beta^\star,H^\star}(s \mid x)/F_T^{\beta^\star,H^\star}(\tau \mid x) = F_T^{\beta_0,H_0}(s \mid x)/F_T^{\beta_0,H_0}(\tau \mid x)$ for almost every $s \in (0,\tau)$, for P_X-almost every $x \in \mathcal{X}_\tau$; both ratio maps admit continuous versions, extending the equality to all $s \in (0,\tau]$. The two proper PH distributions therefore differ on the capped horizon only by a factor depending on the covariate value, and the analytic argument of Proposition 2, with the two covariate points just supplied, gives $\beta^\star = \beta_0$ and $H^\star = H_0$ on $[0,\tau]$, whence $h^\star = e^{\eta_0}$ almost everywhere. Since the logarithm is Lipschitz on $[e^{-K}, e^K]$, $\eta_n \to \eta_0$ in $L_1([0,\tau])$, and since $\|\eta_n - \eta_0\|_\infty \le 2K$ the same holds in $L_2([0,\tau])$. Together with $\beta_n \to \beta_0$ this contradicts $\|\beta_n - \beta_0\| + \|\eta_n - \eta_0\|_{L_2([0,\tau])} \ge \varepsilon$, and the lemma follows. □

Proof of Theorem 5. Consistency. By Lemma S3 the population criterion $M(\nu) = Pm_\nu$ has a well-separated maximum at $\nu_0 = (\beta_0, \eta_0)$ over the growing sieves. The global form of the common entropy bound gives $\sup_{\nu \in \mathcal{B}\times\mathcal{H}_n} |(\mathbb{P}_{n_1} - P)m_\nu| = O_p\{\sqrt{J_n/n_1} + J_n/n_1\} = o_p(1)$ under (S4). The sieve argmax theorem therefore gives $\|\hat\nu - \nu_0\|_{\mathcal{V}} = o_p(1)$.

For the rate, let $\nu_n^* = (\beta_0, \eta_n^*)$, with η_n^* from Remark S1. Then $\|\nu_n^* - \nu_0\|_{\mathcal{V}} \lesssim J_n^{-p}$ and a population Taylor expansion gives $M(\nu_0) - M(\nu_n^*) \lesssim J_n^{-2p}$. On the neighbourhood reached by consistency, Lemma S2(b), after rescaling $\|\cdot\|_{\mathcal{V}}$, supplies the quadratic margin uniformly over the actual sieve. Let Θ_n^{loc} be its intersection with the sieve and set $\tilde\nu_n = \hat\nu$ there and $\tilde\nu_n = \nu_n^*$ otherwise. For all large n, exact maximization gives $\mathbb{P}_{n_1} m_{\tilde\nu_n} \ge \mathbb{P}_{n_1} m_{\nu_n^*}$, and consistency gives $\mathbb{P}(\tilde\nu_n = \hat\nu) \to 1$. Theorem 3.4.1 of van der Vaart and Wellner (2023, pp. 429–430), which allows ν_0 outside the sieve, therefore applies on Θ_n^{loc} with zero maximization tolerance. Take $\phi_{n_1}(\delta) = C\{\delta\sqrt{J_n} + J_n/\sqrt{n_1}\}$ and $\delta_J = L\{J_n^{-p} + \sqrt{J_n/n_1}\}$. For sufficiently large L, ϕ_{n_1} is increasing, $\phi_{n_1}(\delta)/\delta$ is decreasing, $M(\nu_0) - M(\nu_n^*) \lesssim \delta_J^2$, and $\phi_{n_1}(\delta_J) \le \sqrt{n_1}\delta_J^2$. The theorem gives $\|\hat\nu - \nu_0\|_{\mathcal{V}} = O_p(\delta_J)$, which is $O_p(n_1^{-p/(2p+1)})$ under the choice of J_n in Theorem 5.

Uniform rate for the cumulative functionals. The distribution function depends on η only through the integral $H(t) = \int_0^t e^\eta$. By (S3b) the marginal density p_+ of $T \mid S^\tau = 1$ is bounded below on $[0,\tau]$, so the $L_2(P_\tau)$ norm in the curvature bound dominates the Lebesgue L_2 norm on $[0,\tau]$ up to a constant. Since the elements of the Hölder ball and of the sieve are uniformly bounded, a mean-value expansion and the Cauchy–Schwarz inequality give, uniformly in $t \in [t_0, \tau]$,

$$\left|\hat{H}(t) - H_0(t)\right| \le \int_0^\tau \left|e^{\hat{\eta}} - e^{\eta_0}\right| du \ \lesssim\ \|\hat{\eta} - \eta_0\|_{L_2(du)} \ \lesssim\ \|\hat{\eta} - \eta_0\|_{L_2([0,\tau])},$$

and the map $(\beta, H) \mapsto F_T(t \mid x) = 1 - \exp(-e^{x'\beta}H(t))$ is Lipschitz in $(\beta, H(t))$ uniformly over bounded x and $t \in [t_0, \tau]$. The uniform rate $O_p(n^{-p/(2p+1)})$ for $\hat{F}_T$ follows without passing through $\|\hat{\eta} - \eta_0\|_\infty$; no inverse estimate is used.

Uniform rate for the log-hazard. For $\hat{\eta}$ itself, apply the spline inverse estimate $\|g\|_\infty \lesssim \sqrt{J_n}\,\|g\|_{L_2}$ to the spline difference $\hat{\eta} - \eta_n^*$. The bounds above give

$$\|\hat{\eta} - \eta_0\|_\infty \lesssim \sqrt{J_n}\{\|\hat{\eta} - \eta_0\|_{L_2([0,\tau])} + \|\eta_n^* - \eta_0\|_{L_2([0,\tau])}\} + \|\eta_n^* - \eta_0\|_\infty = O_p\Big(n_1^{-(p-1/2)/(2p+1)}\Big),$$

which is the slower uniform rate stated in the theorem. □

Rates at general sieve dimension. The preceding argument already gives $\delta_J = J_n^{-p} + \sqrt{J_n/n_1}$ for any $J_n \to \infty$ with $J_n^2/n_1 \to 0$. Applying the spline inverse inequality only to $\hat{\eta} - \eta_n^*$ gives

$$\|\hat{\eta} - \eta_0\|_\infty \lesssim \sqrt{J_n}\{\delta_J + J_n^{-p}\} + J_n^{-p} = O_p(\sqrt{J_n}\,\delta_J) = O_p(J_n^{1/2-p} + J_n/\sqrt{n_1}).$$

This is $o_p(1)$ for $p > 1/2$ and $J_n = o(n_1^{1/2})$; the coefficient box is then inactive (Remark S1), and the sup-norm neighbourhood required by Theorem S2 is reached.

Lemma S4 (Stability of information on sieve)**.** *Let $\mathcal{V}_n = \mathbb{R}^d \times \mathrm{span}\{B_1, \ldots, B_{J_n}\}$ be the linear joint sieve tangent space and equip $a = (b, g)$ with*

$$\|a\|_{\mathcal{V}}^2 = \|b\|^2 + \|g\|_{L_2([0,\tau],du)}^2.$$

For a bilinear form $\mathbb{B}$ on $\mathcal{V}_n$, write $\|\mathbb{B}\|_{\mathrm{op},n} = \sup_{\|a\|_{\mathcal{V}} \le 1, \|\bar{a}\|_{\mathcal{V}} \le 1} |\mathbb{B}(a, \bar{a})|$. Define the population information form and empirical Hessian by

$$\mathbb{A}_0(a, \bar{a}) = -P\ddot{\ell}_{\nu_0}[a, \bar{a}] = P\{\dot{\ell}_{\nu_0}[a]\dot{\ell}_{\nu_0}[\bar{a}]\},$$

$$\widehat{\mathbb{A}}_n(a, \bar{a}) = -\mathbb{P}_{n_1}\ddot{\ell}_{\hat{\nu}}[a, \bar{a}].$$

Under (S1), (S2), (S3a)–(S3b), *and* (S4), *for either choice of J_n in Theorem 6,*

$$\|\widehat{\mathbb{A}}_n - \mathbb{A}_0\|_{\mathrm{op},n} = O_p\left(J_n^{1/2-p} + \frac{J_n}{\sqrt{n_1}}\right) = o_p(1). \tag{S8}$$

Consequently, with probability tending to one,

$$\inf_{a \in \mathcal{V}_n : \|a\|_{\mathcal{V}} = 1} \widehat{\mathbb{A}}_n(a, a) \ge c_0/2, \tag{S9}$$

where c_0 is the population coercivity constant of Lemma S2(a). Under (S5)–(S6)*, let $\mathcal{I}_{\beta,n}$ and $\widehat{\mathcal{I}}_{\beta,n}$ be the population and empirical Schur complements after profiling the nuisance directions in* $\operatorname{span}\{B_1,\dots,B_{J_n}\}$*. Then*

$$\mathcal{I}_{\beta,n} \longrightarrow \mathcal{I}(\beta_0), \qquad \widehat{\mathcal{I}}_{\beta,n} - \mathcal{I}_{\beta,n} = o_p(1). \tag{S10}$$

Proof. Normalized spline coordinates. Let $B(u) = (B_1(u),\dots,B_{J_n}(u))'$ and $G_n = \int_0^\tau B(u)B(u)'\,du$. Quasi-uniformity and fixed spline order give the Riesz stability

$$c\,J_n^{-1}\|v\|^2 \le v'G_nv \le C\,J_n^{-1}\|v\|^2,$$

with constants independent of n. Equivalently, the locally supported basis $\varphi_{kn} = J_n^{1/2}B_k$ has a Gram matrix whose eigenvalues are bounded above and below. Thus Euclidean operator calculations in the (β,φ_n) coordinates are equivalent, uniformly in n, to the form norm $\|\cdot\|_{\mathrm{op},n}$.

Empirical operator bound. At ν_0, choose coordinates obtained from the scaled local basis above. Boundedness of X, the two-sided bound on h_0, and $C \ge c_{\mathrm{L}}$ imply that the normalizing argument $e^{X'\beta_0}H_0(C)$ stays in a compact subset of $(0,\infty)$. The first two derivatives of (S1) are therefore bounded combinations of $b'X$, $g(T)$, and the primitive linear and bilinear forms

$$G_g(s) = \int_0^s e^{\eta_0(u)}g(u)\,du, \qquad G_{g\bar g}(s) = \int_0^s e^{\eta_0(u)}g(u)\bar g(u)\,du.$$

The Cauchy–Schwarz inequality bounds the primitive forms by $C\|g\|_{L_2}$ and $C\|g\|_{L_2}\|\bar g\|_{L_2}$. Local support leaves only a fixed number of overlapping spline pairs at each time, so the sum of the second moments of the Hessian entries is $O(J_n^2)$ in the scaled coordinates. Hence

$$\mathbb{E}\left\|(\mathbb{P}_{n_1} - P)\ddot\ell_{\nu_0}\right\|_{\mathrm{F}}^2 \lesssim \frac{J_n^2}{n_1},$$

and the empirical fluctuation is $O_p(J_n/\sqrt{n_1})$ in operator norm.

It remains to replace ν_0 by $\hat\nu$. On the coefficient-bounded set, direct differentiation of (S1) and $|e^u - e^v| \le C|u-v|$ for bounded u,v give, uniformly over unit directions,

$$|\ddot\ell_\nu[a,\bar a] - \ddot\ell_{\nu_0}[a,\bar a]| \le C\{\|\beta-\beta_0\| + \|\eta-\eta_0\|_\infty\}.$$

The rates at general sieve dimension established above give

$$\|\hat\beta - \beta_0\| + \|\hat\eta - \eta_0\|_\infty = O_p\left(J_n^{1/2-p} + \frac{J_n}{\sqrt{n_1}}\right).$$

Combining these bounds proves (S8). Lemma S2(a) gives $\mathbb{A}_0(a,a) \ge c_0\|a\|_{\mathcal{V}}^2$ on every $\mathcal{V}_n$, so (S9) follows directly.

Profiled information. For each coordinate of the β-score, let a_n^* be its information projection on the nuisance score space generated by the spline directions. Orthogonal projection and boundedness of the score operator give

$$\|a_n^* - a^*\|_{\mathcal{I}} \le C \inf_{g_n \in \mathrm{span}(B)^d} \|g_n - a^*\|_{L_2(du)} \le C\|a_{n,\infty}^* - a^*\|_{L_2(du)} = O(J_n^{-s_*})$$

by (S6). The population Schur complement is the covariance of $\dot\ell_\beta - \dot\ell_\eta[a_n^*]$, so the first convergence in (S10) follows. Equations (S8)–(S9) give uniformly bounded inverses of the population and empirical nuisance blocks in the G_n-normalized coordinates. The inverse perturbation identity, applied to these blocks, then gives the second convergence in (S10). Because a Schur complement is invariant under a nonsingular change of nuisance coordinates, this is also the Schur complement computed from the raw (β, ξ) Hessian.

