Nonparametric analysis of delayed treatment effects using single-crossing constraints

Nicholas C. Henderson | Kijoeng Nam | Dai Feng

1Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
2BARDS, Merck & Co., Inc., North Wales, Pennsylvania, USA
3Data and Statistical Sciences, AbbVie Inc., North Chicago, Illinois, USA

Abstract
Clinical trials involving novel immuno-oncology therapies frequently exhibit survival profiles which violate the proportional hazards assumption due to a delay in treatment effect, and, in such settings, the survival curves in the two treatment arms may have a crossing before the two curves eventually separate. To flexibly model such scenarios, we describe a nonparametric approach for estimating the treatment arm-specific survival functions which constrains these two survival functions to cross at most once without making any additional assumptions about how the survival curves are related. A main advantage of our approach is that it provides an estimate of a crossing time if such a crossing exists, and, moreover, our method generates interpretable measures of treatment benefit including crossing-conditional survival probabilities and crossing-conditional estimates of restricted residual mean life. Our estimates of these measures may be used together with efficacy measures from a primary analysis to provide further insight into differences in survival across treatment arms. We demonstrate the use and effectiveness of our approach with a large simulation study and an analysis of reconstructed outcomes from a recent combination therapy trial.

KEYWORDS
censored data, clinical trial, constrained estimation, immuno-oncology, nonproportional hazards

1 INTRODUCTION
Recent advances in immuno-oncology (IO) therapies for the treatment of cancer have led to the development of a variety of treatments which show great potential for improving long-term patient outcomes. While very promising, patient response to such immunotherapies is often quite different when compared to more traditional cytotoxic agents such as chemotherapy. Indeed, it is well-recognized that IO drugs frequently exhibit a clear delay in treatment effect when they are compared with standard chemotherapies, and, moreover, the nature of the delay in this treatment effect is often such that the estimated survival curves in the two treatment arms have a crossing at some time point after randomization. Due to this feature of immunotherapies, traditional comparisons between IO drugs and chemotherapies can have a number of limitations.
For time-to-event endpoints in randomized clinical trials, the log-rank test is a conventional choice for comparing survival curves in an active treatment arm versus a control arm, and whenever the log-rank test passes a threshold for statistical significance, a hazard ratio is often reported as one of the chief measures of treatment efficacy. However, in settings with survival curve crossings where the proportional hazards assumption is plainly violated, the interpretation of both the log-rank test and estimated hazard ratio from a Cox proportional hazards model could be unclear. A variety of alternatives to the log-rank test and the hazard ratio have been suggested in the context of evaluating the efficacy of immunotherapies or other contexts where the proportional hazards assumption regularly fails. In the context of improving the power of testing the equality of the arm-specific survival curves under possible delayed treatment effects, a range of weighted log-rank tests and related tests have been proposed in the statistical literature. These include, for example, the piecewise weighted log-rank test (APPLE) (Xu et al., 2017) and weighted log-rank tests from the Fleming–Harrington family which place more weight on later time points (Harrington & Fleming, 1982; Rahman et al., 2019). Combination tests which combine multiple weighted log-rank test statistics have also been proposed for analyzing IO trials (Lin et al., 2020). In addition to alternatives to the unweighted log-rank test, a number of treatment effect measures have been explored and evaluated as alternatives to hazard ratios in the delayed treatment context. These include, for example, differences in restricted mean survival time (RMST; Pak et al., 2017; Zhao et al., 2016), differences in milestone survival and reporting such measures of treatment efficacy can certainly be valuable, imposing estimation constraints on how the survival curves may cross can provide additional information by which to evaluate a treatment that exhibits delayed benefit. To this end, in this paper, we both propose a nonparametric estimation approach for the treatment-specific survival functions that allows for at most one crossing of the two survival functions without requiring that the crossing time be prespecified, and we propose several efficacy measures which utilize the crossing-time parameters to provide greater context for interpreting the impact of the active treatment on patient outcomes. The approach we describe for estimating the survival curves under a single-crossing constraint can be thought of as a two-stage procedure. In the first stage, one finds conditional estimates of the survival function by maximizing a nonparametric log-likelihood that is conditional on the value of two crossing parameters. Then, in the second stage, one estimates these crossing parameters by maximizing a profile log-likelihood function. This produces estimates of the two survival curves which satisfy the single-crossing constraint, and it generates an estimate of the crossing time and an estimate of which survival curve shows better long-term survival. While flexible semiparametric methods can allow for crossings of the treatment-specific survival function, such methods will entail certain assumptions about how the active treatment arm modifies the baseline survival function. In contrast, our approach makes no assumptions about how the two survival curves are related other than that they cross at most once. Though a number of available methods target inference for the crossing times of hazards with a focus on the time of first crossing without necessarily imposing any constraints on the number of crossings, our approach focuses primarily on the problem of crossing survival curves under a single-crossing constraint without assuming a prespecified crossing time, and we provide a likelihood-based estimate of the crossing time if a positive crossing time is determined to be present. In addition to this, we show in Section 2.4 how our approach can be directly extended to handling a single-crossing constraint on the hazard functions if the hazard constraint seems more appropriate in a particular application.

One of the advantages of imposing a single-crossing constraint is that scenarios with a delay in treatment effect and improved long-term survival in the active treatment arm are often better modeled as having a single crossing after which the survival curve in the active treatment arm consistently dominates the survival curve of the control arm. Such patterns of delayed treatment effect have been observed in many recent IO trials, and hence, constrained modeling in such trials has the potential to improve estimation performance and provide greater interpretability when comparing the two arm-specific estimates of the survival functions. An important advantage of using single-crossing constraints is that it yields estimates and uncertainty intervals for the time that the survival curves cross. The estimate of the crossing-time parameter can be useful in both assessing when the active treatment begins to show superiority and as an interpretable measure that can be used in conjunction with other measures of treatment efficacy that are designed for scenarios with delayed treatment effects. Making inferences on these additional measures can be valuable for gaining further insights about key questions of interest and for providing a more complete picture of differences in survival across the two treatment arms. In Section 3, we outline several of these additional measures of efficacy and discuss their interpretation. These efficacy measures include the proportion of patients surviving up to crossing, crossing-time conditional restricted residual mean life, crossing-time conditional survival probabilities, and pre-/postcrossing average hazard ratios.
This paper is organized as follows. In Section 2, we describe a two-stage nonparametric estimation procedure which assumes that the treatment arm-specific survival curves cross at most once but otherwise makes no additional assumptions about the forms of the survival functions, and we briefly outline several potential extensions of this approach. Section 3 discusses a number of interpretable estimands of interest that can capture relevant concerns about the treatment impact in cases where survival curves may cross, and we describe how our method can be used to estimate these terms. Section 3 also describes hypothesis tests of interest under delayed treatment effects. Section 4 shows the results from a simulation study which evaluates the estimation performance of our method across six piecewise-exponential simulation settings. Section 5 shows an application of our method to a recent trial involving a novel combination immunotherapy in the treatment of non-small-cell lung cancer, and Section 6 concludes with a brief discussion.

2 | NONPARAMETRIC ESTIMATION FOR SURVIVAL CURVES WITH SINGLE CROSSINGS

2.1 | Survival curve profiles and notation

We assume that \( n \) patients have been enrolled in a randomized clinical trial consisting of two treatment arms. The primary outcome is the time to some event of interest, and for the \( i \)th individual in the study, we let \( T_i \) denote the time-to-failure for this event of interest. Instead of directly observing \( T_i \), we observe the follow-up time \( Y_i = \min\{T_i, C_i\} \) and the event indicator \( \delta_i = I(T_i \leq C_i) \), where \( C_i \) denotes the censoring time and \( I(\cdot) \) denotes the indicator function. We let \( A_i = 1 \) denote that patient \( i \) was assigned to the active treatment arm, and we let \( A_i = 0 \) denote that patient \( i \) was assigned to the control arm. We also assume that censoring is noninformative in the sense that the terms from the censoring distribution can be factored out of the likelihood function (Lawless, 2011), and we assume that \( C_i \) and the treatment arm assignment \( A_i \) are independent.

We let \( S_a(t) = P(T_i > t | A_i = a) \) denote the survival function for those patients assigned to treatment group \( a \in \{0, 1\} \). We consider an analysis where these survival functions are allowed to exhibit four distinct “survival profiles” according to whether or not the survival curves cross. Specifically, we allow for the possibility that the two survival functions \( S_0(t) \) and \( S_1(t) \) “cross,” but we limit the number of crossings so that they can occur at most once. The crossing restriction implies that we will have one of four survival profiles, where each profile refers to a distinct crossing pattern of the survival functions. These four possible survival profiles are as follows: (1) a survival profile where treatment arm \( a = 1 \) completely dominates \( a = 0 \), namely \( S_1(t) \geq S_0(t) \) for all \( t \geq 0 \); (2) a survival profile where \( a = 0 \) dominates \( a = 1 \) before some crossing time \( \theta \) but \( a = 1 \) dominates \( a = 0 \) afterwards, that is, \( S_0(t) \geq S_1(t) \) for \( 0 \leq t \leq \theta \) and \( S_1(t) \geq S_0(t) \) for \( t > \theta \); (3) a survival profile where \( a = 1 \) dominates \( a = 0 \) before some crossing time \( \theta \) but \( a = 0 \) dominates \( a = 1 \) afterwards; and (4) a survival profile where \( a = 0 \) completely dominates \( a = 1 \). Figure 1 illustrates each of these four possible survival profiles.

