Journal of Data Science 0 (0), 1–14
??? 2025
DOI: 10.6339/25-JDS1199
Statistical Data Science

Pseudo Partial Likelihood Method for Proportional Hazards Models when Time Origin Is Missing for Control Group with Applications to SARS-CoV-2 Seroprevalence Study

Yunro Chung[1,2], Vel Murugan[2], Kassu Mehari Beyene[3], and Ding-Geng Chen[1,4,*]

[1]*College of Health Solutions, Arizona State University, Phoenix, AZ, U.S.A.*
[2]*Virginia G. Piper Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ, U.S.A.*
[3]*Department of Neurology, Barrow Neurological Institute, Phoenix, AZ, U.S.A.*
[4]*Department of Statistics, University of Pretoria, Pretoria, Gauteng, South Africa*

Abstract

Time-to-event data analysis without a well-defined time origin commonly occurs in observational studies that retrospectively collect survival endpoints. For instance, after enrolling participants who have or have not received a specific treatment, an event status can be observed for all participants; however, the start date of treatment is only observable for the treatment group. The corresponding time origin does not exist for the control group, resulting in missing survival time data. Complete-case analysis is often considered the standard approach, but it disregards information from all participants in the control group and does not allow us to compare their survival distributions. To address this challenge, we propose a novel semiparametric proportional hazards model by regarding these missing time origins as nuisance parameters. We approximate the risk sets as cumulative normal distributions to deal with these nuisance parameters and develop estimation and inference procedures for our proposed estimator. We study the asymptotic properties of this model and conduct the simulation studies to validate its finite sample property. Analysis of data from a recent SARS-CoV-2 seroprevaluence study illustrates the applicability of our methods. The proposed methods are implemented in the R package *coxphm*.

Keywords *COVID-19; missing data; observational study; right censoring; semiparametric regression; vaccine efficacy*

1 Introduction

After the outbreak of coronavirus disease in 2019 (COVID-19), many serological prevalence survey (serosurvey) studies have been conducted across various regions and countries to determine the prevalence of antibodies against the SARS-CoV-2 virus in their target populations (Havers et al., 2020; Mercado-Reyes et al., 2022; Venugopal et al., 2021; Nah et al., 2021; Vusirikala et al., 2021; Anand et al., 2020; Moreira-Soto et al., 2021; Lombardi et al., 2021). Many of these were snapshot or cross-sectional studies, recruiting participants within relatively short time frames and collecting two types of data: 1) SARS-CoV-2 infection status and 2) self-reported survey information, such as age, gender, date of COVID-19 vaccinations, and other relevant details. The data from these studies would be valuable resources for assessing the efficacy of COVID-19

*Corresponding author. Email: Ding-Geng.Chen@asu.edu.

Received December 7, 2024; Accepted September 18, 2025

vaccinations in a broader and more representative population, especially when conducting large vaccine trials and obtaining timely results, which is a challenge during emergency situations such as the COVID-19 pandemic.

To achieve this goal, one can utilize these serosurvey datasets to compare SARS-CoV-2 infection rates between vaccinated and unvaccinated groups. However, this approach may be too simplistic because it does not account for the fact that the vaccine efficacy diminishes over time. After receiving COVID-19 vaccines such as Moderna or Pfizer (Baden et al., 2021; Polack et al., 2020), antibodies against SARS-CoV-2 typically peak four weeks post-vaccination and decline thereafter. Higher antibody levels against SARS-CoV-2 are linked to greater protection against infection. Consequently, individuals who recently received these vaccines may have differing antibody levels and infection risks compared to those who were vaccinated a long time ago. To draw more comprehensive and informative conclusions, it is essential to evaluate vaccine efficacy over time, but this is also challenging because there is no comparable starting point. The date of vaccination is only available for vaccinated individuals and does not exist for those unvaccinated. Therefore, comparing their vaccine efficacy over time is impossible due to lacking a corresponding reference date.

The same problem can arise in cross-sectional studies that retrospectively collect survival endpoints. After enrolling subjects who have or have not received a specific treatment, an event status can be observed for all subjects. However, the start date of treatment was observable only for those who had received treatment. The corresponding start date does not exist for those who did not receive treatment, which results in missing survival time data. Complete-case analysis is often regarded as the standard approach. However, it disregards information from participants who did not receive the treatment and does not allow us to compare their survival distributions.