□

Proof of Theorem 6. We use the sieve M-estimation arguments of Shen (1997, Theorem 1) and Chen (2007, Theorem 4.3). *Least-favourable direction.* The score for β and the score operator in the η-direction are $\dot\ell_\beta$ and $\dot\ell_\eta[g]$ from (S3). The least-favourable direction a^* is the minimizer of $P\{(\dot\ell_\beta - \dot\ell_\eta[a])^{\otimes 2}\}$ over a in the closed linear span of the η-tangent set; equivalently a^* solves the normal equation $P\{\dot\ell_\eta[g](\dot\ell_\beta - \dot\ell_\eta[a^*])\} = 0$ for all g, i.e. a^* is the coordinatewise $L_2(P)$ projection of $\dot\ell_\beta$ onto the η-scores. The existence of a^* and nonsingularity of $\mathcal{I}(\beta_0)$ are (S5), and the rate $J_n^{-s_*}$ for its spline approximation follows from (S6); the efficient score is $\tilde\ell_\beta = \dot\ell_\beta - \dot\ell_\eta[a^*]$ with information $\mathcal{I}(\beta_0) = P[\tilde\ell_\beta^{\otimes 2}]$.

Estimating equations. Since $\beta_0 \in \mathrm{int}(\mathcal{B})$, consistency and Remark S1 imply that, with probability tending to one, the sieve estimator satisfies the interior first-order conditions of (S2) in the β-coordinates and along the coordinatewise spline direction $a_{n,\infty}^*$ obtained under (S6), which is uniformly bounded in supremum norm:

$$\mathbb{P}_{n_1}\Big(\dot\ell_{\hat\beta,\hat\eta,\beta} - \dot\ell_{\hat\beta,\hat\eta,\eta}[a_{n,\infty}^*]\Big) = 0.$$

Write $\Psi_n(\beta, \eta) = \mathbb{P}_{n_1}(\dot\ell_{\beta,\eta,\beta} - \dot\ell_{\beta,\eta,\eta}[a_{n,\infty}^*])$ and $\Psi(\beta, \eta) = P(\dot\ell_{\beta,\eta,\beta} - \dot\ell_{\beta,\eta,\eta}[a^*])$.

Population expansion. By the information identity and the projection property of a^*, the derivative of Ψ at ν_0 is $-\mathcal{I}(\beta_0)$ in the β-direction and zero in every nuisance direction. Thus, for $\delta_J = J_n^{-p} + \sqrt{J_n/n_1}$, bounded second derivatives give $\Psi(\hat\nu) = -\mathcal{I}(\beta_0)(\hat\beta - \beta_0) + O_p(\delta_J^2)$. Under the default dimension, $\sqrt{n_1}\delta_J^2 = o(1)$ because $p > 1/2$; this is the quadratic no-bias condition.

Empirical process and direction approximation. Take $a_n = a_{n,\infty}^*$ and $\mathcal{F}_n = \mathcal{S}_{n,C\delta_J}(a_n)$. Theorem 5 implies that $\hat\nu$ lies in the defining neighbourhood with probability tending to one, and a_n is uniformly bounded by (S6), so the common score-class bounds apply. Replacing a^* by $a_{n,\infty}^*$ leaves $\mathbb{G}_{n_1}\dot\ell_{\nu_0,\eta}[a_{n,\infty}^* - a^*] = O_p(J_n^{-s_*})$ and, by differentiability and the mean-zero score identity, $\sqrt{n_1}P\dot\ell_{\hat\nu,\eta}[a_{n,\infty}^* - a^*] = O_p(\sqrt{n_1}J_n^{-s_*}\delta_J)$; both are $o_p(1)$ under (S6) and the rate calculations below. Moreover, $\mathbb{E}^* \sup_{f\in\mathcal{F}_n} |\mathbb{G}_{n_1} f| \lesssim \delta_J\sqrt{J_n} + J_n/\sqrt{n_1} = o(1)$ under the default choice whenever $p > 1/2$, as (S2) ensures. Hence, after absorbing the direction-approximation terms, $\mathbb{G}_{n_1} Z_{\hat{\nu},a_n} = \mathbb{G}_{n_1}\tilde{\ell}_\beta + o_p(1)$.

Asymptotic linearity. The population expansion, the bounds for approximating the least-favourable direction, the empirical-process result, the interior score equation, and $n_1/n \to p_{S^\tau}$ give

$$\sqrt{n_1}(\hat{\beta} - \beta_0) = \mathcal{I}(\beta_0)^{-1}\,\mathbb{G}_{n_1}\tilde{\ell}_\beta + o_p(1) \ \xrightarrow{d}\ N\Big(0, \mathcal{I}(\beta_0)^{-1}\Big),$$

by the conditional i.i.d. central limit theorem and $P[\tilde{\ell}_\beta^{\otimes 2}] = \mathcal{I}(\beta_0)$.

For efficiency with the infinite-dimensional baseline, the same derivative bounds give differentiability in quadratic mean along bounded joint directions (b, g), with score $b'\dot{\ell}_\beta + \dot{\ell}_\eta[g]$; density and continuity extend the tangent closure to the L_2 baseline directions. Its projection gives $\tilde{\ell}_\beta$, while $E(\tilde{\ell}_\beta \mid D, X, S^\tau = 1) = 0$ gives orthogonality to the unrestricted (D, X) tangent space of Theorem S1, and (S5) gives nonsingular information. The Taylor and equicontinuity bounds are stable under bounded $n_1^{-1/2}$ paths in the conditional model; for local (D, X) paths, Le Cam's third lemma adds the zero covariance with their score. Hence the expansion is regular in the semiparametric model for labelled positive observations and attains $\mathcal{I}(\beta_0)^{-1}$.

An undersmoothed sieve. For any admissible J_n, the argument above gives $\|\hat{\nu} - \nu_0\|_{\mathcal{V}} = O_p(\delta_J)$, where $\delta_J = J_n^{-p} + \sqrt{J_n/n_1}$, together with the sup-norm bound before Lemma S4. The β expansion requires exactly

$$\sqrt{n_1}\delta_J^2 \to 0, \qquad \delta_J\sqrt{J_n} + J_n/\sqrt{n_1} \to 0, \qquad \sqrt{n_1}J_n^{-s_*}\delta_J \to 0.$$

Here $\sqrt{n_1}J_n^{-s_*}\delta_J = \sqrt{n_1}J_n^{-(p+s_*)} + J_n^{1/2-s_*}$. For the default choice, $p > 1/2$ and $s_* > 1/2$ make all three quantities vanish. Under the undersmoothing condition in Theorem 6, $\sqrt{n_1}J_n^{-p} \to 0$ and $J_n^2/n_1 \to 0$ give the first two, and the preceding identity with $s_* > 1/2$ gives the third. Hence the same efficient expansion holds on the sieve used by Theorem S2. □

Proof of Theorem S2. Condition on $n_1 = m$ and put $\delta_J = J_n^{-p} + \sqrt{J_n/n_1}$. The proof of Theorem 5, the spline inverse inequality, and undersmoothing give

$$\|\hat{\nu} - \nu_0\|_{\mathcal{V}} = O_p(\delta_J), \qquad \|\hat{\eta} - \eta_0\|_\infty = O_p(\sqrt{J_n}\delta_J), \qquad \sqrt{n_1}\delta_J^2 + \sqrt{J_n}\delta_J \longrightarrow 0.$$

Coercivity in Lemma S2(a) makes the information and product norms equivalent. The radii in (R2) can therefore be chosen as stated, and Remark S1 gives the empirical interior first-order conditions.

For $a=(b,g)$, let $G_g(s)=\int_0^s e^{\eta_0(u)}g(u)\,du$. Direct differentiation gives

$$\begin{aligned}
\dot H_{\eta_0}(t)[a] &= G_g(t),\\
\dot F_{\nu_0}(t\mid x)[a] &= e^{x'\beta_0}\{1-F_T(t\mid x)\}\{H_0(t)x'b+G_g(t)\},\\
\dot\chi_{w,\nu_0}[a] &= \int_0^\tau w(s)f_T(s\mid x)[x'b+g(s)-e^{x'\beta_0}\{H_0(s)x'b+G_g(s)\}]\,ds,\\
\dot\pi_\tau(x)[a] &= \int \dot F_{\nu_0}(c\wedge\tau\mid x)[a]\,dG_C(c\mid x).
\end{aligned}$$

For bounded w, these derivatives are bounded by a constant times $\|b\|+\|g\|_{L_2}$, and hence by a constant times $\|a\|_{\mathcal{I}}$. The Riesz theorem on the completed joint tangent space proves (R1) for all four functionals. On the neighbourhood in (R2), $e^{x'\beta}H_\eta(C)$ remains in a compact subset of $(0,\infty)$. The second derivatives of the exponential and log-likelihood maps are then uniformly bounded, while $\eta(T)$ is linear. The Cauchy–Schwarz inequality bounds the resulting products and primitive integrals by a constant times $\|\nu-\bar\nu\|_{\mathcal{I}}^2$. The same calculation, followed by integration against bounded w or $G_C(\cdot\mid x)$, proves both remainders in (R2). The calculation also applies between ν_0 and the spline approximant introduced below, although ν_0 need not belong to $\mathcal{B}\times\mathcal{H}_n$.

Let η_n^* be the spline approximant in Remark S1, put $\nu_n^*=(\beta_0,\eta_n^*)$, and let $\nu_{0,n}=(\beta_{0,n},\eta_{0,n})$. The identity $P\dot\ell_{\nu_0}[a]=0$, the bound just established for ℓ_ν, maximality, Lemma S3, and Lemma S2(b) give

$$\begin{gathered}
0\le P\ell_{\nu_0}-P\ell_{\nu_{0,n}}\le P\ell_{\nu_0}-P\ell_{\nu_n^*}=O(J_n^{-2p}),\\
\|\nu_{0,n}-\nu_0\|_{\mathcal{V}}=O(J_n^{-p}),\qquad \|\eta_{0,n}-\eta_0\|_\infty=O(J_n^{1/2-p})=o(1),\\
\sqrt{n_1}|\chi(\nu_{0,n})-\chi(\nu_0)|=O(\sqrt{n_1}J_n^{-p})=o(1).
\end{gathered}$$

The last line follows because each displayed functional is locally Lipschitz in the product norm. Coefficient stability and $\beta_0\in\operatorname{int}(\mathcal{B})$ also place $\nu_{0,n}$ in the interior of $\mathcal{B}\times\mathcal{H}_n$, so $P\dot\ell_{\nu_{0,n}}[b_n]=0$ for every $b_n\in\mathcal{V}_n$.

Let $\mathbb{A}_0(a,b)=P\{\dot\ell_{\nu_0}[a]\dot\ell_{\nu_0}[b]\}$. The derivative bounds used in Lemma S4, the preceding convergence, and the displayed derivative formulae imply

$$\|\mathbb{A}_{0,n}-\mathbb{A}_0\|_{\mathrm{op},n}+\sup_{\substack{b_n\in\mathcal{V}_n\\ \|b_n\|_{\mathcal{I}}\le 1}}|\dot\chi_{\nu_{0,n}}[b_n]-\dot\chi_{\nu_0}[b_n]|\longrightarrow 0.$$

Uniform coercivity and the defining finite-sieve equations consequently give

$$\|a_n^\chi-a^\chi\|_{\mathcal{I}}\le C\inf_{b_n\in\mathcal{V}_n}\|a^\chi-b_n\|_{\mathcal{I}}+o(1)=o(1),$$

because the spline spaces are dense in $L_2([0,\tau])$ and the β component is represented exactly.

Thus the representer approximation in (R3) holds.

Write $a_n^\chi = (b_n, g_n)$. The preceding convergence gives $\|b_n\| + \|g_n\|_{L_2} = O(1)$. Put $\lambda_n = e^{X'\beta_{0,n}}$, $A_n = \lambda_n H_{\eta_{0,n}}(C)$, and $v(a) = (e^a - 1)^{-1}$. Differentiating (S1) gives, for every $z \in \mathrm{span}\{B_1, \dots, B_{J_n}\}$,

$$\mathbb{A}_{0,n}((b_n, g_n), (0, z)) = \int_0^\tau \{\omega_n g_n + K_n g_n + c_n' b_n\}(s) z(s)\, ds,$$
$$\omega_n(s) = e^{\eta_{0,n}(s)} P[\lambda_n\{\mathbf{1}(s \le T) + v(A_n)\mathbf{1}(s \le C)\}].$$

Here $(K_n g)(s) = \int_0^\tau K_n(s,u) g(u)\, du$, with $\sup_{n,s,u} |K_n(s,u)| < \infty$, and c_n is a uniformly bounded vector function. These bounds follow from bounded X, the coefficient box, and $C \ge c_{\mathrm{L}}$, which keep A_n in a compact subset of $(0, \infty)$. The atom at $C = \tau$ also gives constants $0 < c_0 \le \omega_n(s) \le C_0 < \infty$, independent of n and s. For each displayed functional, its derivative in a baseline direction is $\dot\chi_{\nu_{0,n}}[(0,z)] = \int_0^\tau k_n^\chi(s) z(s)\, ds$, where $\sup_n \|k_n^\chi\|_\infty < \infty$, by the derivative formulae above. Thus the defining equation for a_n^χ yields

$$g_n = \Pi_n^{\omega_n} \left\{ \frac{k_n^\chi - K_n g_n - c_n' b_n}{\omega_n} \right\},$$

where $\Pi_n^{\omega_n}$ is the orthogonal projection onto $\mathrm{span}\{B_1, \dots, B_{J_n}\}$ in $L_2(\omega_n(s)\, ds)$. The function inside braces is uniformly bounded. The bounds on ω_n, the fixed spline order, and the quasi-uniform mesh satisfy the conditions of Huang (2003, Appendix, Theorem A.1), so this projection is uniformly bounded in supremum norm. Consequently $\|g_n\|_\infty = O(1)$.