To enforce one of the four possible profiles implied by Figure 1, we introduce the two crossing parameters \( \theta \geq 0 \) and \( \gamma \in \{-1, 1\} \). The parameter \( \theta \) represents the crossing time of the survival functions, and the discrete parameter \( \gamma \) determines which treatment arm exhibits better long-term survival. Specifically, when \( \gamma = 1 \) the survival functions are assumed to exhibit the following behavior:

\[
S_0(t) \geq S_1(t) \quad \text{for } 0 \leq t \leq \theta \quad \text{and} \quad S_0(t) \leq S_1(t) \quad \text{for } t > \theta, \tag{1}
\]

and when \( \gamma = -1 \) the survival functions are assumed to obey the following inequalities:

\[
S_0(t) \leq S_1(t) \quad \text{for } 0 \leq t \leq \theta \quad \text{and} \quad S_0(t) \geq S_1(t) \quad \text{for } t > \theta. \tag{2}
\]

If there is a \( \theta > 0 \) such that either constraints (1) or constraints (2) hold, we will say that \( \theta \) is a nontrivial crossing time of the survival functions \( S_0 \) and \( S_1 \). If there is no such \( \theta > 0 \), we must either have \( S_0(t) \geq S_1(t) \) or \( S_1(t) \geq S_0(t) \) for all \( t \geq 0 \) in which case we set \( \theta \) to the trivial crossing time \( \theta = 0 \). Because we are assuming the survival curves follow one of the four profiles shown in Figure 1, we are excluding the possibility of multiple crossing times, and, hence, the crossing time parameter \( \theta \) represents the unique crossing time of the two survival curves whenever \( \theta > 0 \).

When constructing estimates of the survival functions, we can only enforce the infinite-dimensional constraints implied by (1) and (2) at a finite number of time points. In our implementation, we enforce the survival function constraints (1) and (2) at each of the observed event times. To this end, we let \( 0 < t_1 < t_2 < \cdots < t_m \) denote the collection of unique, ordered
event times from both treatment arms, and, hence, \( m \) represents the total number of unique event times occurring in either of the treatment arms. Note that \( m \) will equal the total number of events occurring in either treatment arm (i.e., \( \sum_{i=1}^{n} \delta_i \)) if there are no ties among the event times and will be less than \( \sum_{i=1}^{n} \delta_i \) if any ties occur. For \( \gamma = 1 \) and a fixed value of \( \theta \), we enforce the following constraints:

\[
S_0(t_j) \geq S_1(t_j) \quad \text{for all } j \text{ such that } t_j \leq \theta \\
S_0(t_j) \leq S_1(t_j) \quad \text{for all } j \text{ such that } t_j > \theta. 
\]  

Similarly, for \( \gamma = -1 \) and a fixed value of \( \theta \), we enforce the following constraints:

\[
S_0(t_j) \leq S_1(t_j) \quad \text{for all } j \text{ such that } t_j \leq \theta \\
S_0(t_j) \geq S_1(t_j) \quad \text{for all } j \text{ such that } t_j > \theta. 
\]  

Note that (3) and (4) together imply that we will have a different set of constraints for each choice of \((\theta, \gamma)\). In the next subsection, we describe our procedure for estimating the arm-specific survival functions when both \( \theta \) and \( \gamma \) are assumed to be fixed.

### 2.2 Estimating the survival functions with fixed crossing parameters

We first describe the nonparametric estimation of the survival functions \( S_0(t) \) assuming both \( \theta \) and \( \gamma \) are known. As detailed in Park et al. (2012) in the context of finding the constrained nonparametric maximum likelihood estimate of
survival functions under a stochastic ordering constraint, the constrained maximum likelihood estimates of the survival functions will be discrete with potential jumps only at the combined event times $t_1, \ldots, t_m$. Because of this, we define $u_{j,a}(\theta, \gamma)$ as the jump of $\log S_a(t)$ at the point $t_j$ when it is assumed that the true crossing-time parameter is $\theta$ and the true superior long-term survival parameter is $\gamma$. More specifically, $u_{j,a}(\theta, \gamma) = \log[S_a(t_j)] - \log[S_a(t_j)]$, where $S_a(t_j)$ denotes the left limit of $S_a(t)$ at time $t$. The term $h_{j,a}(\theta, \gamma) = 1 - \exp[u_{j,a}(\theta, \gamma)] = 1 - S_a(t_j)/S_a(t_j)$ can be interpreted as the discrete hazard at time point $t_j$. Following Park et al. (2012) and Johansen (1978), the log-likelihood function to be maximized in this context is

$$
\log L[\mathbf{u}_0(\theta, \gamma), \mathbf{u}_1(\theta, \gamma)] = \sum_{a=0}^{1} \sum_{j=1}^{m} [d_{ja} \log(1 - \exp(u_{j,a}(\theta, \gamma))) + (R_{ja} - d_{ja})u_{j,a}(\theta, \gamma)],
$$

where $\mathbf{u}_0(\theta, \gamma)$ and $\mathbf{u}_1(\theta, \gamma)$ are the $m \times 1$ column vectors having components $u_{j,0}(\theta, \gamma)$ and $u_{j,1}(\theta, \gamma)$, respectively. In (5), $R_{j0} = \sum_{i=1}^{n}(1 - A_i)I(Y_i \geq t_j)$ denotes the number of individuals at risk in the control arm at time $t_j$, and $R_{j1} = \sum_{i=1}^{n}A_iI(Y_i \geq t_j)$ denotes the number of individuals at risk in the active treatment arm at time $t_j$. In (5), $d_{j0} = \sum_{i=1}^{n}(1 - A_i)\delta_iI(Y_i = t_j)$ denotes the number of events in the control arm at time $t_j$ while $d_{j1} = \sum_{i=1}^{n}A_i\delta_iI(Y_i = t_j)$ denotes the number of events in the active treatment arm at time $t_j$.

In the absence of any crossing constraints on the survival functions, the only constraints on the vector $\mathbf{u}_a(\theta, \gamma)$ would be $u_{j,a}(\theta, \gamma) \leq 0$, for $j = 1, \ldots, m$. Because the survival functions $S_0(t)$ and $S_1(t)$ associated with the vectors $\mathbf{u}_0(\theta, \gamma)$ and $\mathbf{u}_1(\theta, \gamma)$ have the form $S_a(t) = \exp\{\sum_{j=1}^{m}u_{j,a}(\theta, \gamma)I(t_j \leq t)\}$, the crossing constraints (3) and (4) can be expressed as a collection of linear inequality constraints. Specifically, we can represent both the crossing constraints and the inequality constraints $u_{j,a}(\theta, \gamma) \leq 0$ as $[a_k(\theta, \gamma)]^T \mathbf{u}(\theta, \gamma) \geq 0$ for $k = 1, \ldots, 3m$, where $a_k(\theta, \gamma) \in \mathbb{R}^{2m}$ and $\mathbf{u}(\theta, \gamma) = (\mathbf{u}_0(\theta, \gamma)^T, \mathbf{u}_1(\theta, \gamma)^T)^T$. For $k \leq m$, the vectors $a_k(\theta, \gamma)$ are constructed to enforce the crossing constraints (3) or (4). If we define $v(\theta) = \max\{j : t_j \leq \theta\}$ if $\theta \geq t_1$ and $v(\theta) = 0$ if $\theta < t_1$, then $a_k(\theta, -1)$ for $k \leq m$ is given by

$$
a_k(\theta, -1) = \begin{cases} 
(1_k^T, 0_{m-k}^T)^T, & \text{if } k \leq v(\theta) \\
(-1_k^T, 0_{m-k}^T)^T, & \text{if } v(\theta) < k \leq m,
\end{cases}
$$

where $1_k$ denotes an $l \times 1$ vector containing all ones, $0_l$ denotes an $l \times 1$ vector containing all zeros, and $0_0$ denotes a vector of “length 0” that should be ignored. Similarly, $a_k(\theta, -1)$ is given by

$$
a_k(\theta, 1) = \begin{cases} 
(1_k^T, 0_{m-k}^T)^T, & \text{if } k \leq v(\theta) \\
(-1_k^T, 0_{m-k}^T)^T, & \text{if } v(\theta) < k \leq m.
\end{cases}
$$

In order to enforce the constraints $u_{j,a}(\theta, \gamma) \leq 0$, $a_k(\theta, \gamma)$ is, for any value of $(\theta, \gamma)$, defined as $a_k(\theta, \gamma) = (0_{k-m-1}^T, -1, 0_{3m-k}^T)^T$, for $k = m + 1, \ldots, 3m$.