To address this challenge, we propose a novel semiparametric proportional hazards model by regarding these missing time origins as nuisance parameters. Partial likelihood analysis is difficult in our setting because partial likelihood, including the nuisance parameters via risk sets, is a non-differentiable step-function. Alternatively, we developed a pseudo-partial likelihood by approximating the risk sets as cumulative normal distributions. We created an efficient algorithm to maximize the pseudo-partial likelihood. The closest relevant work is that of Chen et al. (Chen et al., 2024) who suggested a full parametric Weibull proportional hazards model for this problem. However, their approach was sensitive to the misspecification of the baseline hazard function. Xiong et al. (2021) suggested using an auxiliary longitudinal variable to estimate the missing time origin, but their approach is not practical when such an auxiliary variable is not available.

The remainder of this paper is organized as follows. Section 2 introduces the problems and proposes a pseudo-partial likelihood estimator. In Section 3, we describe the design and conduct simulation studies under various scenarios. In Section 4, we analyze a SARS-CoV-2 serosurvy study to demonstrate the applicability of our method. Finally, Section 5 concludes the article with some discussions.

2 Method

2.1 Notation

Let O_i, T_i^O, and C_i^O denote the calendar dates of the time origin, event of interest, and censoring, respectively, for the ith subject, $i = 1, 2, \ldots, n$. Define T_i and C_i as survival and censoring times, respectively, where $T_i = T_i^O - O_i$ and $C_i = C_i^O - O_i$ for $i = 1, 2, \ldots, n$. Needless to say, the

time origin O_i has to be well-defined for all subjects. Otherwise, T_i and C_i are not definable. Let $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})$ be a p-dimensional covariate vector for the ith subject, respectively, for $i = 1, 2, \ldots, n$. Following the traditional assumption that T_i and C_i are conditionally independent given Z_i, and suppose that the first m subjects belong to the control group, and the remaining $n - m$ subjects belong to the treatment groups.

Define $\mathbb{C} = \{i \mid Z_{i1} = 0, i = 1, 2, \ldots, n\}$ and $\mathbb{T} = \{j \mid Z_{j1} = 1, j = 1, 2, \ldots, n\}$ as disjoint sets of control and treatment groups, respectively, where $Z_{i1} = 1$ or $Z_{i1} = 0$ indicates that the ith subject belongs to the treatment or control group for $i = 1, \ldots, n$. Under the right-censoring mechanism, (Y_j, δ_j, Z_j) is observed if $j \in \mathbb{T}$, where $Y_j = \min(T_j, C_j)$ and $\delta_j = I(T_j \leqslant C_j)$ are the observed survival time and event indicator. However, if $i \in \mathbb{C}$, then O_i is not definable, as discussed in Section 1. Subsequently, O_i is missing for entire control group, and T_i and C_i are not observable for $i \in \mathbb{C}$. Instead, (Y_i^O, δ_i, Z_i) is observed if $i \in \mathbb{C}$, where $Y_i^O = \min(T_i^O, C_i^O)$ is the date of the observed survival time, and event indicator $\delta_i = I(T_i \leqslant C_i) = I(T_i^O \leqslant C_i^O)$ is observed without O_i. Without loss of generality, we assume the first m and the remaining $n - m$ subjects belong to the control and treatment groups, respectively, so $\mathbb{C} = \{1, 2, \ldots, m\}$ and $\mathbb{T} = \{m + 1, m + 2, \ldots, n\}$.

Remark 1. Although the time origins and survival times are missing for subjects in the control group, their other variables, i.e., (Y_i^O, δ_i, Z_i) for $i \in \mathbb{C}$, are fully observed. Therefore, we do not assume that any truncation occurs.

2.2 Limitations of Complete-Case and Cross-Section Analyses

Consider the proportional hazards model, which assumes

$$\lambda(t \mid Z) = \lambda_0(t) e^{\beta^\top Z}, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard function, and $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is a p dimensional regression coefficients. Note that (1) is commonly used to estimate vaccine efficacy over time (Baden et al., 2021) when, for example, Z_{i1} is an indicator variable for vaccine completion at baseline. Suppose for a moment that O_i is observed for $i \in \mathbb{C}$, so Y_i is completely observed for all $i = 1, 2, \ldots, n$. Under model (1), the Cox partial likelihood (Cox, 1972) is defined as

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta^\top Z_i}}{\sum_{j=1}^{n} I(Y_j \geqslant Y_i) e^{\beta^\top Z_j}} \right\}^{\delta_i}.$$

Partial likelihood has been extensively studied in the literature. Let $\beta_0 \in \mathbb{B} \in \mathbb{R}^p$ be the true parameter.