This bound and differentiation of (S1) bound the quadratic remainder of $P\dot\ell_\nu[a_n^\chi]$ about $\nu_{0,n}$ by $C\|\nu - \nu_{0,n}\|_{\mathcal{V}}^2$, uniformly on $\mathcal{N}_n$. Hence the population remainder in (R3) is $O(\sqrt{n_1} r_n^2) = o(1)$.

For the empirical remainder, the term $X'b_n + g_n(T)$ cancels in $\dot\ell_\nu[a_n^\chi] - \dot\ell_{\nu_0}[a_n^\chi]$. The remaining integral terms are uniformly Lipschitz in ν for the product norm. The entropy calculation in Section S4.2.1, applied to these differences, therefore gives

$$\mathbb{E}^* \sup_{\nu \in \mathcal{N}_n} \left| \mathbb{G}_{n_1}\{\dot\ell_\nu[a_n^\chi] - \dot\ell_{\nu_0}[a_n^\chi]\} \right| \lesssim r_n \sqrt{J_n} + J_n/\sqrt{n_1} = o(1).$$

The last equality follows from $\sqrt{n_1} r_n^2 \to 0$ and $J_n^2/n_1 \to 0$. This verifies the remaining two requirements in (R3) for all four functionals.

Put $d_n = \hat\nu - \nu_{0,n}$. The preceding bounds place $\hat\nu, \nu_{0,n} \in \mathcal{N}_n$ with probability tending to one. The first equality below follows from their first-order conditions and the last two requirements in (R3). The second follows from the quadratic remainder for the functional and the defining equation for a_n^χ:

$$\mathbb{A}_{0,n}(a_n^\chi, d_n) = (\mathbb{P}_{n_1} - P)\dot\ell_{\nu_0}[a_n^\chi] + o_p(n_1^{-1/2}),$$
$$\chi(\hat\nu) - \chi(\nu_{0,n}) = \mathbb{A}_{0,n}(a_n^\chi, d_n) + O_p(\delta_J^2).$$

It follows from the bias, $\sqrt{n_1}\delta_J^2 \to 0$, and the representer approximation that

$$\sqrt{n_1}\{\chi(\hat{\nu}) - \chi(\nu_0)\} = \mathbb{G}_{n_1}\dot{\ell}_{\nu_0}[a_n^\chi] + o_p(1) = \mathbb{G}_{n_1}\widetilde{\ell}_\chi + o_p(1).$$

This is the standard sieve M-estimation expansion; see Shen (1997, Theorem 1) and Chen (2007, Theorem 4.3). □

Proof of Proposition S3. Lemma S4 gives invertibility of the nuisance block and convergence of the empirical Schur complement $\widehat{\mathcal{I}}_{\beta,n}$ to $\mathcal{I}(\beta_0)$. The block-inverse identity gives $[\hat{\mathbb{I}}^{-1}]_{\beta\beta} = \widehat{\mathcal{I}}_{\beta,n}^{-1}$, and the result follows.

□

S4.3 Identification and information

S4.3.1 Weak identification of baseline scale

Proof of Proposition 4. Take the constant direction $g \equiv 1$. Then $G(s) = \int_0^s e^{\eta_0} = H_0(s)$, so $e^\psi G(t) = B_t$ and $e^\psi G(c)\, e^{-A}/(1-e^{-A}) = A\, e^{-A}/(1-e^{-A}) = w(A)$ with $A = e^\psi H_0(c)$; hence by (S3)

$$\dot{\ell}_\eta[1] = 1 - B_t - w(A), \qquad \dot{\ell}_\beta = X\,\dot{\ell}_\eta[1].$$

Condition on $(C,X) = (c,x)$. Since $1 - w(A)$ is then constant, $\mathrm{Var}(\dot{\ell}_\eta[1] \mid c,x) = \mathrm{Var}(B_t \mid c,x)$. Set $U = H_0(T)/H_0(c) \in (0,1)$, so $B_t = A\,U$. Under $q(t \mid c,x) \propto h_0(t)\, e^{-e^\psi H_0(t)}$ on $(0,c)$, the change of variables $H_0(t) = u\, H_0(c)$, $h_0(t)\, dt = H_0(c)\, du$, gives the density of U proportional to e^{-Au} on $(0,1)$, i.e. a truncated Exponential(A) with normalizer $A/(1-e^{-A})$. Writing $v(A) = \mathrm{Var}(U)$, we obtain $\mathrm{Var}(B_t \mid c,x) = A^2 v(A)$. The conditional mean of $\dot{\ell}_\eta[1]$ is zero (it is a conditional-likelihood score), so by the law of total variance

$$\mathcal{I}_{\mathrm{sc}} = \mathbb{E}\Big[\,\mathrm{Var}(\dot{\ell}_\eta[1] \mid C,X) \mid S^\tau = 1\Big] = \mathbb{E}\Big[A^2 v(A) \mid S^\tau = 1\Big].$$

As $A \to 0$, U converges to Uniform$(0,1)$ and $v(A) \to 1/12$; moreover $A = -\log\{1 - F_T(C \mid X)\} = F_T(C \mid X)\{1 + O(F_T)\}$, whence $\mathcal{I}_{\mathrm{sc}} = \frac{1}{12}\mathbb{E}[F_T(C \mid X)^2 \mid S^\tau = 1]\,(1 + o(1))$. As $A \to \infty$, U concentrates near 0 with $v(A) \sim A^{-2}$, so $A^2 v(A) \to 1$. □

S4.3.2 Local asymptotic minimax bound under limited follow-up

Theorem S3 (Local asymptotic minimax bound under limited follow-up)**.** *Consider the one-dimensional submodel $H_{0,\delta}(t) = e^\delta H_0(t)$ with $\beta = \beta_0$ and $\delta_0 = 0$. For a fixed capped experiment, hold the distribution of $(C,X) \mid S = 1$ fixed and suppose $A(C,X) = e^{X'\beta_0}H_0(C)$ is bounded above almost surely and positive with positive probability. Let $I_\delta = \mathcal{I}_{\mathrm{sc}}$ be the information calculated in Proposition 4.*

The model for labelled positive observations is locally asymptotically normal at δ_0 with information I_δ. Hence the local asymptotic minimax risk for every estimator sequence and every nonnegative loss with convex, symmetric sublevel sets is bounded below by the risk under $N(0, I_\delta^{-1})$. If a regular estimator has a finite-variance limit L_{δ_0} under $\sqrt{n_1}$ scaling, then $\mathrm{Var}(L_{\delta_0}) \geq I_\delta^{-1}$.

These are fixed-design bounds. Along any sequence of designs for which $\mathbb{E}\{F_T(C \mid X)^2 \mid S = 1\} \to 0$, one has $I_\delta^{-1} \to \infty$; if also $\sup_{c,x} A(c, x) \to 0$, then

$$I_\delta^{-1} \sim \frac{12}{\mathbb{E}\{F_T(C \mid X)^2 \mid S = 1\}}.$$

Under Conditions (E1)–(E3), the bound is unchanged under $\sqrt{n_1}$ scaling when both labelled and unlabelled observations are used. This conclusion need not hold when the labelling mechanism is restricted, since the labelling indicator can then contain information about δ.

Proof. The score, the conditional density of U, and the information were obtained in the proof of Proposition 4. Under the scale submodel, the conditional density is $f_{e^\delta A}(u)$. The map

$$(z, u) \longmapsto \left\{ \frac{z e^{-zu}}{1 - e^{-z}} \right\}^{1/2}, \qquad z \geq 0, \quad 0 < u < 1,$$

with value 1 at $z = 0$, is smooth on compact sets. Boundedness of A therefore gives a common square-integrable envelope for its first two derivatives near $\delta = 0$. The model is differentiable in quadratic mean with the stated score and information. Theorem 7.2 of van der Vaart (1998) gives local asymptotic normality for m independent labelled-positive observations; Theorems 8.11 and 8.8 of that reference give the minimax and convolution bounds. Since $n_1 \to \infty$ in probability, the same conclusions hold with $m = n_1$.

For the comparison across fixed designs, put

$$R(A) = \frac{A^2 \,\mathrm{Var}_A(U)}{(1 - e^{-A})^2}, \qquad R(0) = \frac{1}{12}.$$

Direct integration gives $A^2 \,\mathrm{Var}_A(U) = 1 - A^2 e^A/(e^A - 1)^2$. Thus R is continuous, with limits $1/12$ at zero and 1 at infinity, and is bounded by some finite constant K. Consequently,

$$I_\delta \leq K \,\mathbb{E}\{F_T(C \mid X)^2 \mid S = 1\},$$

which proves divergence of I_δ^{-1} whenever that expectation tends to zero. Proposition 4 gives the sharper equivalence when $\sup_{c,x} A(c, x) \to 0$. Finally, by Theorem 4, $S\dot{\ell}_\delta$ is efficient and the information when both labelled and unlabelled observations are used is $p_S I_\delta$. Since $n_1/n \to p_S$, the lower bound under $\sqrt{n_1}$ scaling is again I_δ^{-1}. When the labelling mechanism is restricted, the labelling-indicator contribution need not vanish, so this argument does not apply. □

S4.3.3 Proof of Proposition 2

(a) Nonparametric rescaling. This is Theorem 2(b): equality of q on $\{0 < t < c \leq \tau_x\}$ forces $F_1(\cdot \mid x)/F_1(\tau_x \mid x) = F_0(\cdot \mid x)/F_0(\tau_x \mid x)$, i.e. $F_1(\cdot \mid x) = a(x)F_0(\cdot \mid x)$ with $a(x) = F_1(\tau_x \mid x)/F_0(\tau_x \mid x) > 0$, and the level $a(x)$ is free.

(b) Proper PH injectivity under a nondegenerate covariate effect. Suppose $q_{\beta,H}(t \mid c, x) = q_{\beta_0,H_0}(t \mid c, x)$ for $P_{C,X|S=1}$-almost every (c, x) and Lebesgue-almost every $t \in (0, c)$, with both distributions proper PH. There is then a measurable $\mathcal{X}_0 \subseteq \operatorname{supp}(X \mid S = 1)$ with $\mathbb{P}(X \in \mathcal{X}_0 \mid S = 1) = 1$ such that, for every $x \in \mathcal{X}_0$, the equality holds for Lebesgue-almost every pair $0 < t < c \leq \tau_x$. Integrating in t over $(0, s]$ gives $F_T^{\beta,H}(s \mid x)/F_T^{\beta,H}(c \mid x) = F_T^{\beta_0,H_0}(s \mid x)/F_T^{\beta_0,H_0}(c \mid x)$ for Lebesgue-almost every supported pair (s, c); both ratio maps admit continuous versions, because the distribution functions are continuous and the denominators are positive, so the equality extends to every supported pair, and Theorem 2(b) gives $F_T^{\beta,H}(\cdot \mid x) = a(x)\, F_T^{\beta_0,H_0}(\cdot \mid x)$ on $\mathcal{T}_x$ for some $a(x) > 0$. The comparisons below are carried out for $x \in \mathcal{X}_0$. The PH structure forces $a(x) \equiv 1$ without a terminal normalization. Write $\lambda = e^{x'\beta_0}$, $\widetilde{\lambda} = e^{x'\beta}$, and $v = H_0(s)$; by Assumption 3(a) each $\mathcal{T}_x$ is an interval $(0, \tau_x]$, so for any two covariate points x_1, x_2 the identities below hold jointly for v in the common interval $(0, \bar{v}_{12})$, where $\bar{v}_{12} = H_0(\min\{\tau_{x_1}, \tau_{x_2}\}) > 0$. The rescaling identity reads $1 - \exp\{-\widetilde{\lambda}\, H(s)\} = a(x)\, [1 - e^{-\lambda v}]$, so that

$$H(s) = \varphi(a(x), \lambda, \widetilde{\lambda}; v) := -\widetilde{\lambda}^{-1} \log\Big\{1 - a(x)\, (1 - e^{-\lambda v})\Big\}.$$

The left side is one fixed function of s, and the rescaling identity shows that it depends on s only through $v = H_0(s)$; since the conditional distributions admit densities, H_0 is continuous and nondecreasing with $H_0(0) = 0$, and Assumption 3(b) gives $H_0(s) > 0$ for $s > 0$, so the image of every interval $(0, \bar{s}]$ under H_0 contains $(0, H_0(\bar{s})]$, and the right side must be the same function of v, on a common interval adjoining zero, for every x considered.