The maximum likelihood estimates $\hat{\mathbf{u}}_a(\theta, \gamma)$ of $\mathbf{u}_a(\theta, \gamma)$ can be expressed as the solution to the following optimization problem with linear inequality constraints

$$
\text{maximize } \log L[\mathbf{u}_0(\theta, \gamma), \mathbf{u}_1(\theta, \gamma)] \quad \text{subject to } A_{\theta, \gamma} \begin{bmatrix} \mathbf{u}_0(\theta, \gamma) \\ \mathbf{u}_1(\theta, \gamma) \end{bmatrix} \geq \mathbf{0}_{3m},
$$

where $A_{\theta, \gamma}$ is the $3m \times 2m$ matrix whose $k$th row is $a_k(\theta, \gamma)^T$. The above optimization problem involves maximizing a concave function subject to linear inequality constraints, and, hence, any local maximum is also guaranteed to be a global maximum (Boyd & Vandenberghe, 2004). In our implementation, we use sequential quadratic programming (Nocedal & Wright, 2006) to compute the solution $(\hat{\mathbf{u}}_0(\theta, \gamma)^T, \hat{\mathbf{u}}_1(\theta, \gamma)^T)^T$ of (6). Initialization of $\mathbf{u}_0(\theta, \gamma)$ and $\mathbf{u}_1(\theta, \gamma)$ is done by minimizing, subject to the single-crossing constraint, the squared discrepancy $\sum_{a=0}^{1} \sum_{j=1}^{m} [u_{j,a}(\theta, \gamma) - \log[S_a^K(t_j)]/\log[S_a^K(t_j - 1)]]^2$, where $S_a^K(t_j)$ is the Kaplan–Meier estimate of the survival function in treatment arm $a$. One possible limitation of this computational strategy is that sequential quadratic programming can become very computationally demanding for large values of $m$. One remedy for this in large-$m$ situations would be to group the follow-up times $Y_1, \ldots, Y_m$ into a collection of small “bins” and set $Y_i$ to the midpoint of the bin to which it is assigned.
Algorithm 1 (Single-crossing constrained estimates of survival curves).

**Input:** Follow-up times $Y_1, \ldots, Y_n$. Event indicators $\delta_1, \ldots, \delta_n$. Treatment-arm assignments $A_1, \ldots, A_n$.

1. Compute unique ordered $t_1 < t_2 < \cdots < t_m$ follow-up times from both treatment arms.
2. Set $\gamma_1 = -1$ and $\gamma_2 = 1$.
3. Compute a grid $\delta_1, \ldots, \delta_m$ of potential crossing times: $\delta_1 = 0, \delta_2 = t_1, \ldots, \delta_m = t_{m-1}$.
4. **for** $j = 1, 2, \ldots, m$ **do**
   5. **for** $k = 1, 2$ **do**
      6. Compute estimates $\hat{u}_0(\delta_j, \gamma_k)$, $\hat{u}_1(\delta_j, \gamma_k)$ by solving the optimization problem (6) with the single-crossing constraints implied by the crossing parameters $(\delta_j, \gamma_k)$.
      7. Using $\hat{u}_0(\delta_j, \gamma_k)$, $\hat{u}_1(\delta_j, \gamma_k)$, compute the profile log-likelihood $\ell_{jk} = \ell^P(\delta_j, \gamma_k)$ at $(\delta_j, \gamma_k)$.
      8. Compute $\ell^* = \max_{j,k} \{\ell_{jk}\}$.
      9. For the pair $(\delta_{j^*}, \gamma_{k^*})$ such that $\ell^P(\delta_{j^*}, \gamma_{k^*}) = \ell^*$, set $\hat{\delta}_{sc} = \delta_{j^*}$ and $\hat{\gamma}_{sc} = \gamma_{k^*}$.
   10. Compute estimates $\hat{S}_{sc}^0(t)$, $\hat{S}_{sc}^1(t)$ of the survival curves using
       $$
       \hat{S}_{sc}^a(t) = \exp \left\{ \sum_{j=1}^m \hat{u}_{ja}(\hat{\delta}_{sc}, \hat{\gamma}_{sc}) I(t_j \leq t) \right\},
       $$
       where $\hat{u}_{ja}(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ is the $j$th component of $\hat{u}_a(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$.

2.3 Estimates of crossing-time parameters $\theta$ and $\gamma$

The estimated vectors $\hat{u}_0(\theta, \gamma)$ and $\hat{u}_1(\theta, \gamma)$ will generate estimates of the two survival curves for fixed values of $(\theta, \gamma)$. To find the best values of the crossing parameters $(\theta, \gamma)$, we maximize the profile log-likelihood function $\ell^P(\theta, \gamma)$ associated with $\hat{u}_0(\theta, \gamma)$ and $\hat{u}_1(\theta, \gamma)$

$$
\ell^P(\theta, \gamma) = \sum_{a=0}^1 \sum_{j=1}^m \left[ d_{ja} \log(1 - \exp(\hat{u}_{ja}(\theta, \gamma))) + (R_{ja} - d_{ja}) \hat{u}_{ja}(\theta, \gamma) \right].
$$

(7)

We refer to the values $\hat{\delta}_{sc}, \hat{\gamma}_{sc}$ which maximize $\ell^P(\theta, \gamma)$ as the single-crossing constrained estimates of the crossing parameters $\theta$ and $\gamma$. Because the conditional estimates $\hat{u}_a(\theta, \gamma)$ do not change as $\theta$ varies over each of the intervals $(t_{j-1}, t_j)$ and can only change at each $t_j$, the single-crossing constrained estimates $\hat{\delta}_{sc}$ and $\hat{\gamma}_{sc}$ can be found by solving the following discrete optimization problem:

$$
(\hat{\delta}_{sc}, \hat{\gamma}_{sc}) = \arg \max_{\theta \in \{0, t_1, \ldots, t_{m-1}\}, \gamma \in \{-1, 1\}} \ell^P(\theta, \gamma).
$$

(8)

To find $(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ in (8), we only need to first compute $\ell^P(\theta, \gamma)$ for each of the $2m$ possible choices of $\theta \in \{0, t_1, \ldots, t_{m-1}\}, \gamma \in \{-1, 1\}$ and then to pick the values $(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ such that $\ell^P(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ is the largest among the $2m$ possible choices of $(\theta, \gamma)$. The reason for only considering values of $\theta$ up to $t_{m-1}$ is because both $\theta = 0$ and $\theta = t_m$ refer to situations where one survival function dominates the other survival function at every time point $t_1, \ldots, t_m$. Thus, including $\theta = t_m$ as a possible crossing time is superfluous as $(\theta, \gamma) = (0,1)$ and $(\theta, \gamma) = (0,-1)$ cover both scenarios where one survival curve dominates the other at each of the event times $t_j$. The steps involved in computing the single-crossing constrained estimates $(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ are summarized in Algorithm 1.

The vectors $\hat{u}_a(\hat{\delta}_{sc}, \hat{\gamma}_{sc})$ generate the following estimates of the survival functions:

$$
\hat{S}_{sc}^a(t) = \exp \left\{ \sum_{j=1}^m \hat{u}_{ja}(\hat{\delta}_{sc}, \hat{\gamma}_{sc}) I(t_j \leq t) \right\}.
$$
We refer to $\hat{S}_{0}^{sc}(t)$ and $\hat{S}_{1}^{sc}(t)$ as the single-crossing constrained estimates of the survival functions. Note that both $\hat{S}_{0}^{sc}(t)$ and $\hat{S}_{1}^{sc}(t)$ are flat for $t \geq t_{m}$, and hence if $\hat{S}_{0}^{sc}(t), \hat{S}_{1}^{sc}(t)$ satisfy the single-crossing constraint over $[0, t_{m}]$, they will satisfy it for all time points.

### 2.4 | Alternative single-crossing constraints

The estimation strategy outlined in Sections 2.1–2.3 focuses on single-crossing constraints for the survival functions, but other related single-crossing constraints could potentially be incorporated using a similar approach. We briefly mention a few interesting possible extensions below. While we explore nonsmooth estimation of hazard functions with single-crossing constraints in our application in Section 5, we do not explore the other mentioned extensions further as they lie beyond the scope of this paper.