Denote $\bar{\beta}$ as the maximizer of $PL(\beta)$ over $\mathbb{B}$. Under the mild conditions, the maximum partial likelihood estimator $\bar{\beta}$ is a consistent estimator of β_0 and asymptotically normal (Fleming and Harrington, 2013). However, in our setting, $PL(\beta)$ is not defined because Y_i is missing for $i \in \mathbb{C}$, and $\bar{\beta}$ cannot be directly estimated from $PL(\beta)$. Alternatively, complete-case analysis, which disregards information from the missing data, does not work. It defines $PL(\beta)$ based only on (Y_j, δ_j, Z_j) for $j \in \mathbb{T}$, but the corresponding treatment indicator variable $Z_{i1} = 1$ for $i \in \mathbb{T}$. Consequently, β, especially β_1, cannot be estimated.

Due to these limitations, the most practical approach is to use a cross-sectional analysis without considering the survival time Y_i. For example, a logistic regression model can be used to model the probability of $\delta = 1$ as a function of Z. However, it completely ignores Y_i, even

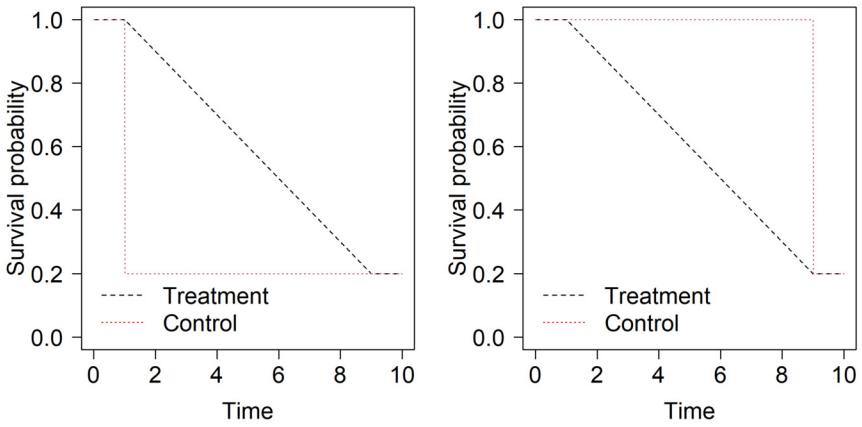

Figure 1: Hypothetical examples. Left: the survival probabilities for the treatment and control groups decrease linearly over time and drop once at time 1, respectively. Right: the survival probabilities for the treatment and control groups decrease linearly over time and drop once at time 9, respectively.

though Y_i is observed for $i \in \mathbb{T}$, which produces misleading results. Figure 1 demonstrates two extreme examples, where the left (or right) panel shows that the treatment group has a better (or worse) survival function than the control group. However, if the survival time is ignored, the event rates of the treatment and control groups are the same at 20%, which erroneously concludes that the treatment is not effective compared to the control, or vice versa.

2.3 Proposed Pseudo-Partial Likelihood Method

To overcome the challenge described in the previous subsection, we regard Y_i as a nuisance parameter η_i for $i \in \mathbb{C}$. This is equivalent to considering O_i as a nuisance parameter because O_i is expressed as $Y_i^O - \eta_i$ for $i \in \mathbb{C}$. We call η_i and $O_i(\eta_i)$ the pseudo survival time and pseudo time origin, respectively. By replacing Y_i with η_i in $PL(\beta)$, we propose a pseudo-partial likelihood, which is