Derivatives at $v = 0$. Under the assumptions of the proposition the displayed identity holds for all v in the common interval $(0, \bar{v}_{12})$. The value $a(x) = F_T^{\beta,H}(\tau_x \mid x)/F_T^{\beta_0,H_0}(\tau_x \mid x)$ satisfies $a(1 - e^{-\lambda H_0(\tau_x)}) = F_T^{\beta,H}(\tau_x \mid x) \leq 1$ by properness of $F_T^{\beta,H}$, so the map $v \mapsto \varphi(a, \lambda, \widetilde{\lambda}; v)$ is real-analytic on the open interval $I_{a,\lambda} = \{v : 1 - a(1 - e^{-\lambda v}) > 0\}$, which is connected and contains $(-\epsilon, \bar{v}_{12})$ for some $\epsilon > 0$, hence a neighbourhood of $v = 0$. For two covariate points x_1, x_2 as below, the functions $\varphi(a_1, \lambda_1, \widetilde{\lambda}_1; \cdot)$ and $\varphi(a_2, \lambda_2, \widetilde{\lambda}_2; \cdot)$ agree on $(0, \bar{v}_{12})$ and are real-analytic on the connected intersection $I_{a_1,\lambda_1} \cap I_{a_2,\lambda_2}$, so all their derivatives at $v = 0$ coincide, and the Taylor-coefficient comparison below is legitimate. The expansion at $v = 0$ is

$$\widetilde{\lambda}\, \varphi = a\lambda\, v + \tfrac{\lambda^2}{2}\, a(a - 1)\, v^2 + \tfrac{\lambda^3}{6}\, a(a - 1)(2a - 1)\, v^3 + O(v^4).$$

Because $x'\beta_0$ is nonconstant on the support, every neighbourhood of a support point carries positive probability under the law of X given $S = 1$, and λ is continuous, the full-measure set

$\mathcal{X}_0$ contains two points x_1, x_2 with $\lambda_1 \neq \lambda_2$; write $(a_i, \lambda_i, \widetilde{\lambda}_i)$ for the corresponding parameters. Matching Taylor coefficients of the two representations of H:

$$\frac{a_1\lambda_1}{\widetilde{\lambda}_1} = \frac{a_2\lambda_2}{\widetilde{\lambda}_2}, \qquad \frac{\lambda_1^2 a_1(a_1-1)}{\widetilde{\lambda}_1} = \frac{\lambda_2^2 a_2(a_2-1)}{\widetilde{\lambda}_2},$$
$$\frac{\lambda_1^3 a_1(a_1-1)(2a_1-1)}{\widetilde{\lambda}_1} = \frac{\lambda_2^3 a_2(a_2-1)(2a_2-1)}{\widetilde{\lambda}_2}.$$

Dividing the second relation by the first gives $\lambda_1(a_1-1) = \lambda_2(a_2-1) =: m$, and dividing the third by the first gives $\lambda_1^2(a_1-1)(2a_1-1) = \lambda_2^2(a_2-1)(2a_2-1)$. If $m \neq 0$, substituting $a_i - 1 = m/\lambda_i$ into the last display yields $\lambda_1(2a_1-1) = \lambda_2(2a_2-1)$, and since $\lambda_i(2a_i-1) = 2m+\lambda_i$, this forces $\lambda_1 = \lambda_2$, a contradiction. Hence $m = 0$, i.e. $a_1 = a_2 = 1$. Pairing each $x \in \mathcal{X}_0$ with a point of $\mathcal{X}_0$ whose λ differs (such a point exists by the argument above) gives $a(x) = 1$ for every $x \in \mathcal{X}_0$. Thus $F_T^{\beta,H}(\cdot \mid x) = F_T^{\beta_0,H_0}(\cdot \mid x)$ on $\mathcal{T}_x$ for every $x \in \mathcal{X}_0$. Under PH, $\log\{-\log(1 - F_T(s \mid x))\} = x'\beta + \log H(s)$ on $\mathcal{T}_x$. For any $x_1, x_2 \in \mathcal{X}_0$ choose $0 < s < \min\{\tau_{x_1}, \tau_{x_2}\}$; with $a \equiv 1$ on $\mathcal{X}_0$, the two representations give $e^{x_j'\beta}H(s) = e^{x_j'\beta_0}H_0(s)$ for $j = 1, 2$, and taking logarithms and differencing yields $(x_1 - x_2)'(\beta - \beta_0) = 0$. Thus $x'(\beta - \beta_0)$ is constant on $\mathcal{X}_0$, and nonsingularity of $\mathrm{Var}(X \mid S = 1)$ forces $\beta = \beta_0$, whence $H = H_0$ on $\mathcal{T}_x$ for every $x \in \mathcal{X}_0$. It remains to extend the equality from $\mathcal{X}_0$ to the full support. Let $x \in \mathrm{supp}(X \mid S = 1)$ and $t < \tau_x$. By the assumed lower semicontinuity of $z \mapsto \tau_z$, the set $\{z : \tau_z > t\}$ contains a neighbourhood of x; every neighbourhood of a support point has positive probability and therefore meets the full-measure set $\mathcal{X}_0$, so there is $z \in \mathcal{X}_0$ with $\tau_z > t$. Since the baseline is common across covariate values, $H(t) = H_0(t)$. Letting $t \uparrow \tau_x$ and using continuity of the distribution functions gives $F_T^{\beta,H}(\cdot \mid x) = F_T^{\beta_0,H_0}(\cdot \mid x)$ on $\mathcal{T}_x$; the two distributions therefore coincide for every x in the support.

(c) Defective PH. Write $F(t \mid x) = p(x)G_{\beta,H_0}(t \mid x)$ with $G_{\beta,H_0}(t \mid x) = 1 - \exp\{-e^{x'\beta}H_0(t)\}$ a proper PH distribution, a defective (long-term-survivor) model in the sense of Maller and Zhou (1996). Since the factor $p(x)$ cancels from the ratio, $q(t \mid c, x) = \frac{g_{\beta,H_0}(t|x)}{G_{\beta,H_0}(c|x)}$ is independent of $p(x)$. Thus two candidate levels $p(x)$ and $\tilde{p}(x)$ correspond to the rescaling in Theorem 2(b), with $a(x) = \tilde{p}(x)/p(x)$. Equality of q therefore identifies the proper event-time component (β, H_0) exactly under the covariate conditions in part (b), while $p(x)$ is unconstrained by q and requires additional information or restrictions for identification. $\square$

S4.3.4 Proof of Proposition 3

(a) Global flatness. Absorb the constant effect into the baseline, so $F_0(t) = 1 - e^{-H_0(t)}$ with $F_0(\tau) < 1$. Fix $a \in (0, 1/F_0(\tau))$ and set $\widetilde{F}(t) = a\,F_0(t)$ on $[0, \tau]$; then $\widetilde{F}(\tau) = aF_0(\tau) < 1$, and $\widetilde{F}$ extends beyond τ to a proper continuous distribution (e.g. linearly, then smoothly to 1). The function $\widetilde{H}_a(t) = -\log\{1 - \widetilde{F}(t)\} = -\log\{1 - aF_0(t)\}$ is continuous, nondecreasing, and finite

on $[0,\tau]$, hence a valid proper PH baseline. For $0 < t < c \le \tau$,

$$\widetilde{q}(t \mid c, x) = \frac{\widetilde{f}(t)}{\widetilde{F}(c)} = \frac{a f_0(t)}{a F_0(c)} = \frac{f_0(t)}{F_0(c)} = q(t \mid c, x),$$

so q is unchanged while the terminal level $\widetilde{F}(\tau) = aF_0(\tau)$ ranges over $(0,1)$ as a ranges over $(0, 1/F_0(\tau))$.

(b) Local flatness. With the constant effect absorbed, $\psi = 0$ and $A = H_0(c)$ in (S3). Take $g^*(u) = e^{H_0(u)}$, which is bounded on $[0,\tau]$ by $e^{H_0(\tau)} = 1/\{1 - F_0(\tau)\} < \infty$, hence an admissible direction. Then $G(s) = \int_0^s e^{\eta_0(u)} g^*(u)\, du = \int_0^s e^{H_0(u)}\, dH_0(u) = e^{H_0(s)} - 1$, and (S3) gives

$$\begin{aligned}\dot{\ell}_\eta[g^*] &= g^*(t) - G(t) - G(c)\,\frac{e^{-A}}{1 - e^{-A}} = e^{H_0(t)} - \Big(e^{H_0(t)} - 1\Big) - \Big(e^{A} - 1\Big)\frac{e^{-A}}{1 - e^{-A}} \\ &= 1 - \frac{1 - e^{-A}}{1 - e^{-A}} = 0\end{aligned}$$

identically in (t, c, x). (The path $a \mapsto \widetilde{H}_a$ of part (a) has $\partial_a \widetilde{H}_a|_{a=1}(s) = F_0(s)/\{1 - F_0(s)\} = e^{H_0(s)} - 1 = G(s)$, so g^* is precisely the tangent of the global flat family.) Hence the information operator is singular at g^* and the coercivity of Lemma S2 fails. By part (a), the same observed law arises from distinct values of $H_0(\tau)$, the absolute $F_T(t \mid x)$, and $\pi(x)$. Therefore no estimator can be consistent for these functionals at every law in the observationally equivalent family; in particular, no regular estimator of these functionals exists. $\square$

S4.4 Reverse-time partial likelihood

Suppose the reverse hazard satisfies the proportional reverse-hazards model

$$r_T(t \mid x) = r_0(t) e^{x'\beta}, \qquad \text{equivalently} \quad F_T(t \mid x) = F_0(t)^{\exp(x'\beta)}. \tag{S11}$$

Then

$$q(t \mid c, x) = e^{x'\beta} r_0(t) \exp\Big\{ - e^{x'\beta} \int_t^c r_0(u)\, du \Big\}, \qquad 0 < t < c, \tag{S12}$$

where $0 < t < c$. Fix an analysis horizon τ interior to the support of the censoring time ($\tau < \tau_C$, with τ_C as in Lemma S1). The reverse-time partial likelihood uses all labelled-positive events in $(0, \tau]$:

$$\mathrm{PL}_\tau(\beta) = \prod_{\substack{i: S_i = 1 \\ T_i \le \tau}} \frac{e^{x_i'\beta}}{\sum_{j \in \widetilde{\mathcal{R}}(T_i)} e^{x_j'\beta}}, \tag{S13}$$

where

$$\widetilde{\mathcal{R}}(u) = \{ j : T_j \le u \le C_j,\ S_j = 1 \} \tag{S14}$$

is the reverse risk set; $\hat{\beta}$ maximizes $\mathrm{PL}_\tau(\beta)$ over the compact set $\mathcal{B}$ of (PL1). For a labelled positive, put $u_v = \tau e^{-v}$ and $Y(v) = \mathbf{1}\{T \le u_v < C\}$ for $v \ge 0$, the at-risk indicator on the unbounded reverse-time scale $v = \log(\tau/u)$, which describes membership of $\widetilde{\mathcal{R}}(u_v)$ up to the null entry boundary $\{C = u_v\}$; write $\lambda_0^*(v) = u_v\, r_0(u_v)$ for the baseline hazard on this scale, and define $s^{(k)}(v,\beta) = \mathbb{E}[Y(v)e^{X'\beta}X^{\otimes k} \mid S = 1]$ for $k = 0, 1, 2$. We impose the following regularity conditions.

(PL1) X is bounded and β_0 is interior to a compact parameter set $\mathcal{B}$;

(PL2) for every $\epsilon \in (0,\tau)$, the baseline increment $A_0(\epsilon) = \int_\epsilon^\tau r_0(u)\,du$ is finite and $\mathbb{P}(T \le \epsilon,\ C > \tau \mid S = 1) > 0$, the latter being equivalent to $\mathbb{P}\{Y(v) = 1 \text{ for all } v \in [0, L] \mid S = 1\} > 0$ for every finite L;

(PL3) $\Sigma(\beta_0) = \int_0^\infty \left\{ s^{(2)}(v,\beta_0)/s^{(0)}(v,\beta_0) - \left(s^{(1)}(v,\beta_0)/s^{(0)}(v,\beta_0) \right)^{\otimes 2} \right\} s^{(0)}(v,\beta_0)\,\lambda_0^*(v)\,dv$ is finite and positive definite.