#### 2.4.1 | Nonsmooth estimation of hazard functions under single-crossing constraints

In many cases, it is also sensible to place single-crossing constraints on the hazards rather than on the survival curves. In this context, without imposing any smoothness conditions we would want the discrete hazards $h_{ja}(\theta, \gamma) = 1 - S_{a}(t_{j})/S_{a}(t_{j-})$ in one treatment arm to be larger (smaller) before some crossing time $\theta$ and remain smaller (larger) for $j$ such that $t_{j} > \theta$. Because of the connection $h_{ja}(\theta, \gamma) = 1 - \exp[u_{ja}(\theta, \gamma)]$ between the $u_{ja}(\theta, \gamma)$ and the discrete hazards $h_{ja}(\theta, \gamma)$, we can express the single-crossing constraints on the discrete hazards as

$$
u_{ja}(\theta, \gamma) \geq \nu_{ja}(\theta, \gamma) \quad \text{for all } j \text{ such that } t_{j} \leq \theta$$

$$
u_{ja}(\theta, \gamma) \leq \nu_{ja}(\theta, \gamma) \quad \text{for all } j \text{ such that } t_{j} > \theta,$$

if $\gamma = 1$ with both inequalities reversed whenever $\gamma = -1$. As in the case of estimating survival functions under single-crossing constraints, one would first, for fixed values of $(\theta, \gamma)$, find conditional maximum likelihood estimates of $u_{ja}(\theta, \gamma)$ by maximizing the log-likelihood (5) subject to constraints (9). After this, one would find estimates of the crossing parameters $(\theta, \gamma)$ by maximizing the associated profile log-likelihood function (7).

#### 2.4.2 | Smooth hazard functions with single-crossing constraints

To find smoothly estimated hazard functions, one could consider hazard functions $h_{a}(\cdot|\theta, \gamma)$ of the form

$$h_{a}(t|\theta, \gamma) = 1 - \exp\left\{ \sum_{j=1}^{m_{a}} s_{ja}(t)u_{ja}(\theta, \gamma) \right\} = 1 - \exp\left\{ s_{a}(t)^{T}u_{a}(\theta, \gamma) \right\},$$

for a choice of smoothing weights $s_{a}(t) = (s_{1a}(t), \ldots, s_{ma}(t))^{T}$. A common choice of smoothing weights, for example, would be $s_{ja}(t) = 1/2K[(t - t_{j})/b]$ for some symmetric kernel function $K(\cdot)$ and bandwidth $b > 0$. Under formulation (10), inequalities for the hazard functions at time points $t_{j}$ can be expressed as linear inequality constraints of the form $s_{1}(t_{j})^{T}u_{1}(\theta, \gamma) \geq s_{a}(t_{j})^{T}u_{a}(\theta, \gamma)$ and, hence, to compute smooth estimates of the hazard functions one could estimate the $\hat{u}_{ja}(\theta, \gamma)$ by maximizing the log-likelihood function (5) subject to the linear inequality constraints implied by the form of the hazard functions in (10). One would then find estimates of the crossing-time parameters $(\theta, \gamma)$ by maximizing the associated profile log-likelihood function (7).

An alternative to this approach would be to simply smooth estimated discrete hazards that have been found using the approach outlined in Section 2.4.1. While this may work well in many situations, this approach would not guarantee that the smoothed hazard function estimates will satisfy the single-crossing constraint.
3 | MEASURES OF TREATMENT EFFICACY AND MODEL INFERENCE

3.1 | Estimands of interest

Milestone survival probabilities. Comparing differences in estimated survival probabilities at one or several prespecified time points $\hat{S}^{sc}_1(t^*_1) - \hat{S}^{sc}_0(t^*_1), \ldots, \hat{S}^{sc}_1(t^*_q) - \hat{S}^{sc}_0(t^*_q)$ can be a useful way of characterizing the treatment effect over time without relying on any assumptions about proportional hazards. Because $\hat{\theta}_{sc}$ provides an estimate of precisely where the sign change in $S_1(t) - S_0(t)$ occurs, augmenting the survival probability differences at the milestones $t^*_1, \ldots, t^*_q$ with the estimated crossing time $\hat{\theta}_{sc}$ can provide additional context when interpreting the estimated differences $\hat{S}^{sc}_1(t^*_j) - \hat{S}^{sc}_0(t^*_j)$.

The proportion surviving up to crossing. The proportion surviving up to the crossing time in treatment arm $a$ is represented by the parameter $S_a(\theta)$. If both $S_1$ and $S_0$ are continuous, we will have $S_1(\theta) = S_0(\theta)$ and if either $S_1$ or $S_0$ is not continuous, these will be approximately equal as long as the true survival curves do not have large jumps. For this reason, we use $S_a(\theta)$ to denote the proportion surviving up to crossing, and, in practice, we estimate this parameter with $\{n_1 \hat{S}^{sc}_1(\hat{\theta}_{sc}) + n_0 \hat{S}^{sc}_0(\hat{\theta}_{sc})\}/n$, where $n_1 = \sum_{i=1}^n A_i$ and $n_0 = n - n_1$ denote the number of patients who have been randomized to the active and the control treatment arm, respectively. The quantity $S_a(\theta)$ could be of particular interest if one is concerned about a substantial fraction of patients experiencing early events that occur before the crossing time. In these cases, reporting an estimate of $S_a(\theta)$ provides a measure of the fraction of patients who will survive long enough to reach the point at which the survival curve in the active treatment arm begins to dominate the control-arm survival curve.

Restricted mean survival time. The RMST (Royston & Parmar, 2013) is defined as the expected time under follow-up for an individual assuming you only follow individuals up to some prespecified time point $\tau$. Specifically, the RMST for individuals in treatment arm $a$ is defined as

$$\text{RMST}_a(\tau) = E\{\min(T_i, \tau) | A_i = a\}.$$ 

RMST$_a(\tau)$ is equal to the area under the survival curve $S_a(t)$ between the time points $t = 0$ and $t = \tau$, and the difference RMST$_1(\tau) - \text{RMST}_0(\tau)$ provides an interpretable measure of treatment effect regardless of whether or not the proportional hazards assumption holds. While providing an interpretable measure of treatment effect, the difference in RMST can mask important differences in survival that occur at earlier time points. One way of addressing this is to also examine differences in RMST for different choices of $\tau$ with differences in RMST$_a(\theta)$ perhaps being of key interest.

Restricted residual mean life. If one is interested in differences in survival for those who are longer survivors, the restricted residual mean life (RRML) function (Cortese et al., 2017) is an appealing measure. The RRML function for treatment arm $a$ is defined at time $t$ as

$$\text{RRML}_a(t, \tau) = E\{\min(T_i, \tau) - t | A_i = a, T_i \geq t\} = \int_t^\tau \frac{S_a(u)}{S_a(t)} du.$$ 

The quantity RRML$_a(\theta, \tau)$ represents, for those in treatment arm $a$, the remaining expected on-study survival up to time $\tau$ conditional on the fact that one has survived up to the crossing time $\theta$. RRML$_a(t, \tau)$ bears some relation to the window mean survival time (Paukner & Chappell, 2021) (WMST) which measures the expected survival time over a certain specified time window. Whereas WMST would measure the expected length of survival within the time window $(t, \tau)$, RRML$_a(t, \tau)$ is a conditional expectation which measures the expected survival time over $(t, \tau)$ conditional on the fact that one has survived up to time $t$.

Crossing-time conditional survival curves. In cases of delayed treatment where the two survival curves cross, it may be of interest to also plot survival probabilities conditional on surviving up to the point of crossing. Such conditional probabilities give the probability of surviving past a point of interest conditional on the fact that one has survived up to the crossing time. This conditional survival curve for patients in treatment arm $a$ is defined, for $t > \theta$, as

$$S_{a,\text{cond}}(t) = P(T_1 > t | A_i = a, T_i > \theta) = \frac{S_a(t)}{S_a(\theta)}.$$ 

The conditional survival curves may be estimated directly using $\hat{S}_{a,\text{cond}}^{sc}(t) = \hat{S}_a^{sc}(t)/\hat{S}_a^{sc}(\hat{\theta}_{sc})$. It is worth mentioning that

$$\text{RRML}_a(\theta, \tau) = \int_{\theta}^{\tau} S_{a,\text{cond}}(u) du.$$
Pre- and post-crossing average hazard ratios. Comparing the average hazard ratios over the time periods before and after the crossing can provide an interpretable measure of treatment efficacy for longer survivors and can provide a good comparison for the relative improvement in treatment efficacy between earlier and later time points. Assuming arm-specific hazard functions $h_a(t)$ exist, we define, as in Kalbfleisch and Prentice (1981), the average hazard ratio using the “active treatment-to-total” hazard ratio which measures the average ratio between the active treatment-arm hazard $h_1(t)$ and the total hazard $h_0(t) + h_1(t)$ across time. Specifically, for a truncation time $\tau$ and $\theta \in (0, \tau)$, we define the pre- and postcrossing average hazard ratios as

$$\lambda_{\text{pre}} = \frac{1}{\theta} \int_0^\theta \frac{h_1(t)}{h_0(t) + h_1(t)} dt \quad \text{and} \quad \lambda_{\text{post}} = \frac{1}{\tau - \theta} \int_\theta^\tau \frac{h_1(t)}{h_0(t) + h_1(t)} dt,$$