$$\begin{aligned} PL^P(\beta, \eta) &= \prod_{a \in \mathbb{C}} \left\{ \frac{e^{\beta^\top Z_a}}{\sum_{i \in \mathbb{C}} I(\eta_i \geqslant \eta_a) e^{\beta^\top Z_i} + \sum_{j \in \mathbb{T}} I(Y_j \geqslant \eta_a) e^{\beta^\top Z_j}} \right\}^{\delta_a} \\ &\quad \times \prod_{b \in \mathbb{T}} \left\{ \frac{e^{\beta^\top Z_b}}{\sum_{i \in \mathbb{C}} I(\eta_i \geqslant Y_b) e^{\beta^\top Z_i} + \sum_{j \in \mathbb{T}} I(Y_j \geqslant Y_b) e^{\beta^\top Z_j}} \right\}^{\delta_b} \\ &= \prod_{i=1}^{n} \left\{ \frac{e^{\beta^\top Z_i}}{\sum_{j=1}^{n} I\{Y(\eta_j) \geqslant Y(\eta_i)\} e^{\beta^\top Z_j}} \right\}^{\delta_i}, \end{aligned}$$

where $\eta = (\eta_1, \eta_2, \ldots, \eta_m)$ is a m dimensional nuisance parameters, and $Y(\eta_i) = \eta_i$ for $i \in \mathbb{C}$. Here, $\eta_{m+1}, \eta_{m+2}, \ldots, \eta_n$ are not parameters because they are observed as $Y_{m+1}, Y_{m+1}, \ldots, Y_n$, respectively, but for notational convenience, we use the following notation: $Y(\eta_j) = \eta_j = Y_j$ for $j \in \mathbb{T}$.

Denote $(\tilde{\beta}, \tilde{\eta})$ as the maximizer of $PL^P(\beta, \eta)$ over $\mathbb{B} \times \mathbb{H} \in \mathbb{R}^p \times \mathbb{R}^m$. Let $\eta_0 = (Y_1, Y_2, \ldots, Y_m) \in \mathbb{H}$ be the true parameter. Since $PL(\beta)$ is expressed as $PL^P(\beta, \eta_0)$, $\tilde{\beta}$ is asymptotically equivalent to $\bar{\beta}$ as long as $\tilde{\eta}$ is a consistent estimator of η_0. In fact, this is still true even if $\tilde{\eta}$ is not

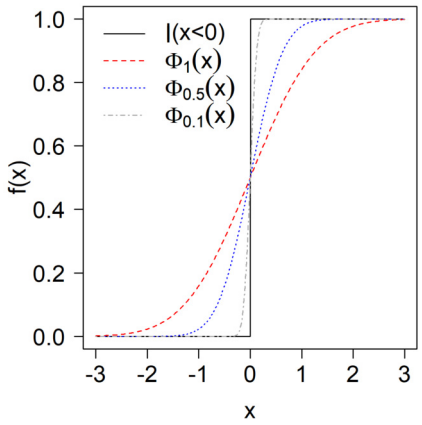

Figure 2: Indicator function $I(x) = I(x \leqslant 0)$ (solid) and cumulative normal density $\Phi_s(x)$ with $s = 1, 0.5, 0.1$ (dashed, dotted, dotdashed).

a consistent estimator. The nuisance parameter η is included only in the risk sets of $PL^P(\beta, \eta)$, and the asymptotic equivalent holds as long as

$$I\{Y(\tilde{\eta}_j) \geqslant Y(\tilde{\eta}_i)\} \tag{2}$$

is sufficiently close to $I(Y_j \geqslant Y_i)$ for $(i, j) \in (\mathbb{T}, \mathbb{C}), (\mathbb{C}, \mathbb{T}), (\mathbb{C}, \mathbb{C})$, where $Y(\tilde{\eta}_j) = \tilde{\eta}_j = Y_j$ for $j \in \mathbb{T}$.

In practice with a finite n, it is a challenge to compute $(\tilde{\beta}, \tilde{\eta})$ because $PL^P(\beta, \eta)$, where η is in the risk sets, is a non-differentiable step-function over $\mathbb{B} \times \mathbb{H}$. Alternatively, we consider an approximation of $I(x \geqslant 0)$ by $\Phi_s(x)$ on $x \in \mathbb{R}$, where $\Phi_s(x) = \Phi(x/s)$, and Φ is the standard normal cumulative density function. The scaling parameter s controls the approximation error, where $\Phi_s(x)$ approaches $I(x \geqslant 0)$ as s goes to zero, as shown in Figure 2. With this approximation, we propose a smoothed pseudo-partial likelihood, which is defined as

$$PL_s^P(\beta, \eta) = \prod_{i=1}^{n} \left\{ \frac{e^{\beta^\top Z_i}}{\sum_{j=1}^{n} I_s\{Y(\eta_j) \geqslant Y(\eta_i)\} e^{\beta^\top Z_j}} \right\}^{\delta_i},$$

where

$$I_s\{Y(\eta_j) \geqslant Y(\eta_i)\} = \begin{cases} \Phi_s\{Y(\eta_j) - Y(\eta_i)\} \text{ if } (i, j) \in (\mathbb{T}, \mathbb{C}), (\mathbb{C}, \mathbb{T}), (\mathbb{C}, \mathbb{C}), \\ I(Y_j \geqslant Y_i) \text{ otherwise.} \end{cases}$$