Proposition S4 (Reverse-time partial likelihood)**.** *Under Assumptions 1–2, the proportional reverse-hazards model, and Conditions (PL1)–(PL3),*

(a) *profiling the reverse-time multiplicative-intensity likelihood of the labelled-positive processes at horizon τ over the baseline reverse cumulative-hazard measure yields the partial likelihood* (S13) *for β;*

(b) *for each fixed labelling mechanism $\rho(c, x)$ whose induced labelled-positive law satisfies* (PL1)–(PL3)*, the partial-likelihood score is conditionally unbiased at β_0, and $\hat{\beta}$ is consistent with $\sqrt{n_1}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0)^{-1})$; the labelling mechanism does not enter the conditional hazard, but may change the labelled-positive law, and with it $s^{(k)}$ and $\Sigma(\beta_0)$;*

(c) *a Breslow-type estimator consistently estimates the baseline reverse cumulative-hazard increment $A_0(t) = \int_t^\tau r_0(u)\,du$ uniformly on $[\epsilon,\tau]$ for every fixed $\epsilon \in (0,\tau)$, equivalently the relative functional $F_0(t)/F_0(\tau)$; the absolute baseline level is not identified within* (S11).

The estimator is the reverse-time analogue of Cox regression: under (S11), the partial likelihood (S13) is the right-truncation partial likelihood of Kalbfleisch and Lawless (1991), in the reverse-time treatment of right-truncated data introduced by Lagakos et al. (1988). The proof in Section S4.4.1 shows that the labelling mechanism does not enter the reverse-time conditional hazard and applies the infinite-time-interval theory of Andersen and Gill (1982, Section 4).

Whether the partial likelihood attains the semiparametric efficiency bound for β is not established here. Theorem 6 concerns the forward proportional-hazards model and does not apply to the proportional reverse-hazards model.

S4.4.1 Proof of Proposition S4

Reverse-time processes and filtration. Relabel the labelled positives $i = 1, \ldots, n_1$. A forward counting process has already jumped wherever the subject is at reverse-time risk, so a forward filtration cannot carry the reverse-hazard intensity; apply instead the reverse-time transformation $v = \log(\tau/u) \in [0, \infty)$ from the statement, with $u_v = \tau e^{-v}$. Define the reverse entry time $E_i = \log(\tau/C_i)$, the reverse event time $V_i = \log(\tau/T_i)$, the counting process $N_i(v) = \mathbf{1}\{T_i \leq \tau,\ V_i \leq v\}$, and the at-risk indicator $Y_i(v) = \mathbf{1}\{E_i < v \leq V_i\} = \mathbf{1}\{T_i \leq u_v < C_i\}$, membership of $\widetilde{\mathcal{R}}(u_v)$ with the entry boundary excluded; an event time falls on $\{u_v = C_j\}$ with probability zero under Assumption 1, so the closed risk set of (S14) gives the same partial likelihood almost surely. Take the filtration

$$\mathcal{F}_v = \sigma\Big\{(C_i, X_i, \mathbf{1}\{T_i \leq \tau\}) : i \leq n_1\Big\} \vee \sigma\{N_i(w) : w \leq v,\ i \leq n_1\},$$

whose time-zero marks are observed for every labelled positive; Y_i and the constant covariates are left continuous with right-hand limits, so Y_i is predictable. A labelled positive with $T_i > \tau$ is never at risk and has no event. The labelling mechanism ρ determines the distribution of the marks but not the conditional hazard derived in Step 1.

Step 1: conditional hazard after reverse-time entry. By Theorem 1, conditional on $(C_i, X_i, S_i = 1)$ the event time T_i has density $q(t \mid C_i, X_i) = f_T(t \mid X_i)/F_T(C_i \mid X_i)$ on $(0, C_i)$. For $u \leq \tau \wedge C_i$,

$$\mathbb{P}\Big(T_i \in (u - du, u] \,\Big|\, T_i \leq u,\, C_i, X_i, S_i = 1\Big) = \frac{f_T(u \mid X_i)/F_T(C_i \mid X_i)}{F_T(u \mid X_i)/F_T(C_i \mid X_i)}\, du = r_T(u \mid X_i)\, du :$$

the truncation denominator cancels, and conditioning further on $\{T_i \leq \tau\} \supseteq \{T_i \leq u\}$ does not change this conditional probability. Since $|du_v/dv| = u_v$, on $\{Y_i(v) = 1\}$ the conditional hazard of V_i at v given $\mathcal{F}_{v-}$ is therefore $\alpha_i(v) = u_v\, r_T(u_v \mid X_i) = e^{X_i'\beta_0}\, \lambda_0^*(v)$ under (S11): entry determines who is at risk, not the hazard.

Step 2: martingale structure and score unbiasedness. By Step 1 the processes N_i have the multiplicative intensity $\lambda_i(v) = Y_i(v)\, e^{X_i'\beta_0}\, \lambda_0^*(v)$ with respect to $(\mathcal{F}_v)$, so $M_i(v) = N_i(v) - \int_0^v \lambda_i(w)\, dw$ are orthogonal square-integrable martingales on $[0, \infty)$: each N_i has at most one jump and its compensator mass is at most $e^{X_i'\beta_0} A_0(T_i) < \infty$ almost surely, because $F_0(T_i) > 0$. Writing $S^{(k)}(v, \beta) = n_1^{-1} \sum_{j \leq n_1} Y_j(v) e^{X_j'\beta} X_j^{\otimes k}$ and $\bar{X}(v, \beta) = S^{(1)}/S^{(0)}$, the partial-likelihood score is

$$U(\beta_0) = \sum_{i:S_i=1} \int_0^\infty \{X_i - \bar{X}(v, \beta_0)\}\, dN_i(v) = \sum_{i:S_i=1} \int_0^\infty \{X_i - \bar{X}(v, \beta_0)\}\, dM_i(v),$$

the compensator terms cancelling by the definition of $\bar{X}$. The integrand is bounded and predictable, so $\mathbb{E}\{U(\beta_0) \mid \mathcal{F}_0\} = 0$: the score is conditionally unbiased given the realized marks, hence unbiased under every labelling mechanism $\rho(c, x)$, the unbiasedness part of claim (b).

Step 3: profiling and claim (a). By (S12) and Step 1, the likelihood of the processes of Step 1 given the time-zero marks is $\prod_{i:T_i \le \tau} e^{x_i'\beta} r_0(t_i) \exp\{-e^{x_i'\beta} \int_{t_i}^{\tau \wedge c_i} r_0\}$, the likelihood of left-truncated data in reverse time with intensity λ_i; events with $T_i \le \tau$ contribute jumps and every labelled positive contributes its at-risk exposure. Replacing the continuous baseline reverse cumulative-hazard measure by a discrete measure with jumps $a_k \ge 0$ at the finitely many observed reverse event times v_k, the log-likelihood is maximized over a_k at $\hat{a}_k(\beta) = d_k / \sum_{j \le n_1} Y_j(v_k) e^{X_j'\beta}$, with d_k the multiplicity at v_k ($d_k = 1$ almost surely under Assumption 1); substituting back yields $\log \mathrm{PL}_\tau(\beta)$ of (S13) plus terms free of β, the monotone parameter-free time change preserving event order and risk sets. The profile likelihood for β is the partial likelihood, and $\hat{\beta}$ does not involve r_0: claim (a).

Step 4: Cox theory on an infinite time interval and claim (b). Conditional on $S = 1$, the processes (N_i, Y_i) with the constant covariates $Z_i(v) = X_i$ are i.i.d. and left continuous with right-hand limits; every limit below is taken under $\mathbb{P}(\cdot \mid S = 1)$, so $s^{(k)}$ and $\Sigma(\beta_0)$ may depend on ρ even though the conditional hazard does not. The hypotheses of Theorem 4.2 of Andersen and Gill (1982) hold: X is bounded (PL1); $N_i(\infty) \le 1$, so $\mathbb{E}\{N_i(\infty) \mid S_i = 1\} < \infty$; for each finite L, $\mathbb{P}\{Y_i(v) = 1 \text{ for all } v \le L \mid S_i = 1\} > 0$ and $\int_0^L \lambda_0^*(v)\, dv = A_0(\tau e^{-L}) < \infty$ by (PL2); and $\Sigma(\beta_0)$ is positive definite (PL3). The tail is negligible in their sense because

$$\mathbb{E}\Big[\int_L^\infty Y_i(v) e^{X_i'\beta_0} \lambda_0^*(v)\, dv \,\Big|\, S_i = 1 \Big] = \mathbb{E}\{N_i(\infty) - N_i(L) \mid S_i = 1\} \longrightarrow 0 \qquad (L \to \infty),$$

and bounded X extends the same control to the tails of the score and of the information. Theorem 4.2 of Andersen and Gill (1982), together with the extensions of Lemma 3.1, Corollary 3.3, and Theorem 3.4 recorded immediately before it, then gives $\hat{\beta} \to_p \beta_0$, $\sqrt{n_1}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0)^{-1})$, and consistency of the observed-information estimator: claim (b). Simultaneous jumps, excluded there, have probability zero here.

Step 5: the Breslow estimator and claim (c). Define the reverse-time Breslow estimator $\widehat{\Lambda}_0^*(w) = \int_0^w \{\sum_{j \le n_1} Y_j(v) e^{X_j'\hat{\beta}}\}^{-1}\, d\bar{N}(v)$ for $w \in [0, \infty)$, with $\bar{N} = \sum_{i \le n_1} N_i$, and set $\widehat{A}_0(t) = \widehat{\Lambda}_0^*(\log(\tau/t))$ for $t \in (0, \tau]$. For every fixed $\epsilon \in (0, \tau)$ the v-interval $[0, \log(\tau/\epsilon)]$ is finite, so the extension of Theorem 3.4 recorded in Section 4 of Andersen and Gill (1982), with $\hat{\beta} \to_p \beta_0$, gives uniform consistency of $\widehat{A}_0$ on $[\epsilon, \tau]$ for $A_0(t) = \int_t^\tau r_0(u)\, du = \log F_0(\tau) - \log F_0(t)$, so that $\exp\{-\widehat{A}_0(t)\}$ consistently estimates the relative baseline $F_0(t)/F_0(\tau)$ there. The absolute level is not identified: replacing F_0 by cF_0 ($c > 0$, $cF_0(\tau) \le 1$) multiplies $F_T(\cdot \mid x)$ by $c^{\exp(x'\beta)}$ and leaves q, the partial likelihood, and A_0 unchanged, so the proportional reverse-hazards family is closed under the level rescaling of Theorem 2(b): claim (c). $\square$

Table S1: Finite-sample performance of the capped proportional-hazards sieve

		bias		RMSE		95% coverage (percentile)		95% coverage (basic; sensitivity)	
n	$n_{1,\tau}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
500	191	0.23	−0.10	0.90	0.78	0.96	0.96	0.68	0.94
1,000	382	0.12	−0.07	0.52	0.50	0.94	0.93	0.73	0.92
2,000	764	0.05	−0.02	0.30	0.21	0.94	0.95	0.79	0.92

Note. 1,000 replications; coefficient bounds $|\beta_j| \leq 6$ and $|\xi_k| \leq 20$. RMSE means root-mean-squared error. Intervals are 95% weighted-bootstrap percentile intervals; basic reverse-percentile intervals are reported as a sensitivity analysis.

S5 Simulation details and additional results

S5.1 Implementation details for the simulation studies

Capped-experiment finite-sample diagnostic under (S2). Table S1 examines the capped sieve of Section 4.4 in finite samples under a baseline that satisfies (S2): the bounded-below, non-monotone, non-Weibull hazard $h_0(t) = 0.005 + 0.040 \exp\{-(t-30)^2/(2 \cdot 12^2)\}$ (so $h_0 \in [0.005, 0.045]$ is bounded away from 0 on $[0, \tau]$), $\beta_0 = (0.6, -0.4)$, $C \sim \text{Uniform}[10, 120]$, (C, X)-dependent labelling (SCAR fails), with the administrative-capping construction $\tau = 100$, $c_{\text{L}} = 10$, $D = \min(C, \tau)$, $S^\tau = S\,\mathbf{1}\{T \leq \tau\}\,\mathbf{1}\{C \geq c_{\text{L}}\}$. The estimator is the unpenalized box-constrained maximizer of (S2). For the finite-sample diagnostic, we use $B = 199$ normalized Exp(1) multiplier weights. We compute an interval in every replication, with none discarded.

Across the entries of Table S1, percentile coverage ranges from 0.929 to 0.962, whereas the basic interval substantially undercovers $\hat{\beta}_1$ (0.68, 0.73 and 0.79 across n) and undercovers $\hat{\beta}_2$ less severely (0.94, 0.92 and 0.92). The two intervals differ materially in this finite-sample design. Both bias and RMSE decrease with sample size. A Wald interval is formed from the profiled observed information at the multi-start $\hat{\theta}$; it is available only in the replications whose profiled observed information is positive definite and passes a condition-number check, a fraction of 0.87, 0.97 and 1.00 across n. At $n = 500$ a baseline (spline) coefficient lies at the box boundary in about 70% of bootstrap draws, against 40% at $n = 1{,}000$ and 14% at $n = 2{,}000$, so the $n = 500$ results are interpreted with caution. A separate experiment for the recovery of F_T uses $J_n = 12$ and reports grid-averaged squared error over $t \in [2, 100]$ (80 points, 200 draws, $n = 3{,}000$) at two covariate profiles: sieve 0.0219 and 0.0086, against 0.0557 and 0.0128 for the misspecified Weibull, a factor of 1.5 to 2.5. The atom mass $\widehat{\kappa} \approx 0.27$ is stable across n.