respectively. One reason for using the ratio $h_1(t)/[h_0(t) + h_1(t)]$ rather than $h_1(t)/h_0(t)$ to improve estimation stability as very small estimated values of $h_0(t)$ could lead to highly variable estimates of $\lambda_{\text{pre}}$ and $\lambda_{\text{post}}$. When assuming $\tau = \tau_m$, the parameters $\lambda_{\text{pre}}$ and $\lambda_{\text{post}}$ can be estimated by the following quantities:

$$\hat{\lambda}_{\text{pre}} = \frac{1}{\hat{\theta}_{\text{sc}}} \sum_{j=1}^m \frac{\hat{h}_{j1} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}}) (t_j - t_{j-1}) I(t_j \leq \hat{\theta}_{\text{sc}})}{\hat{h}_{j0} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}}) + \hat{h}_{j1} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}})}$$

and

$$\hat{\lambda}_{\text{post}} = \frac{1}{\tau_m - \hat{\theta}_{\text{sc}}} \sum_{j=1}^m \frac{\hat{h}_{j1} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}}) (t_j - t_{j-1}) I(t_j > \hat{\theta}_{\text{sc}})}{\hat{h}_{j0} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}}) + \hat{h}_{j1} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}})},$$

where $t_0 = 0$ and $h_{j0} (\theta, \gamma) = 1 - \exp\{u_{j0}(\theta, \gamma)\}$ is as defined in Section 2.2. Depending on the context, one could either use discrete hazard estimates $\hat{h}_{j0} (\hat{\theta}_{\text{sc}}, \hat{\gamma}_{\text{sc}})$ under the single-crossing constraint on the survival functions or the single-crossing constraint on the hazard functions as described in Section 2.4.1.

### 3.2 Hypothesis testing

While we can compute confidence intervals for certain parameters of interest that do not involve the crossing parameters $(\theta, \gamma)$, it can be useful to perform inference with respect to both $\theta$ and another parameter of interest $\phi$ (or collection of parameters) that represents a measure of treatment efficacy. For example, in traditional settings where it is assumed that proportional hazards hold, $\phi$ would frequently be a hazard ratio, but in settings with delayed treatment effect choosing $\phi$ to be an alternative estimand such as difference in RMST may be more appealing. Combining $\theta$ and $\phi$ in a joint hypothesis test can address concerns about having a scenario where the estimated value of $\phi$ indicates overall treatment effectiveness, but substantial time elapses before the two survival curves clearly separate. Cases such as these may lead to concerns that most of the observed treatment benefit is mainly due to differences in long-term survivors.

In the aforementioned context, one possible hypothesis of interest is that both the efficacy parameter $\phi$ is sufficiently large and the crossing time $\theta$ does not occur too late. This can be expressed more formally as

$$H_0 : \phi \leq \phi^* \text{ or } \theta \geq \theta^* \quad \text{versus} \quad H_A : \phi > \phi^* \text{ and } \theta < \theta^*, \quad (13)$$

where $\phi^*$ and $\theta^*$ are prespecified values of $(\theta, \phi)$, which are determined to be clinically meaningful. Alternatively, if it is difficult to specify a time point before which the crossing should occur, one could instead require that survival in the active treatment arm should be sufficiently large whenever the survival curves cross. The hypothesis test of interest in this case would be

$$H_0 : \phi \leq \phi^* \text{ or } S_1(\theta) \leq p^* \quad \text{versus} \quad H_A : \phi > \phi^* \text{ and } S_1(\theta) > p^*. \quad (14)$$

One could test either (13) or (14) using a permutation test with the test statistics $\hat{\phi}$ and $\hat{\theta}_{\text{sc}}$ or $\hat{\phi}$ and $\hat{S}_{\text{SC}}(\hat{\theta}_{\text{sc}})$, respectively. Another approach would be to test (13) or (14) by using a bootstrap procedure to construct one-sided confidence intervals for either the parameter $\eta_1 = \min(\phi - \phi^*, \theta^* - \theta)$ or the parameter $\eta_2 = \min(\phi - \phi^*, S_1(\theta) - p^*)$. If the lower bound of the confidence interval for $\eta_1$ is greater than zero, one would reject $H_0$ in (13). Likewise, a lower bound for the confidence inter-
val for $\eta_2$ greater than zero would imply that one should reject $H_0$ in (14). Simulation studies evaluating the type-I error and power for tests of hypotheses of the form (14) are described and shown in Appendix D of the Supporting Information.

One should note that the purpose of reporting a confidence interval for $\eta_1$ or $\eta_2$ is to test $H_0$ in either (13) or (14), and one should only check whether or not this confidence interval covers zero or not. The range of values covered by the confidence intervals for either $\eta_1$ or $\eta_2$ does not have a clear interpretation as both $\eta_1$ and $\eta_2$ are minimums of two parameters which could be on quite different scales. If one is more interested in assessing the magnitudes of each of the individual parameters (i.e., $\phi - \phi^*$ and $\theta^* - \theta$ for (13) or $\phi - \phi^*$ and $S_1(\theta) - p^*$ for (14)), one could separately report estimates and confidence intervals for these two parameters rather than only reporting a confidence interval for $\eta_1$ or $\eta_2$.

In the context of delayed treatment effects, another main question of interest is whether or not the active treatment arm provides superior long-term survival. In our framework, this can formalized as the following hypothesis test $H_0 : \gamma = -1$ versus $H_A : \gamma = 1$. One approach for performing this hypothesis test is to use a bootstrap likelihood ratio test of $H_0 : \gamma = -1$ versus $H_1 : \gamma = 1$ in a similar way to the approach described in McLachlan (1987) in the context of testing the number of components in a Gaussian mixture model. Specifically, one could simulate the null distribution (i.e., the distribution under $H_0 : \gamma = -1$) of the log-likelihood ratio test statistic by first estimating the two survival curves under the restriction that $\gamma = -1$ and then repeatedly sampling from these two survival function estimates. Details of this bootstrap sampling procedure are provided in Appendix C of the Supporting Information.

4 | SIMULATIONS

4.1 | Estimation performance with piecewise exponential distributions

We considered six simulation scenarios where, in each scenario, it is assumed that survival follows a piecewise exponential distribution in both treatment arms. The arm-specific survival curves for these six scenarios are depicted in Figure 2.
The top-left graph in Figure 2a depicts Scenario 1 where the survival curves never cross, and the survival curve for the active treatment arm always dominates the control-arm survival curve. Figure 2b depicts Scenario 2 where there is a clear, unambiguous single crossing of the two survival curves at time point 5.0. Figure 2c depicts Scenario 3 where the two survival curves have a single crossing near time point 2.0. While Scenario 3 has a single, distinct crossing, when compared with Scenario 2 the two survival curves in Scenario 3 do not have as much separation before the survival curves cross. Figure 2d shows the survival curves in Scenario 4 where there is a single crossing at time point 0.75, but, in this scenario, there is almost no separation between the curves before the crossing time. In Scenario 5, there is a single crossing at time point 1.5 with little separation before the crossing and a diminishing treatment benefit that occurs towards the end of the time interval considered. In Scenario 6, there are two crossing times, but the later crossing is more “distinct” than the first in the sense that the separation between the two curves is larger immediately before and after the crossing point. Simulation results for an additional, seventh simulation scenario can be found in the Supporting Information.

For these simulations, we set the total number of patients to \( n = 200 \), \( n = 400 \), and \( n = 800 \) with the number of patients split evenly between the two treatment arms for each choice of \( n \). For each of the six simulation scenarios and choice of \( n \), we ran 200 simulation replications. The censoring distribution used in each of the six scenarios was a uniform distribution from 4 to 8. This feature of the simulation design can be thought of as representing a situation where one has uniform enrollment over a period of 4 months and where all patients have a minimum follow-up period of 4 months without any dropouts. While the percentage of survival outcomes which were observed event times varied across simulation scenarios, the percentage of observed events was between 55% and 75% for each of the six settings, and the percentage of observed events was typically 0–15% larger in the control arm than in the active treatment arm. For each simulation setting, we evaluated the performance of the single-crossing constrained (SCC) procedure in estimating the following measures: the difference in RMST at time point 7, the differences in the survival function at the time points 2 and 4, the crossing time \( \theta \), the proportion surviving up to crossing \( S_a(\theta) \), and the difference in RRML using the time points \( \theta \) and 7. While the Kaplan–Meier estimates do not directly generate estimates of \( \theta \) and other parameters which depend on \( \theta \) since the Kaplan–Meier estimates can cross multiple times, we compared the SCC procedure with a Kaplan–Meier-based estimator of \( \theta \) that is defined to be the time of the first crossing of the two Kaplan–Meier curves.