Denote $(\hat{\beta}, \hat{\eta})$ to be the maximizer of $PL_s^P(\beta, \eta)$ over $\mathbb{B} \times \mathbb{H}$. Since $PL_s^P(\beta, \eta)$ is a differentiable smooth function on $\mathbb{B} \times \mathbb{H}$, we can efficiently compute $(\hat{\beta}, \hat{\eta})$, which is further described in the subsection below.

Remark 2. Unlike $PL(\beta)$, which is a differential function, $PL^P(\beta, \eta)$ is a non-differentiable function due to the inclusion of the additional parameter η. Maximizing such non-differentiable functions, including $PL^P(\beta, \eta)$, is typically computationally challenging, but instead, maximizing $PL_s^P(\beta, \eta)$ is computationally efficiently. Under the regularity conditions, these two approaches are asymptotically equivalent, as formally established in Theorem 1.

2.4 Computation

Denote the logarithms of the pseudo-partial likelihood and smoothed pseudo-partial likelihood functions as

$$\ell_n^P(\beta,\eta) \equiv \log PL^P(\beta,\eta) = \sum_{i=1}^n \delta_i \left\{ \beta^\top Z_i - \log\left(\sum_{j=1}^n I_{ij}(\eta) e^{\beta^\top Z_j} \right) \right\},$$

$$\ell_{n,s}^P(\beta,\eta) \equiv \log PL_s^P(\beta,\eta) = \sum_{i=1}^n \delta_i \left\{ \beta^\top Z_i - \log\left(\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j} \right) \right\},$$

respectively, where $I_{ij}(\eta) = I\{Y(\eta_j) \geqslant Y(\eta_i)\}$, and $I_{ij,s}(\eta) = I_s\{Y(\eta_j) \geqslant Y(\eta_i)\}$. Maximizing $\ell_n^P(\beta,\eta)$ over $\mathbb{B}\times\mathbb{H}$ is computationally expensive because $\ell_n^P(\beta,\eta)$ is a non-differentiable function on $\mathbb{B}\times\mathbb{H}$. Instead, we propose to maximize $\ell_{n,s}^P(\beta,\eta)$, a differentiable function on $\mathbb{B}\times\mathbb{H}$, as stated in following steps:

Step 1. Set an initial value $(\beta^{(0)},\eta^{(0)}) \in \mathbb{B}\times\mathbb{H}$.

Step 2. Update $\beta^{(r)} = \beta^{(r-1)} - U\left(\beta^{(r-1)},\eta^{(r-1)}\right) H(\beta^{(r-1)},\eta^{(r-1)})^{-1}\left(\beta^{(r-1)},\eta^{(r-1)}\right)$.

Step 3. Update $\eta^{(r)} = \arg\max_{\eta\in\mathbb{H}} \ell_{n,s}^P(\beta^{(r)},\eta)$ using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

Step 4. Repeat Steps 2-3 for $r = 1, 2, \ldots,$ until

$$\left| \frac{\ell_{n,s}^P(\beta^{(r)},\eta^{(r)}) - \ell_{n,s}^P(\beta^{(r-1)},\eta^{(r-1)})}{\ell_{n,s}^P(\beta^{(r-1)},\eta^{(r-1)})} \right| < \xi,$$

for a small $\xi > 0$.