The simulation studies in Section 5 use the Weibull conditional likelihoods below, optimized over $(\log a, \log \lambda, \beta)$ by Nelder–Mead or by quasi-Newton (L-BFGS-B, with a Nelder–Mead fallback), depending on the experiment. All optimizations in these simulation studies start at the data-generating $(\log a, \log \lambda)$ with zero regression coefficients, a common starting rule across the compared estimators; the `pusurv` package instead uses a data-driven start based on the median censoring time of the labelled positives. For the proposed estimator,

$$\ell_n(\theta) = \sum_{i:S_i=1} \Big[\log f_\theta(T_i \mid X_i) - \log F_\theta(C_i \mid X_i) \Big];$$

the positive-only estimator drops the $-\log F_\theta(C_i \mid X_i)$ truncation term; the S-as-event comparator maximizes $\sum_i [S_i \log f_\theta(T_i \mid X_i) + (1-S_i) \log\{1 - F_\theta(C_i \mid X_i)\}]$, retaining labelled event times but treating unlabelled subjects as right-censored at C_i; the full-data MLE replaces S_i by the true Y_i and uses every event time. Event-before-censoring probabilities are computed as $\hat{\pi}(x) = \mathbb{E}_{G_C}[F_{\hat{\theta}}(C \mid x)]$, with Wald standard errors obtained from the inverse observed information. The semiparametric sieve estimator of Section 4.4 uses a log-hazard B-spline with Gauss–Legendre evaluation of H_0.

S5.2 Additional simulation results

The joint-likelihood comparison uses all n units, with labelled contribution $\rho(c, x; \alpha) F_\theta(c \mid x) q_\theta(t \mid c, x) = \rho(c, x; \alpha) f_\theta(t \mid x)$ and unlabelled contribution $1 - \rho(c, x; \alpha) F_\theta(c \mid x)$; its misspecified variant fits a constant ρ. Table S2 reports bias and RMSE for these fits and for the complete labelling comparison.

Modelled labelling and complete labelling. With ρ correctly specified, the joint likelihood (5,000 replications, $n = 3{,}000$, main design) improves on the conditional likelihood by only a few percent in RMSE. In this correctly specified finite-dimensional labelling model, the labelling-indicator score retains a nonzero residual after adjustment for the labelling nuisance, so the regular joint maximum-likelihood estimator uses additional information; the gain is specific to this design and need not be strict in every restricted-labelling model or parameter direction. With ρ misspecified as constant it is severely biased while the conditional likelihood, which never models ρ, is unaffected. At complete labelling ($\rho \equiv 1$, so $S = \Delta$ and the data are ordinary right-censored), the full censored-data likelihood roughly halves the RMSE relative to the conditional likelihood. This setting is used for the internal comparison in Section S6.5.

Classical right truncation. Applied to the labelled positives, the Lynden-Bell estimator (Lynden-Bell, 1971; Woodroofe, 1985) targets the marginal relative distribution $R = F_T(20)/F_T(40)$ and is the $\beta = 0$ case of the reverse-time estimator of Section S4.4. Where its conditions hold (no covariate effects, constant labelling, so T and C are quasi-independent among labelled positives) it is essentially unbiased and the conditional likelihood without covariates matches it at smaller RMSE (0.011 against 0.017); in the main design the (C, X)-dependent labelling and covariate effects induce marginal T–C dependence and it acquires a systematic bias at approximately twice the RMSE of the conditional likelihood with covariates

Table S2: Conditional and joint-likelihood comparisons

Design	Estimator	$\log\lambda$ bias	$\log\lambda$ RMSE	β_1 bias	β_1 RMSE
Main	conditional likelihood	+0.001	0.064	+0.001	0.059
Main	joint, correct ρ	+0.000	0.062	+0.001	0.055
Main	joint, misspecified ρ	+0.464	0.469	+0.171	0.175
Complete ($\rho \equiv 1$)	full censored-data likelihood	—	0.024	—	0.024
Complete ($\rho \equiv 1$)	conditional likelihood	—	0.052	—	0.049

Note. $n = 3{,}000$; 5,000 replications. In the complete labelling rows, bias is omitted and the dash indicates that only RMSE is reported.

Table S3: Consistency and coverage of conditional likelihood

n	$\log\lambda$ bias	$\log\lambda$ RMSE	$\log\lambda$ cov_{95}	β_1 bias	β_1 RMSE	β_1 cov_{95}
500	0.011	0.162	0.946	0.021	0.155	0.945
1,000	0.004	0.110	0.948	0.009	0.103	0.954
3,000	0.000	0.062	0.951	0.001	0.058	0.946

Note. All sample sizes use 5,000 replications; the Monte Carlo standard error of each coverage entry is about 0.003.

Table S4: Bias under dependent censoring

ϱ_{TC}	$\log\lambda$ bias	$F_T(20 \mid x_0)$ bias	$\pi(x_0)$ bias
0.0	+0.001	0.000	0.000
0.2	+0.192	−0.081	−0.067
0.4	+0.412	−0.169	−0.151

Note. The data use a Gaussian copula with conditional correlation ϱ_{TC}; $n = 3{,}000$, 5,000 replications. The proposed estimator is fitted at $\gamma = 0$.

(bias +0.010, RMSE 0.019, against +0.000 and 0.010 over 5,000 replications at $n = 3{,}000$).

Table S3 reports consistency and coverage of the conditional likelihood estimator in the main design: bias and RMSE fall with n, and the 95% Wald coverage from the inverse observed information of Theorem 3 is within sampling error of the nominal level for $\log\lambda$ and β_1 at all sample sizes.

Wald and delta-method coverage for β_2, $F_T(20 \mid x_0)$, and $\pi(x_0)$ ranges from 0.93 to 0.96 at $n = 500$, 1,000, and 3,000; these additional targets are not tabulated.

Dependent censoring. Breaking Assumption 1 with a Gaussian copula of correlation ϱ_{TC} between T and C given X, the absolute bias of the $\gamma = 0$ estimator increases with ϱ_{TC} (Table S4). The results illustrate the sensitivity of the estimator to violations of conditional independent censoring.

Table S5 ($n = 2{,}000$) repeats the main experiment under four distributions of C, including

Table S5: Bias under non-uniform censoring-time distributions

		Conditional likelihood		Positive-only
Distribution of C	$\mathbb{E}[F_T(C \mid X)]$	$\log\lambda$	$F_T(20 \mid x_0)$	$\log\lambda$
Uniform	0.64	0.001	0.000	−0.276
Calendar (beta)	0.75	0.004	−0.001	−0.246
Seasonal (mixture)	0.62	0.002	−0.001	−0.261
Age-dependent entry	0.63	0.004	−0.001	−0.340

Note. n = 2,000 and 1,000 replications per distribution of C; Monte Carlo standard errors for the PU $\log\lambda$ biases are at most 0.003.

a covariate-dependent one. Assumption 1 restricts only the dependence between T and C, not the shape of G_C; accordingly the proposed estimator is essentially unbiased in every case, while the positive-only estimator is biased throughout, confirming that the uniform distribution of C used elsewhere is a convenient illustration rather than a requirement.

A further misspecification experiment keeps the main design's labelling model and censoring distribution but generates event times from a gamma-baseline proportional-hazards model with baseline shape 2 and scale 14, so that the fitted Weibull family is misspecified. Over 1,000 replications at n = 2,000, the Weibull conditional likelihood has bias 0.001 (RMSE 0.027) for $F_T(20 \mid x_0)$ and bias 0.018 (RMSE 0.026) for $\pi(x_0)$; the positive-only Weibull fit has biases 0.087 and 0.074, and the correct-family gamma full censored-data MLE 0.000 and 0.000 (RMSE 0.013 and 0.009). In this design, the Weibull conditional likelihood estimator has smaller absolute bias than the positive-only Weibull fit for both targets.

Figure 1 traces the weak-identification mechanism of Proposition 4 by varying $c_{\max}$ in $C \sim \text{Uniform}[2, c_{\max}]$ so that $\mathbb{E}[F_T(C \mid X)]$ increases from 0.13 to 0.68 (n = 3,000, 5,000 replications). As $\mathbb{E}[F_T(C \mid X)]$ increases, the bias of the proposed estimator in $\log\lambda$ falls from very large values to essentially zero, and the median observed-Hessian standard error among fits with invertible information decreases. The positive-only estimator's bias persists throughout. The bias and standard-error panels aggregate the replications in which the observed information was invertible, a fraction 0.88 at the smallest value of $\mathbb{E}[F_T(C \mid X)]$ and 1.00 from the fourth value onwards; the dropped replications are themselves a symptom of weak identification, which is why the coverage panel is computed unconditionally. The profile (likelihood-ratio) interval for $\log\lambda$ remains defined when the information degenerates and, as identification weakens, is permitted to be unbounded. Panel C reports its coverage over all replications: at the smallest value of $\mathbb{E}[F_T(C \mid X)]$ more than half of the intervals are unbounded above, the finite-sample signature of a scale that the data cannot bound from above, and the lengths of the finite intervals shrink by more than an order of magnitude as $\mathbb{E}[F_T(C \mid X)]$ increases, so that the intervals become bounded and the coverage approaches the nominal level.

Figure S1 quantifies departures from Assumption 2 by generating labels from the sensitivity model of Remark 1 itself, $\mathbb{P}(S = 1 \mid T, C, X, Y = 1) \propto \rho(C, X) \exp\{\gamma a(T)\}$ with $a(t)$ the

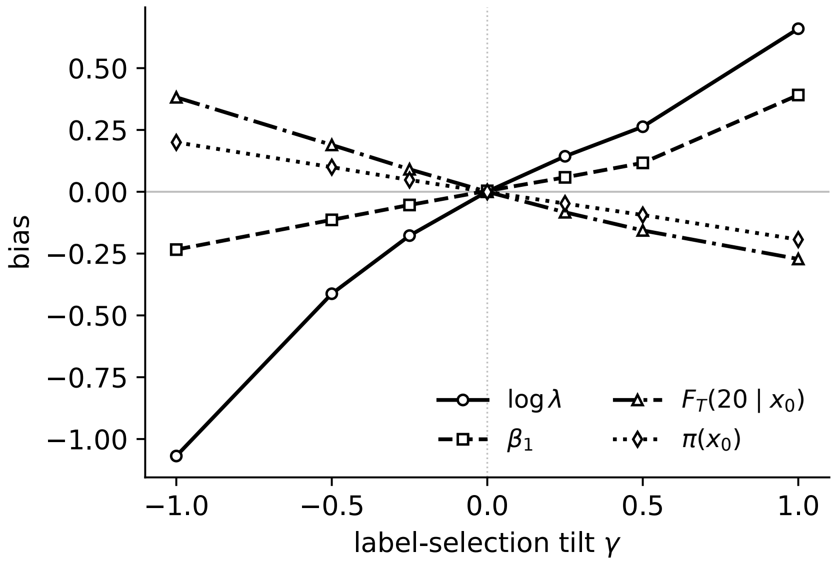

Figure S1: Bias of the $\gamma = 0$ estimator against the label-selection tilt γ for four targets.
Alt text: Line chart with label-selection tilt γ on the horizontal axis and bias on the vertical axis. Four curves show bias in $\log \lambda$, β_1, $F_T(20 \mid x_0)$, and $\pi(x_0)$. All four Monte Carlo biases are close to zero at $\gamma = 0$; absolute bias increases as γ moves away from zero, with $\log \lambda$ changing most strongly.

population-standardized log event time truncated to $[-3, 3]$ (normalized by $e^{-3|\gamma|}$ so the tilt is a valid probability pointwise), so that $\gamma \neq 0$ makes labelling depend on T given $(C, X, Y = 1)$ ($n = 3{,}000$, 5,000 replications per γ). The Monte Carlo bias of the $\gamma = 0$ estimator is negligible at $\gamma = 0$ and increases in absolute value with $|\gamma|$, reaching 0.09 in F_T and 0.05 in π at $\gamma = -0.25$. The tilt also shrinks the labelled sample, from about 1,100 positives per replication at $\gamma = 0$ to between 50 and 130 at $|\gamma| = 1$, and the 6% of replications at $\gamma = 1$ with a degenerate fit are excluded.