Overall, the simulation results for root-mean-squared error (RMSE) show that, for the parameters not involving the crossing time, one can typically expect modest improvements in RMSE over the Kaplan–Meier estimates if the single-crossing assumption is not clearly violated (Scenarios 1–3), slightly worse or similar performance in Scenario 4 when the two survival curves have a long period of near overlap and notably worse performance in Scenarios 5 and 6 where the single-crossing assumption very clearly does not hold. For the crossing time-related parameters \( \hat{\theta}, S_a(\hat{\theta}) \), and \( \Delta \text{RRML}(\theta, 7) \), the relative performance of the SCC method was best when both the single-crossing constraint was satisfied and when the crossing occurred after time zero (i.e., Scenarios 2 and 3). Table 1 shows the RMSE for the SCC estimates in Scenarios 1–3 from our six piecewise-exponential simulation scenarios and the three choices of sample size. These are the three scenarios that have either no crossing or a distinct, single crossing. As shown in Table 1, in the three scenarios which plainly

| \( n \) | Scenario | \( \Delta \text{RMST}(7) \) | \( S_1(2) - S_0(2) \) | \( S_1(4) - S_0(4) \) | \( \theta \) | \( S_a(\theta) \) | \( \Delta \text{RRML}(\theta, 7) \) |
|-------|---------|-----------------|-----------------|-----------------|-------|-----------------|-----------------|
|       | SCC     | KM              | SCC             | KM              | SCC   | KM              | SCC             | KM              | SCC |
| 200   | 1       | 0.367           | 0.373           | 0.066           | 0.061 | 0.061           | 0.806           | 0.472           | 0.149 |
|       |         |                 |                 |                 |       |                 | 0.137           |                 | 0.393 |
|       |         |                 |                 |                 |       |                 | 0.404           |                 |       |
| 400   | 1       | 0.268           | 0.269           | 0.048           | 0.048 | 0.044           | 0.605           | 0.184           | 0.096 |
|       |         |                 |                 |                 |       |                 | 0.086           |                 | 0.282 |
|       |         |                 |                 |                 |       |                 | 0.281           |                 |       |
| 800   | 1       | 0.181           | 0.193           | 0.031           | 0.033 | 0.030           | 0.055           | 0.098           | 0.028 |
|       |         |                 |                 |                 |       |                 | 0.049           |                 | 0.184 |
|       |         |                 |                 |                 |       |                 | 0.196           |                 |       |
| 2     | 0.178   | 0.179           | 0.030           | 0.029           | 0.032 | 0.032           | 0.614           | 2.958           | 0.034 |
|       |         |                 |                 |                 |       |                 | 0.346           |                 | 0.158 |
|       |         |                 |                 |                 |       |                 | 0.539           |                 |       |
| 3     | 0.182   | 0.182           | 0.032           | 0.032           | 0.034 | 0.034           | 0.541           | 1.162           | 0.066 |
|       |         |                 |                 |                 |       |                 | 0.192           |                 | 0.188 |
|       |         |                 |                 |                 |       |                 | 0.264           |                 |       |
satisfy the single-crossing constraint, the SCC-based estimates had RMSE performance, which was generally as good or better than the Kaplan–Meier-based estimates for parameters for which such a comparison could be made. RMSE for the crossing-time estimator \( \hat{\theta}_{sc} \) was lowest in both Scenarios 1 and 3 where there was either no crossing or an early, distinct crossing. The poorer results for \( \hat{\theta}_{sc} \) in Scenario 2 compared to other scenarios are likely due to the fact that, in this scenario, the true crossing time \( \theta \) occurred much later in the study at a time when there would typically be many fewer individuals remaining in this study. Despite this, the estimation performance of the SCC estimator of \( \theta \) was considerably better than our Kaplan–Meier-based estimator of this quantity. Moreover, the estimation performance of the SCC-based estimator of \( S_\theta(\theta) \) was quite good in Scenario 2 as both survival curves are much more flat towards the end of the study period.

Table 2 shows the RMSE results for the SCC estimates in Scenarios 4–6. These are the three scenarios where the single-crossing constraint is either not satisfied or is nearly violated due to either a near double crossing or a very “weak” crossing of the survival curves. Of these three scenarios, Scenario 5 was the one setting where the estimation of the crossing-time parameter was notably poor relative to the first crossing time estimator using the Kaplan–Meier curves. This was mainly due to the strong diminished treatment effect present in Scenario 5, which often resulted in a crossing-time estimate closer to the end of the considered time window rather than the much earlier true crossing time of 1.5. Indeed, the small proportion of cases in Scenario 5 with an estimated very late crossing time drove much of the worse RMSE performance relative to the Kaplan–Meier estimator of the first crossing time. This can be seen by comparing the estimators of \( \theta \) using median absolute deviation instead of RMSE. The median absolute deviation comparison shows that the typical deviation of the SCC estimator from the true value \( \theta \) is much closer to the typical deviation of the Kaplan–Meier estimator from the true \( \theta \). The Supporting Information provides median absolute deviation comparisons for Scenario 5 and several other scenarios. Estimation performance of the SCC-based estimators was overall quite poor in Scenario 6, but this was a scenario where the assumption of a single crossing was plainly violated. In Scenario 6, the RMSE performance of the SCC estimator is consistently around 25–30% worse than the Kaplan–Meier estimator for each of the measures and sample sizes considered. The Supporting Information contains results for both the bias and variance across each of these six simulation scenarios.

While Tables 1 and 2 display mean-squared error (MSE) performance for crossing time-related parameters and features of the differences in the arm-specific survival functions, Figure 3 shows MSE performance of the SCC estimator for features of the arm-specific survival functions \( S_0 \) and \( S_1 \) when the sample size is set to \( n = 400 \). Figure 3 shows the modest but consistently better performance of the SCC estimators versus the Kaplan–Meier estimators for the non-crossing-time parameters in Scenarios 1–3. For Scenarios 4–6, the SCC estimator demonstrates considerable robustness with comparable MSE performance to the Kaplan–Meier estimator across most measures and scenarios. The poor performance of the SCC-based estimates of features of the differences in the survival functions shown in Tables 1 and 2 (for Scenarios 4–6) is likely due to associations between \( S_1^{\text{sc}}(t) \) and \( S_0^{\text{sc}}(t) \) that are induced by the single-crossing constraint. More detailed MSE comparisons for features of the arm-specific survival distributions including bias and variance comparisons are provided in the Supporting Information.
### 5 DATA EXAMPLE

In this section, we examine reconstructed survival outcomes from a recently completed phase III trial (Hellmann et al., 2019) examining the efficacy of a combination of immune checkpoint inhibitors, nivolumab plus ipilimumab, for the treatment of non-small-cell lung cancer. This is a context where there is a strong prior rationale for employing a single-crossing constrained estimation procedure as a number of previous trials have shown a delayed treatment effect in comparisons of immune checkpoint inhibitors and chemotherapy. In this trial, patients were assigned to one of three treatment arms: a combination arm where nivolumab plus ipilimumab was administered, a monotherapy arm where nivolumab alone was administered, and a control arm where only chemotherapy was given. The primary endpoint in this study was overall survival (OS) in the combination therapy arm versus the chemotherapy arm in the subpopulation of patients whose tumors had an expression level of the programmed death ligand 1 (PD-L1) that was at least 1%. Among the group of patients who had a PD-L1 expression of 1% or more, 396 patients were assigned to the combination arm, and 397 patients were assigned to the chemotherapy only arm. While there was a notable delay in treatment effect in this study, the analysis of this study reported in Hellmann et al. (2019) concluded that the nivolumab plus ipilimumab treatment resulted in improved overall survival when compared with chemotherapy. In our analysis, we utilized survival outcomes that we reconstructed from the published Kaplan–Meier curves for OS in Hellmann et al. (2019). (See Appendix A of the Supporting Information for further details on our reconstruction algorithm, and see, e.g., Liu et al., 2021, or Guyot et al., 2012, for closely related approaches for reconstructing survival outcomes from published Kaplan–Meier curves.) Due to the resolution of
these published images, our reconstructed survival outcomes are unlikely to be exactly the same as those recorded in this study, but the reconstructed survival outcomes reproduce the published Kaplan–Meier curves quite closely. Using the reconstructed outcomes, the median OS in the nivolumab plus ipilimumab arm was 17.3 months while the median OS in the chemotherapy arm was 15.0 months. While the median OS suggests an overall benefit of the combination therapy, the Kaplan–Meier estimates of OS indicate a delay in treatment effect as the estimated OS survival curve for the chemotherapy arm initially dominates the estimated OS survival curve for the combination therapy arm, and a crossing appears to occur sometime between 6 and 9 months before the two Kaplan–Meier estimates clearly separate at later time points.