Step 2 utilizes an efficient Newton-Raphson algorithm with the score function and Hessian matrix as follows:

$$U(\beta,\eta) = \frac{\partial \ell_{n,s}^P(\beta,\eta)}{\partial\beta} = \sum_{i=1}^n \delta_i \left\{ Z_i - \frac{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j} Z_j}{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j}} \right\},$$

$$H(\beta,\eta) = \frac{\partial^2 \ell_{n,s}^P(\beta,\eta)}{\partial\beta^2} = -\sum_{i=1}^n \delta_i \left[\frac{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j} Z_j^{\otimes 2}}{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j}} - \frac{\{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j} Z_j\}^{\otimes 2}}{\{\sum_{j=1}^n I_{ij,s}(\eta) e^{\beta^\top Z_j}\}^2} \right],$$

with $z^{\otimes 2} = zz^\top$. In Step 3, the Newton-Raphson algorithm can also be used, but computing the corresponding Hessian matrix with respect to η and inverting it is time-consuming or numerically unstable, especially when m is large. Instead, the BFGS algorithm was used to stabilize the computation. Alternatively, L-BFGS-B algorithm can be used when constraints such as either $\eta_L < \eta_i$ or $\eta_i < \eta_U$ or both are imposed for $i \in \mathbb{C}$, where $\eta_L < \eta_U$ are finite constants. The *optim* function from the R *stats* package (R Core Team, 2025) can be used for the BFGS or L-BFGS-B algorithms.

Remark 3. In practice, only the lower bound $\eta_L = 0$ is typically imposed because selecting an appropriate upper bound η_U is not straightforward. Although this often results in a large value of $\hat{\eta}$, the corresponding estimated regression coefficient $\hat{\beta}$ tends to remain largely unaffected. This result is formally established in Theorem 2 below.

The proposed algorithm above is sensitive to the choice of the initial value because it is not guaranteed that $\ell_{n,s}^P(\beta,\eta)$ is a strictly concave function. As a result, $(\hat{\beta},\hat{\eta})$ may be a local maximizer of $\ell_{n,s}^P(\beta,\eta)$ over $\mathbb{B}\times\mathbb{H}$. Nevertheless, $\hat{\beta}$ could be the global maximizer of $PL(\beta)$ over

$\mathbb{B}$ if an initial value close to the true value (β_0, η_0) is chosen in Step 1. For example, given this initial value, suppose that $\eta^{(r)}$, which may or may not close to η_0, satisfies the condition (2). Then, maximizing $\ell^P_{n,s}(\beta, \eta^{(r)})$ over $\mathbb{B}$ in Step 2 is equivalent to maximizing $PL(\beta)$ over β, so $\beta^{(r)}$ should be the same as $\bar{\beta}$.

In general, it is challenging to select an initial value that is close to the true value, or it may often be impossible. However, we can make a reasonable guess in the following situations. Suppose that a new treatment, e.g., COVID-19 vaccine, was developed on the date of A^O, and the ith subject received the treatment on the date of O_i. Recall that O_i is the date of time origin, and survival time is defined as $Y_i = Y_i^O - O_i$. We further define $A_i = O_i - A^O$ as time from the treatment available to the treatment receive, and $B_i = Y_i^O - A^O$ as time from treatment available to an event of interest. Assuming Y_i is independent of A_i, we consider the following regression model, which is

$$A_i = \gamma_0 + \gamma_1 U_i + e_i, \tag{3}$$

where U_i, which can be a subset of Z_i, is observed for all subjects, and e_i is an error term. For example, U_i is an age variable if older people received the treatment earlier than younger people. Clearly, A and B_i are observed for all subjects. However, A_j is observed only for $j \in \mathbb{T}$ because O_i is missing for $i \in \mathbb{C}$. Since $B_i = A_i + Y_i$, we can reformulate

$$\eta_i \equiv \eta_i(\gamma_0, \gamma_1) = B_i - (\gamma_0 + \gamma_1 U_i), \tag{4}$$

for $i \in \mathbb{C}$. We thus estimate $\hat{\gamma}_0$ and $\hat{\gamma}_1$ using (A_j, U_j) for $j \in \mathbb{T}$, assuming the working model (3) is correctly specified for both groups, and set $\eta^{(0)} = (\eta_1(\hat{\gamma}_0, \hat{\gamma}_1), \ldots, \eta_m(\hat{\gamma}_0, \hat{\gamma}_1))$. Choosing β_0 is relatively straightforward. We use a logistic regression to model

$$\text{logit}(\Pr(\delta_i = 1 | Z_i)) = \alpha_0 + \alpha_1 Z_{i1} + \alpha_2 Z_{i2} + \cdots + \alpha_p Z_{ip}.$$

We then set $\beta^{(0)} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p)$ because the logistic regression is closely related to the hazard regression model (Efron, 1988).