Binary positive-unlabelled comparison. To compare the proposed estimator with positive-unlabelled classification, we use the canonical binary estimator of Elkan and Noto (2008), which ignores event times and targets only the class prior (Table S6). The binary formulation cannot produce the event-time distribution $F_T(t \mid x)$ under an unrestricted labelling mechanism. Let $e(x) = \mathbb{P}(Y = 1 \mid X = x)$ and $g(x) = \mathbb{P}(S = 1 \mid X = x)$. Under constant labelling $c = \mathbb{P}(S = 1 \mid Y = 1)$, $g(x) = ce(x)$, while the population limit of the *e1* estimator is

$$\mathbb{E}\{g(X) \mid S = 1\} = c\,\mathbb{E}\{e(X) \mid Y = 1\}.$$

Unless $e(X) = 1$ almost surely among event subjects, this limit is below c, so dividing the labelled prevalence by the *e1* estimate yields an asymptotically inflated class prior. This is a property of the *e1* functional in this design, not the finite-sample absence of subjects with $e(X) = 1$. Under (C, X)-dependent labelling, the scalar identity $g(x) = ce(x)$ itself fails. The proposed estimator does not require the labelling constant: it recovers π with negligible bias in

Table S6: Class-prior performance of binary PU and PU survival estimators

Labelling regime	Estimator	$\pi(x_0)$	π (marginal)	$F_T(t \mid x)$
SCAR-like (constant ρ)	Binary PU, *e1* constant	0.296 (0.296)	0.254 (0.255)	not estimable
	Binary PU, known c	−0.019 (0.033)	−0.001 (0.018)	not estimable
	Conditional likelihood	−0.001 (0.025)	−0.001 (0.026)	estimated
Non-SCAR (ρ on C, X)	Binary PU, *e1* constant	0.170 (0.172)	0.178 (0.178)	not estimable
	Binary PU, known c	−0.082 (0.085)	−0.007 (0.015)	not estimable
	Conditional likelihood	−0.001 (0.022)	−0.002 (0.022)	estimated

Note. $n = 3{,}000$ and 1,000 replications; truth $\pi(x_0) = 0.701$ and marginal $\pi \approx 0.638$.

both regimes from the conditional likelihood and the design distribution of C, and estimates F_T, which the binary formulation cannot produce even when supplied with the true marginal labelling probability; Table S6 reports both the *e1* variant and the variant using the true c.

S6 AACT data construction and robustness

This section gives the data definition, model specification, and robustness results supporting Section 6. The primary endpoint is time to first results disclosure through either a ClinicalTrials.gov registry posting or a qualifying journal publication. For a labelled positive, the analysis uses the time to first results disclosure recorded in AACT, $\widetilde{T}$, and assumes $\widetilde{T} = T$ as detailed in Section S6.1. For an unlabelled trial with $Y = 1$, no results had been posted on ClinicalTrials.gov by time C, but a qualifying journal publication had appeared by time C and is absent from the retained publication links, so its publication date and T remain latent.

S6.1 Data construction and endpoint mapping

The source is the 1 May 2026 AACT monthly archive. From the AACT `studies` table we include interventional studies with `primary_completion_date_type=ACTUAL` and a Primary Completion Date from 1 January 2008 through 30 April 2026. AACT defines Primary Completion Date as the final data-collection date for the primary outcome. The `ACTUAL` restriction applies only to this time origin. It neither substitutes Study Completion Date nor restricts `results_first_posted_date` to its `ACTUAL` type. The initial cohort contains 265,106 trials; 217 have a `results_first_posted_date` on or before Primary Completion Date and are excluded because no positive post-completion delay is defined, leaving 264,889.

For registry postings, time is measured from Primary Completion Date to `results_first_posted_date`. ClinicalTrials.gov defines that field as the date on which summary results first became publicly available after NLM quality-control review, rather than the sponsor's first-submission date. Among the 70,743 registry positives, 50,019 date-type flags are `ACTUAL` and 20,724 are `ESTIMATED`; in the cohort every posting dated on or before 9 February 2017 is `ESTIMATED` and every posting from 10 February 2017 onwards is `ACTUAL`, a single cutover consistent with a legacy recording convention. Both are kept because the endpoint is the posted date recorded in `results_first_posted_date`, and the flag is preserved in the derived cohort for audit. The resulting time therefore includes any submission-to-posting quality-control lag.

For journal publications, the date is obtained from AACT `study_references`. A `RESULT` record is a results publication designated by the responsible party, and a `DERIVED` record is linked automatically from PubMed by the trial's ClinicalTrials.gov identifier. Automatic linkage requires the identifier to appear in the PubMed record, so the linkage is incomplete (Holt et al., 2025). We parse online-first and issue dates, use the earlier available date, and merge duplicate records. We then apply a fixed rule using PubMed Publication Types retrieved on 25 August 2026 (NLM, 2026) and AACT citation text. A link is excluded if its Publication Type is `Clinical Trial Protocol`, `Review`, `Systematic Review`, `Scoping Review`, `Meta-Analysis`, `Network Meta-Analysis`, `Editorial`, or `Comment`. It is also excluded if its citation contains `study protocol`, `trial protocol`, `statistical analysis plan`, `design and rationale`, `rationale and design`, `systematic review`, `scoping review`, `meta-analysis`, or `meta analysis`, after Unicode NFKC normalization, conversion of dashes to hyphens, whitespace collapsing, and case folding. Otherwise a `RESULT` link is retained. A `DERIVED` link is retained only if its Publication Type is `Adaptive Clinical Trial`, `Clinical Trial` or a Phase I–IV form of that type, `Controlled Clinical Trial`, `Equivalence Trial`, `Pragmatic Clinical Trial`, or `Randomized Controlled Trial`. Missing PubMed metadata therefore leaves a `DERIVED` link unretained but does not exclude a `RESULT` link; an unretained link is not evidence of non-disclosure. The candidate links and publication dates remain those in the 1 May AACT archive; the later PubMed retrieval supplies only their classification. A retained date must be strictly after Primary Completion Date and no later than the snapshot. This gives 62,435 publication positives. Both a registry posting and a linked publication occur for 23,158 trials, and their union contains 110,020 positives.

For this endpoint, T is the time to the first qualifying results disclosure after Primary Completion Date through a registry posting or a qualifying journal publication, whether or not a publication is linked in AACT. For $S = 1$, let $\widetilde{T}$ be the time to the first registry posting or linked publication included in the analysis. Under the specificity condition, the records used in the analysis satisfy $S \leq Y$, but this condition does not by itself imply $\widetilde{T} = T$. The AACT conditional likelihood analysis assumes $\widetilde{T} = T$ whenever $S = 1$. Equivalently, the first qualifying results disclosure is included in the records used for the analysis and its disclosure date is correctly ascertained. If an earlier qualifying publication is not linked in AACT, then $\widetilde{T} > T$. Error due to a missing link to an earlier publication is therefore one-sided. The γ-analysis perturbs the labelled-positive conditional density by event time while maintaining the equality $\widetilde{T} = T$. Without this equality, the fitted conditional likelihood need not recover the distribution of the true time to first results disclosure T; instead, it describes the distribution of $\widetilde{T}$ among labelled positives.

We also recomputed the union using issue dates only, the lower and upper endpoints of publication date intervals, and the lower, midpoint, and upper feasible dates for candidates whose intervals cross the Primary Completion Date or snapshot boundary. Across these seven definitions, the number of labelled positives in the union ranged from 110,020 to 110,354, the observed fraction from 0.415 to 0.417, the fitted conditional likelihood scale from 2.82 to 2.94 years, and the excess of that scale over the positive-only scale from 15.3% to 15.7%. These variants change publication-date assignment but do not address qualifying publications absent from the retained links, and therefore do not relax the assumption $\widetilde{T} = T$ for labelled positives.

For the stated disclosure endpoint, $Y = 1, S = 0$ corresponds to a qualifying journal publication that appeared by time C but is absent from the retained publication links; its publication date and T are unobserved. The alternative $Y = 0, S = 0$ corresponds to no qualifying disclosure by C.

The explicit lead sponsor comes from the AACT `sponsors` table row with `lead_or_collaborator=lead`; `agency_class` supplies the sponsor category. The 264,889-trial cohort contains 74,433 industry sponsors and 11,576 government sponsors after grouping `NIH`, `FED`, and `OTHER_GOV`; the remaining 178,880 trials form the reference category. Trial phase is kept in the registry's categories, with `EARLY_PHASE1` grouped with Phase 1. The parametric regression also includes completion year centred at 2016.

S6.2 Parametric analysis of time to first results disclosure recorded in AACT

The pooled Weibull conditional likelihood for the 110,020 labelled positives gives shape 1.344 and scale 2.843 years; the positive-only fit has shape 1.325 and scale 2.467 years. Because disclosure timing and administrative follow-up both vary with completion year, the principal covariate regression adjusts for completion year directly.

The Weibull regression uses all 103,015 labelled positives completed through 2023. It estimates a common shape of 1.447 and a scale of 2.668 years for a Phase 2 trial with another lead sponsor completed in 2016. The completion-year time ratio is 0.951 per year, with a 95% sandwich interval from 0.950 to 0.952. The substantive phase and sponsor contrasts are reported in Table 2.

The completion-year coefficient is a working log-linear adjustment for secular change. Centring changes only the reference value of the scale and does not change fitted year-to-year contrasts. Because administrative follow-up is deterministically related to completion date, this adjustment does not by itself ensure conditional independence. The specification analysis below examines the linear form and uses common follow-up windows to control administrative follow-up by design.

Table S7 re-estimates the regression under three further specifications. The first replaces the linear term by a restricted cubic spline in completion year with knots at 2010, 2014, 2018, and 2022; every phase and sponsor time ratio changes by less than one per cent. The remaining two use common follow-up windows: trials with at least five, respectively three, years of potential follow-up are included and follow-up is administratively capped at that horizon, the capped experiment of Lemma S1 with a truncation point common to every included trial; the labelled positives are those with a disclosure recorded within the window, and the centred completion-year covariate is kept because window inclusion still selects on completion date. Within the included cohort the common truncation point removes completion-date variation in administrative follow-up; it does not remove other calendar or cohort sources of dependence between the disclosure and censoring times. The Phase 2/3 point estimate moves just above one only in the three-year window. The Phase 1/2 and industry estimates remain close to one, with intervals containing one in at least one window; the other contrasts retain their directions.

The Phase 3 ratio attenuates from 0.824 in the full design to 0.871 at five years and 0.925 at three years, while the fitted common shape rises from 1.447 to 1.662 and 1.968. The broad pattern of the phase and sponsor contrasts is similar across the specifications. The common follow-up windows do not test Assumption 1 in general.

Table S7: Time ratios for first results disclosure recorded in AACT across calendar and follow-up specifications

Contrast	Main	Spline	Five-year window	Three-year window
Phase 1 vs Phase 2	1.156 $[1.132, 1.180]$	1.165 $[1.141, 1.190]$	1.169 $[1.140, 1.199]$	1.210 $[1.176, 1.246]$
Phase 1/2 vs Phase 2	1.033 $[1.004, 1.064]$	1.038 $[1.008, 1.069]$	1.008 $[0.977, 1.039]$	1.010 $[0.980, 1.041]$
Phase 2/3 vs Phase 2	0.934 $[0.898, 0.972]$	0.937 $[0.900, 0.976]$	0.953 $[0.914, 0.993]$	1.001 $[0.959, 1.045]$
Phase 3 vs Phase 2	0.824 $[0.809, 0.839]$	0.825 $[0.810, 0.841]$	0.871 $[0.855, 0.887]$	0.925 $[0.908, 0.942]$
Phase 4 vs Phase 2	0.874 $[0.856, 0.892]$	0.875 $[0.857, 0.893]$	0.886 $[0.868, 0.905]$	0.920 $[0.901, 0.939]$
Phase not applicable vs Phase 2	1.046 $[1.031, 1.062]$	1.052 $[1.036, 1.069]$	1.044 $[1.026, 1.061]$	1.072 $[1.054, 1.091]$
Industry vs other lead sponsor	0.986 $[0.974, 0.999]$	0.983 $[0.971, 0.996]$	0.970 $[0.957, 0.984]$	0.998 $[0.985, 1.013]$
Government vs other lead sponsor	0.865 $[0.844, 0.885]$	0.860 $[0.840, 0.881]$	0.904 $[0.881, 0.926]$	0.947 $[0.922, 0.971]$

Note. Entries are time ratios with 95% sandwich intervals in square brackets. The Main column repeats the specification of Table 2; the spline column replaces the linear completion-year term by a restricted cubic spline with knots at 2010, 2014, 2018, and 2022; the window columns include trials with at least five or three years of potential follow-up, cap follow-up at that horizon, and use the 75,185 and 73,740 labelled positives with a disclosure recorded within the window. Completion year, centred at 2016, is included in every specification but is not displayed.

S6.3 Treating S as the event indicator

To mirror the S-as-event comparator in Section 5, we also fit the conventional right-censored Weibull likelihood

$$\ell_{S=Y}(\theta) = \sum_i \Big[S_i \log f_\theta(\widetilde{T}_i \mid X_i) + (1 - S_i) \log\{1 - F_\theta(C_i \mid X_i)\}\Big]$$

with the same completion-period restriction and covariates. This fit uses all 227,779 trials completed through 2023, treating 124,764 unlabelled trials as ordinary right-censored observations. For a trial with $Y_i = 1$ and $S_i = 0$, the first results disclosure is a qualifying journal publication that is absent from the retained, dateable AACT publication links, and no registry posting occurred by time C_i; the fit treats such a trial as right-censored at C_i. It analyses the endpoint defined by a registry posting or linked publication recorded in AACT, which differs from the stated disclosure endpoint.