Figure 4 displays the single-crossing constrained estimates of the combination-arm and chemotherapy-arm survival curves for OS. As shown in this figure, the single-crossing constrained survival curve estimate for the chemotherapy arm shows an earlier superiority over the combination-arm survival curve, while the combination-arm survival curve remains superior after the survival curves cross. The single-crossing constrained estimate $\hat{\theta}_{sc}$ of the crossing time was $\hat{\theta}_{sc} = 7.36$ months, and the corresponding estimate of the superior long-term survival parameter was $\hat{\gamma}_{sc} = 1$. The right-hand panel of Figure 4 shows the single-crossing constrained estimates of the conditional survival curves $S_{0, cond}(t)$ defined in (11). These curves represent estimates of survival probabilities conditional on the fact that one has survived up to the crossing time. The graph of $\hat{S}_{0, cond}(t)$ and $\hat{S}_{1, cond}(t)$ shows a clear superiority of the active treatment arm among those patients who will survive up to approximately seven and a half months. Indeed, the probability for surviving more than 2 years conditional on surviving up to the crossing time is 0.54 in the combination arm and 0.47 in the chemotherapy arm, and the probability for surviving more than 3 years conditional on surviving up to the crossing is 0.44 in the combination arm and 0.29 in the chemotherapy arm.

Table 3 displays single-crossing constrained estimates and their associated 95% confidence intervals for other measures of treatment efficacy. To obtain these confidence intervals, we used a bootstrap with stratified resampling (Davison...
TABLE 3  Single-crossing constrained estimates of different efficacy measures from the reconstructed nivolumab+ipilimumab versus chemotherapy trial.

| Parameter                      | Estimate | 2.5%    | 97.5%   |
|-------------------------------|----------|---------|---------|
| $\theta$                      | 7.36     | 4.15    | 23.81   |
| $S_a(\theta)$                | 0.73     | 0.37    | 0.86    |
| $\text{RMST}_1(36) - \text{RMST}_0(36)$ | 1.48     | -0.21   | 3.46    |
| $S_i(6) - S_o(6)$             | -0.03    | -0.09   | 0.02    |
| $S_i(12) - S_o(12)$           | 0.04     | -0.03   | 0.12    |
| $S_i(24) - S_o(24)$           | 0.06     | -0.01   | 0.13    |
| $S_i(36) - S_o(36)$           | 0.11     | 0.05    | 0.18    |
| $S_{1,\text{cond}}(12) - S_{0,\text{cond}}(12)$ | 0.06     | 0.00    | 0.15    |
| $S_{1,\text{cond}}(24) - S_{0,\text{cond}}(24)$ | 0.08     | 0.00    | 0.16    |
| $S_{1,\text{cond}}(36) - S_{0,\text{cond}}(36)$ | 0.15     | 0.08    | 0.24    |

& Hinkley, 1997) where, in each bootstrap replication, a subsample of the survival outcomes $(Y_i, \delta_i)$ was drawn with replacement from each of the treatment arms. As shown in this table, our estimate of the proportion surviving up to crossing parameter $S_a(\theta)$ was 0.73, suggesting that approximately 73% of individuals in either treatment arm will survive up to the time point where the active treatment arm will begin to have superior survival probabilities. This implies that, while combination therapy can notably improve long-term survival, there is still a substantial fraction of patients that one should expect to die (in either treatment arm) before the time points at which the survival curves begin to more clearly separate. The estimated difference in RMST truncated at 3 years was 1.48 months with an associated 95% confidence interval of $[-0.21, 3.46]$. While this estimated RMST difference is positive, the uncertainty reflected in the confidence interval suggests there is no strong evidence of an overall treatment benefit according to this measure. Thus, although comparisons of survival and conditional survival probabilities at 24 and 36 months, suggest a long-term survival advantage of the active treatment, this information needs to be balanced with both the weak evidence for an overall positive treatment effect (according to difference in RMST) and the mildly worse mortality in the early months before the crossing when making any type of overall assessment of the advantages of combination therapy versus chemotherapy.

In addition to estimating various efficacy parameters, we performed a hypothesis test of $H_0 : \text{RMST}_1(36) - \text{RMST}_0(36) \leq 0$ or $S_1(\theta) \leq 0.6$ using the approach outlined in Section 3.2. Specifically, we computed a point estimate and one-sided 95% confidence interval for the parameter $\eta_2 = \min \{ \text{RMST}_1(36) - \text{RMST}_0(36), S_1(\theta) - 0.6 \}$. Using the single-crossing constrained estimates of the survival curves and $\theta$, the point estimate of $\hat{\eta}_2$ was $\hat{\eta}_2 = 0.129$, and using 200 bootstrap replications, a one-sided 95% confidence interval for $\eta_2$ was $[-0.024, \infty)$. Although the lower bound of this confidence interval is quite close to the threshold of 0, we would, in this case, not reject the null hypothesis that either the difference in 36 month RMST is less than 0 or that the proportion surviving up to crossing is at most 0.6 (or both).

We also computed estimates of crossing-time parameters and the arm-specific hazard functions under a single-crossing constraint on the hazards rather than the survival curves. Here, we used the approach described in Section 2.4.1 where a single-crossing constraint was placed on the discrete hazards with the support of the discrete hazards being placed on the set of observed event times. The left-hand panel of Figure 5 shows the estimated discrete hazards for both treatment arms with the estimated hazard-crossing time of 2.4 months. This crossing-time estimate suggests that, while those in the combination arm initially have a larger hazard than those in the control arm, the elevated hazard for those in the combination arm disappears after roughly two and a half months. Using the crossing time of 2.4 months, estimates of the pre- and postcrossing average hazard ratio parameters described in (12) were 0.77 and 0.32, respectively.

While the hazard-based estimate of the crossing time can be useful, the discrete hazard estimates are very nonsmooth and hard to interpret. The right-hand panel of Figure 5 shows hazard function estimates obtained by smoothing the discrete hazard estimates in the left-hand panel. To smooth the discrete hazards, we used the LOWESS smoother (Cleveland, 1979) with the smoother span set to $2/3$. We did not impose any additional single-crossing constraints when performing this smoothing, and for the time interval of 0–3 years the single-crossing constraint for the smoothed hazard functions was satisfied without the use of additional constraints on the smoothed functions.
6 | CONCLUSION

In this article, we have proposed nonparametric estimators of two survival curves when such curves are constrained to cross at most once. The development of these single-crossing constrained estimators was primarily motivated by clinical trials involving recent cancer immunotherapies where it is common to observe delays in treatment effects. While allowing for more than one crossing could provide additional flexibility, our experience with IO trials suggests that most successful therapies have at most one distinct crossing, and cases where one could argue that multiple crossings are present in the underlying survival curves rarely provide clear evidence of long-term benefit to patients. Though our approach can improve estimation performance in cases where the underlying survival curves conform to a single-crossing constraint, one of the main advantages of our approach is that it directly allows for inference on a number of interpretable and useful measures of treatment efficacy. These include the crossing time itself, the proportion of patients who survive past the crossing time, and crossing-time conditional survival probabilities. Combining estimates of measures such as these with more traditional measures of overall efficacy and with a careful examination of survival differences before a crossing occurs can provide important additional context by which to evaluate the benefits or trade-offs associated with the active treatment. In addition to estimation with single-crossing constraints on the survival functions, we also explored similar nonparametric estimators under single-crossing constraints on the hazard functions. Such constraints may be more plausible in many contexts, and certain efficacy measures such as pre- and postcrossing average hazard ratios may be more interpretable under single-crossing constraints on the hazard functions.

As exemplified in our simulation studies, the single-crossing constrained estimates of the survival curves perform well when the single-crossing assumption is satisfied but often perform poorly in situations where such an assumption is plainly violated. Due to this lack of robustness to certain types of model misspecification, our method is better positioned as a secondary analysis that is useful for estimating the time of a survival curve crossing and for quantifying interpretable measures of treatment benefit under delayed treatment effects. One potential way to make our approach more robust...
against an erroneous model assumption would be to develop an appropriate goodness-of-fit test and refrain from using
the single-crossing constrained estimates when the goodness-of-fit test indicates that a single-crossing assumption is not
warranted. Though we do not explore this idea here, one possibility for developing such a goodness-of-fit test would
be to look at the sum of absolute differences between the leave-one-out estimates of survival from the single-crossing
estimates with the complete-data Kaplan–Meier estimates of survival and to also look at the sum of absolute difference
between the leave-one-out Kaplan–Meier estimates of survival with the complete-data Kaplan–Meier survival estimates.
If the sum of absolute differences from the leave-one-out single crossing estimates versus the complete-data Kaplan–
Meier estimates is significantly larger than the sum of absolute differences from the leave-one-out Kaplan–Meier estimates
versus the complete-data Kaplan–Meier estimates, one would reject the assumption that a single-crossing assumption
was appropriate. Instead of using leave-one-out estimates in the goodness-of-fit test, an alternative would be to use K-fold
cross-validation where the survival functions would be estimated using one of the training datasets rather than on the
dataset where only one observation has been removed. Despite the potential utility of a goodness-of-fit test that could be
developed, any such goodness-of-fit test will necessarily have low power in small-sample settings. In these cases, using
Kaplan–Meier estimates as the primary analysis tool may be a more robust choice unless one has a strong justification for
why the underlying survival curves should satisfy the single-crossing assumption.

Our approach is designed to compare marginal survival distributions in two treatment arms when there is a potential
for a survival curve crossing. However, it is certainly the case that the observed delayed treatment effect could be driven in
part by substantial differences in response to treatment across certain patient subgroups, and the within-subgroup survival
distribution comparisons may differ substantially from that of the marginal treatment-specific survival distributions. For
example, the survival curves for patients in a higher risk subgroup may have a very late crossing while the survival curves
for patients in a lower risk subgroup may have no crossing or have a very early crossing, and the associated marginal survival
curves might have a crossing in between these two extremes. The development and use of improved biomarkers
for predicting patient response and stratifying patients by such biomarkers have the potential to partially explain the
extent of the observed delays in treatment effect. In the context of immunotherapy trials, it would be difficult to account
for any heterogeneity in the extent of delay in treatment effect unless the patient characteristics/biomarkers driving such
heterogeneity have been identified prior to the start of the trial and have been included in the statistical analysis plan. (See,
Cristescu et al., 2018, e.g., for a discussion of biomarkers with the potential to improve prediction of patient response to
IO treatments.) To extend our method to accommodate patient-level biomarkers, one possible approach might be to make
a proportional hazards assumption for the effect of the biomarker within each treatment arm without making a propor-
tional hazards assumption for the effect of the treatment. With this approach, one could still allow the baseline survival
functions to cross even though a proportional hazards assumption for the role of the biomarker is made. Appendix B of
the Supporting Information provides more details for this potential approach.

Though not explored in the present work, the single-crossing constrained estimates of the crossing parameters could
potentially be deployed in the context of an overall test of the equality of the arm-specific survival curves \( S_0(t) \) and \( S_1(t) \).
This could potentially improve power in cases where the active treatment shows a delayed treatment effect and where it is
difficult to prespecific the extent of the delay in treatment effect. One possible testing approach is to use a weighted Kaplan–
Meier test statistic (Pepe & Fleming, 1989) where the test statistic is the integrated difference of the Kaplan–Meier functions
multiplied by a weight function. With this approach, one could choose a weight function that is zero for early time points
and positive at time points after the estimated crossing time. This would allow the test statistic to avoid the negative
contribution of the survival curve differences before the crossing, thereby increasing the power of the test. This weighted
Kaplan–Meier test would closely resemble the test statistic proposed in Logan et al. (2008), who used a prespecified rather
than estimated time point to determine the support of their weight function. Another attractive alternative would be to
consider a weighted log-rank test with a piecewise constant weight function similar to the one proposed in Xu et al. (2017),
where it is argued that using a weight function with a jump at a hazard ratio change point has optimal power. Using a
similar approach, one could, by employing the single-crossing constraint on the hazards, specify the jump of a weight
function in a log-rank test to occur at the estimated crossing time of the hazards. Though establishing the asymptotic
null distribution of either the weighted Kaplan–Meier or weighted log-rank test statistic may be challenging, Monte Carlo
permutation tests could be used to estimate the desired \( p \)-values.

Another context in which the single-crossing constrained estimation procedure could be very useful is in the design
stage of a study. If relevant historical data are available, one could compute estimates of the crossing time of the hazards
or survival functions and other relevant features such as the proportion surviving up to crossing, and these estimates
could be used to better inform parameter choices used in sample size and power calculations. Because studies involving
immunotherapies frequently anticipate nonproportional hazards and a sudden change in the hazard ratio after an initial delayed treatment effect period, estimates of the crossing time and related quantities could be valuable in better assessing the power of a future study.

**CONFLICT OF INTEREST STATEMENT**
The authors declare no potential conflict of interest.

**DATA AVAILABILITY STATEMENT**
The dataset used in this article is publicly available within the R package DelayedSurvFit which may be obtained at https://github.com/nchenderson/DelayedSurvFit.

**OPEN RESEARCH BADGES**
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

**ORCID**
Nicholas C. Henderson https://orcid.org/0000-0003-3184-0945

**REFERENCES**
Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.
Chen, T.-T. (2013). Statistical issues and challenges in immuno-oncology. *Journal for Immunotherapy of Cancer*, 1(1), 18.
Chen, T.-T. (2015). Milestone survival: A potential intermediate endpoint for immune checkpoint inhibitors. *Journal of the National Cancer Institute*, 107(9), djv156.
Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829–836.
Cortese, G., Holmboe, S. A., & Scheike, T. H. (2017). Regression models for the restricted residual mean life for right-censored and left-truncated data. *Statistics in Medicine*, 36(11), 1803–1822.
Cristescu, R., Mogg, R., Ayers, M., Albright, A., Murphy, E., Yearley, J., Sher, X., Liu, X. Q., Lu, H., Nebozhyn, M., Zhang, C., Lunceford, J. K., Joe, A., Cheng, J., Webber, A. L., Ibrahim, N., Pлимack, E. R., Ott, P. A., Seiwert, T. Y., … Kaufman, D. (2018). Pan-tumor genomic biomarkers for PD-1 checkpoint blockade–based immunotherapy. *Science*, 362(6411), eaar3593.
Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics, number 1. Cambridge University Press.
Guyot, P., Ades, A., Ouwens, M. J., & Welton, N. J. (2012). Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 12(1), 1–13.
Harrington, D. P., & Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, 69(3), 553–566.
Hellmann, M. D., Paz-Ares, L., Bernabe Caro, R., Zurawski, B., Kim, S.-W., Carcereny Costa, E., Park, K., Alexandru, A., Lupinacci, L., de la Mora Jimenez, E., Sakai, H., Albert, I., Vergnenegre, A., Peters, S., Syrigos, K., Barlesi, F., Reck, M., Borghaei, H., Brahmer, J. R., … Ramalingam, S. S. (2019). Nivolumab plus ipilimumab in advanced non–small-cell lung cancer. *New England Journal of Medicine*, 381(21), 2020–2031.
Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scandinavian Journal of Statistics*, 5(4), 195–199.
Kalbfleisch, J. D., & Prentice, R. L. (1980). Estimation of the average hazard ratio. *Biometrika*, 68(1), 105–112.
Lawless, J. F. (2011). *Statistical models and methods for lifetime data*. Wiley Series in Probability and Statistics, Vol. 362. John Wiley & Sons.
Lin, R. S., Lin, J., Roychoudhury, S., Anderson, K. M., Hu, T., Huang, B., Leon, L. F., Liao, J. J., Liu, R., Luo, X., Mukhopadhayay, P., Qin, R., Tatsuoka, K., Wang, X., Wang, Y., Zhu, J., Chen, T.-T., Iacona, R., & Cross-Pharma Non-Proportional Hazards Working Group. (2020). Alternative analysis methods for time to event endpoints under nonproportional hazards: A comparative analysis. *Statistics in Biopharmaceutical Research*, 12(2), 187–198.
Liu, N., Zhou, Y., & Lee, J. J. (2021). IPDfromKM: Reconstruct individual patient data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, 21(1), 1–22.
Logan, B. R., Klein, J. P., & Zhang, M.-J. (2008). Comparing treatments in the presence of crossing survival curves: an application to bone marrow transplantation. *Biometrics*, 64(3), 733–740.
McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 36(3), 318–324.
Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.
Pak, K., Uno, H., Kim, D. H., Tian, L., Kane, R. C., Takeuchi, M., Fu, H., Claggett, B., & Wei, L.-J. (2017). Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. *JAMA Oncology*, 3(12), 1692–1696.
Park, Y., Kalbfleisch, J. D., & Taylor, J. M. (2012). Constrained nonparametric maximum likelihood estimation of stochastically ordered survivor functions. *Canadian Journal of Statistics, 40*(1), 22–39.

Paukner, M., & Chappell, R. (2021). Window mean survival time. *Statistics in Medicine, 40*(25), 5521–5533.

Pepe, M. S., & Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics, 45*, 497–507.

Rahman, R., Fell, G., Ventz, S., Arfé, A., Vanderbeek, A. M., Trippa, L., & Alexander, B. M. (2019). Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: Prevalence, associated factors, and implications. *Clinical Cancer Research, 25*(21), 6339–6345.

Royston, P., & Parmar, M. K. (2013). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMCMedical Research Methodology, 13*(1), 1–15.

Schemper, M., Wakounig, S., & Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine, 28*(19), 2473–2489.

Xu, Z., Zhen, B., Park, Y., & Zhu, B. (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. *Statistics in Medicine, 36*(4), 592–605.

Zhao, L., Claggett, B., Tian, L., Uno, H., Pfeffer, M. A., Solomon, S. D., Trippa, L., & Wei, L. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics, 72*(1), 215–221.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Henderson, N. C., Nam, K., & Feng, D. (2024). Nonparametric analysis of delayed treatment effects using single-crossing constraints. *Biometrical Journal, 66*, 2200165.

https://doi.org/10.1002/bimj.202200165