The resulting shape is 0.788 and the 2016 reference scale is 9.304 years, compared with 1.447 and 2.668 under the conditional likelihood. Table S8 gives the covariate contrasts. Assessing empirical bias relative to the stated disclosure endpoint would require the missing publication outcomes.

Table S8: Conditional likelihood and S-as-event time ratios

Contrast	Conditional likelihood	S-as-event full likelihood
Phase 1 vs Phase 2	1.156	4.504
Phase 1/2 vs Phase 2	1.033	1.251
Phase 2/3 vs Phase 2	0.934	1.351
Phase 3 vs Phase 2	0.824	0.824
Phase 4 vs Phase 2	0.874	1.238
Phase not applicable vs Phase 2	1.046	2.462
Industry vs other lead sponsor	0.986	0.622
Government vs other lead sponsor	0.865	0.621
Completion year, per year	0.951	0.932

Note. Both fits use a common Weibull shape and completion year centred at 2016.

S6.4 Sensitivity analysis for the labelling mechanism

Throughout this sensitivity analysis, we assume $\widetilde{T} = T$ for labelled positives. For the pooled analysis, we use the exponential tilt model of Remark 1 with $a(t, c, x) = a(t)$, where

$$a(t) = \frac{\log t - \bar{\ell}}{s_\ell}, \qquad \bar{\ell} = \frac{1}{n_1} \sum_{i:S_i=1} \log \widetilde{T}_i, \qquad s_\ell^2 = \frac{1}{n_1} \sum_{i:S_i=1} \{\log \widetilde{T}_i - \bar{\ell}\}^2.$$

Here $n_1 = 110{,}020$, and t and $\widetilde{T}_i$ are measured in years. The centre and scale are computed once from all labelled positives in the pooled analysis and held fixed for all values of γ; $a(t)$ is not truncated. For the Weibull model with shape $a > 0$, the normalizing integral is finite if and only if $a + \gamma/s_\ell > 0$. We impose this constraint in estimation. Because $a(t)$ is unbounded below as $t \downarrow 0$, negative values of γ are interpreted here as exponential tilts of the labelled-positive conditional density rather than as a globally bounded model for the labelling probability. The publication rule and the specificity condition are held fixed. Table S9 reports the pooled fits.

S6.5 The registry-posting benchmark

For registry posting, the by-snapshot event indicator is completely observed, so the 194,146 trials without a registry posting are valid censored observations for that endpoint. Table S10

Table S9: Sensitivity scenarios for the labelling mechanism in the pooled AACT analysis

Scenario	γ range	Fitted scale (years)
Benchmark	$\gamma = 0$	2.84
Mild	$[-0.15, 0]$	2.84–3.21
Central	$[-0.20, 0]$	2.84–3.33
Wide	$[-0.30, 0]$	2.84–3.56

Note. The scenarios tilt the labelled-positive conditional density by event time in the pooled working model; their γ endpoints are specified by the analyst.

compares the conditional timing fit with full censored-data models. The cure fit and the conditional likelihood give similar finite-delay timing distributions, while the proper full-likelihood model assumes that every trial eventually posts and yields a much longer scale estimate. The cure incidence applies only to registry posting and has no interpretation as the incidence of results disclosure through either reporting route.

Table S10: Conditional and full likelihood fits for AACT registry-posting times

Fit	Uses unposted trials	$\hat{a}$	$\hat{\lambda}$	$G_{\text{post}}(1)$	Incidence
Positive-only fit	no	1.315	2.820	0.226	not estimated
Conditional likelihood	no	1.351	3.414	0.173	not identified
Full likelihood, proper	yes	0.738	34.381	0.071	1 imposed
Full likelihood, cure mixture	yes	1.321	3.314	0.186	0.328 (s.e. 0.001)

Note. Time is in years. $G_{\text{post}}(1)$ is the fitted one-year probability conditional on eventual registry posting.

S6.6 Cox-type registry-posting analysis

We fit the Cox-type proportional reverse-hazards estimator to the 70,743 registry-posting positives. Under $F_T(t \mid x) = F_0(t)^{\exp(x'\beta)}$, a reverse-hazard ratio above one shifts the distribution towards longer posting delays and is not a forward-time hazard ratio. Table S11 uses phase, explicit AACT lead-sponsor class, and completion year centred at 2016, with Phase 2 and other sponsors as references.

The fit describes posting timing among trials that post. By the level invariance of the partial likelihood, it does not estimate the probability of ever posting. Because AACT records posting at day level, the implementation uses the Breslow convention for tied days and is fitted over the full observed support; Proposition S4 is a continuous-time result with a fixed interior horizon τ, and neither the grouped-time convention nor the reported observed-information standard errors are covered by it.

Table S11: Cox-type proportional reverse-hazards fit for AACT registry-posting times

Covariate	$\hat{\beta}$	s.e.	reverse-HR	z
Phase 1	+0.108	0.015	1.11	+7.0
Phase 1/2	+0.018	0.018	1.02	+1.0
Phase 2/3	−0.025	0.029	0.98	−0.9
Phase 3	−0.152	0.012	0.86	−12.5
Phase 4	−0.212	0.014	0.81	−15.5
Phase not applicable	+0.030	0.010	1.03	+3.0
Sponsor: industry	−0.013	0.009	0.99	−1.5
Sponsor: government	−0.045	0.017	0.96	−2.7
Completion year (per year)	−0.033	0.001	0.97	−36.7

Note. Standard errors are from the partial-likelihood information. Phase 2 and lead sponsors other than industry or government are references.

S6.7 Comparison of the two reporting routes

Among the 23,158 trials with both a registry posting and a linked publication, the publication date precedes the registry-posting date in 55.6% of comparisons after excluding ties. The median absolute difference between the two dates is 1.02 years. The observed date difference is descriptive, not an estimate or an upper bound for $\widetilde{T} - T$, because earlier unlinked publications are unobserved.

Within the 62,435 trials with a linked publication, we residualize log publication delay on C, phase and lead-sponsor class in their original AACT categories, and completion year. A logistic regression of the registry-posting indicator on the standardized residual and the same controls gives $\hat{\delta}_{\text{pub}} = -0.165$ ($z = -17.6$); the registry-posting rate falls from 0.42 to 0.28 across residual-delay deciles (Figure S2). This is not a test of Assumption 2: it uses registry posting among trials with a linked publication, whereas the structural analysis concerns labelling through either reporting route and the time to first results disclosure T. This association motivates examining negative values of γ but does not identify a value or range for the sensitivity parameter.

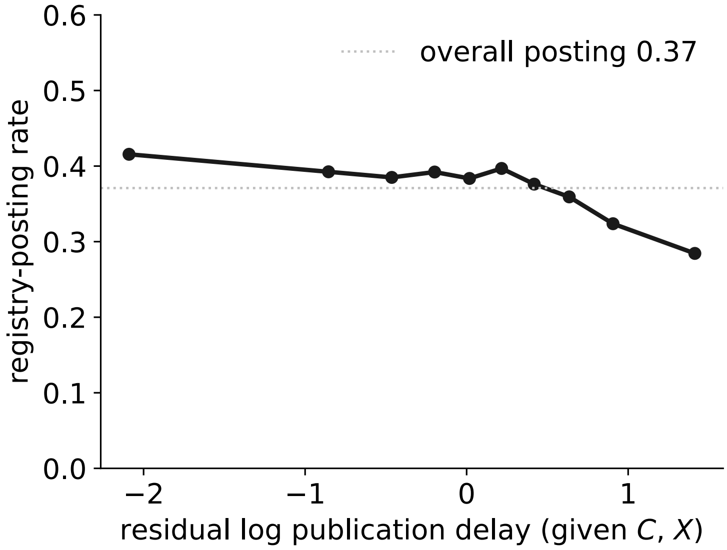

Figure S2: Registry-posting rate by residual publication-delay decile among trials with a linked publication.

Alt text: Line chart with standardized residual log publication delay on the horizontal axis and registry-posting rate on the vertical axis. The rate generally declines from about 0.42 in the shortest-delay decile to 0.28 in the longest-delay decile; a horizontal line marks the overall rate of about 0.37.

S7 Extensions

S7.1 Partially observed censoring times among unlabelled records

Suppose that C is observed for every labelled positive but may be missing for unlabelled records. The conditional density $q(t \mid c, x) = f_T(t \mid x)/F_T(c \mid x)$ and the relative event-time distribution remain identified as in Theorems 1 and 2. Once the terminal level of F_T is fixed, the event-before-censoring probability is also identified if $G_C(\cdot \mid X)$ is known by design or identified from an independent source $J = 1$ satisfying

$$C \mid X, J = 0 \overset{d}{=} C \mid X, J = 1, \qquad \mathbb{P}(X \in \cdot \mid J = 0) \ll \mathbb{P}(X \in \cdot \mid J = 1),$$

where $J = 0$ is the positive-unlabelled source. Alternatively, in the single-cohort frame, let R_C indicate recovery of C for an unlabelled record. If

$$\mathbb{P}(R_C = 1 \mid S = 0, C, X) = \pi_C(X), \qquad \pi_C(X) \geq \epsilon_C > 0,$$

then, for every integrable h,

$$\mathbb{E}\{h(C) \mid X\} = \mathbb{E}\left[\left\{S + \frac{(1-S)R_C}{\pi_C(X)}\right\} h(C) \,\middle|\, X\right].$$

This is the usual inverse probability identity under missing at random (Tsiatis, 2006). In the case-population frame, G_C is recovered from the population sample alone, using $R_C/\pi_C(X)$ there if necessary. In the single-cohort frame, a weighted plug-in estimator of the marginal event-before-censoring probability averages $\{S_i + (1 - S_i)R_{Ci}/\widehat{\pi}_C(X_i)\}F_{\hat{\theta}}(C_i \mid X_i)$ over the cohort. In the case-population frame, the analogous average uses $R_{Ci}/\widehat{\pi}_C(X_i)$ within the population sample. Neither this recovery nor an external source observing only (C, X) fixes the terminal level of F_T, and the efficiency results in Section 4 continue to assume fully observed C.

S7.2 Extensions of conditional likelihood

(a) *Missing censoring times.* If C is missing for labelled positives, the conditional likelihood is generally unavailable. Recovery requires stronger conditions, such as selection independent of C given $(X, Y = 1)$, known labelling probabilities, or validation data on C among positives.

(b) *Interval-censored positive event times.* An event in $(L, R]$, $R \leq C$, contributes $\{F_T(R \mid x) - F_T(L \mid x)\}/F_T(c \mid x)$. The same conditioning gives labelling invariance, but the exact-time estimation and efficiency results do not cover this likelihood.

(c) *Ties and discretely recorded times.* For grid points $\{u_k\}$ and $u_{k-1} < c$, the grouped contribution is $\{F_T(\min\{u_k, c\} \mid x) - F_T(u_{k-1} \mid x)\}/F_T(c \mid x)$, and is zero when $c \leq u_{k-1}$. It inherits labelling invariance; ties in the reverse-time partial likelihood are handled by the Breslow convention.

References

Andersen, P. K. and Gill, R. D. (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics*, Vol. 6B, eds. J. J. Heckman and E. E. Leamer, Amsterdam: Elsevier, pp. 5549–5632.

de Boor, C. (2001), *A Practical Guide to Splines*, revised ed., Applied Mathematical Sciences, Vol. 27, New York: Springer.

Elkan, C. and Noto, K. (2008), "Learning Classifiers from Only Positive and Unlabeled Data," *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 213–220.

Holt, A. M., Troy, A. M. and Smalheiser, N. R. (2025), "Distribution of Trial Registry Numbers within Full-Text of PubMed Central Articles: Implications for Linking Trials to Publications and Indexing Trial Publication Types," *Trials*, 26, 34.

Huang, J. Z. (2003), "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600–1635.

Kalbfleisch, J. D. and Lawless, J. F. (1991), "Regression Models for Right Truncated Data with Applications to AIDS Incubation Times and Reporting Lags," *Statistica Sinica*, 1, 19–32.

Lagakos, S. W., Barraj, L. M. and De Gruttola, V. (1988), "Nonparametric Analysis of Truncated Survival Data, with Application to AIDS," *Biometrika*, 75, 515–523.

Lynden-Bell, D. (1971), "A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars," *Monthly Notices of the Royal Astronomical Society*, 155, 95–118.

Maller, R. A. and Zhou, X. (1996), *Survival Analysis with Long-Term Survivors*, New York: Wiley.

National Library of Medicine (2026), "PubMed User Guide: Publication Types," `https://pubmed.ncbi.nlm.nih.gov/help/`. Accessed 25 August 2026.

Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591.

Shen, X. and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer.

van der Vaart, A. W. (1991), "On Differentiable Functionals," *The Annals of Statistics*, 19, 178–204.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge: Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (2023), *Weak Convergence and Empirical Processes: With Applications to Statistics*, 2nd ed., Cham: Springer.

Woodroofe, M. (1985), "Estimating a Distribution Function with Truncated Data," *The Annals of Statistics*, 13, 163–177.