2.5 Asymptotic Property

Let $\theta_0 = (\beta_0, \eta_0) \in \mathbb{B} \times \mathbb{H} = \Theta$ be the true parameter. Let $\bar{\beta}_n = \arg\max_{\beta \in B} \ell_n(\beta)$, where $\ell_n(\beta) = \log PL(\beta)$ is the logarithm of the partial likelihood function. Let $\tilde{\theta}_n = (\tilde{\beta}_n, \tilde{\eta}_n) = \arg\max_{\theta \in \Theta} \ell^P_n(\beta, \eta)$ and $\hat{\theta}_n = (\hat{\beta}_n, \hat{\eta}_n) = \arg\max_{\theta \in \Theta} \ell^P_{n,s}(\beta, \eta)$. Given a sufficient small s as $n \to \infty$, the smoothed risk set $I_{ij,s}(\eta)$ converges to the true risk sets $I_{ij}(\eta)$ for each $i, j = 1, 2, \ldots, n$. Thus, as $n \to \infty$, $\ell^P_{n,s}(\theta)$, the two maximizers $\tilde{\theta}_n$ and $\hat{\theta}_n$ are asymptotically equivalent, as stated in the theorem below.

Theorem 1. *Let $s = o\{1/\bar{\Phi}(\exp(-n^{-r}))\}$ for $r > 1$, where $\bar{\Phi}$ is the inverse of Φ. Under the assumptions A1-A4 stated in the Supplementary Material, $\ell^P_{n,s}(\theta)$ converges to $\ell^P_n(\theta)$ uniformly over Θ.*

Suppose that the proposed algorithm, described in Subsection 2.4, is used to estimate $\hat{\theta}_n$. Further suppose that the initial value of $\theta^{(0)}$ is sufficiently close to θ_0, so even though $\hat{\eta}_n$ is not a consistent estimator of η_0, they are close in terms of the the risk set approximation. This condition is formally stated in the Supplementary Material. Then the following theorem shows consistency and asymptotic normality for $\hat{\beta}_n$.

Theorem 2. *Under the assumptions A1-A4 stated in the Supplementary Material,* $\hat{\beta}_n$ *is an consistent estimator of* β_0*, and* $\sqrt{n}(\hat{\beta}_n - \beta_0)$ *converges to* $N(0, \Sigma(\beta_0, \eta_0))$ *as* $n \to \infty$*, where* $V(\beta_0, \eta_0) = -n^{-1}E\left(H(\beta_0, \eta_0)\right)$.

A consistent estimator of the asymptotic variance is $\hat{\Sigma}(\hat{\beta}_n, \hat{\eta}_n) = -n^{-1}H(\hat{\beta}_n, \hat{\eta}_n)$. An asymptotic 95% is computed by $\hat{\beta} \pm 1.96\sqrt{\text{diag}(\hat{\Sigma}(\hat{\beta}_n, \hat{\eta}_n))}$, where $\text{diag}(\Sigma)$ is the diagnonal element of Σ. The proofs of Theorems 1–2 are provided in the Supplementary Material.

3 Simulation Studies

We designed and conducted simulation studies to compare the performances of our proposed pseudo Cox versus standard models. We considered the following two scenarios under the proportional hazards model:

Scenario 1: $\lambda(t \mid Z) = \lambda_0(t)e^{\beta_1 Z_1}$,

Scenario 2: $\lambda(t \mid Z) = \lambda_0(t)e^{\beta_1 Z_1 + \beta_2 Z_2}$.

The treatment group variable Z_1 was generated from a binomial distribution with p, where $Z_1 = 1$ and $Z_1 = 0$ indicate treatment and control group, respectively, Z_2 was independently generated from $|N(0, 1)|$. We set $p = 0.9$, so approximately 10% of the samples were control samples with missing survival times, and $(\beta_1, \beta_2) = (1, 1)$. Under each scenario, the failure time was then generated using a Weibull baseline hazard function with shape parameter 3 and scale parameter 100. The censoring time was independently generated from various Weibull distributions with shape parameter 1 and a scale parameter ranging from 28 to 282, which led to censoring rates of approximately 20%, 40%, 60%, and 80%. We then replaced Y with missing values for the control group if $Z_1 = 0$ and fit the following two heuristic models: complete-case Cox model with (Y, δ) if $Z_1 = 1$, and naive Cox model with $(\hat{Y}, \delta)$, where $\hat{Y} = Y$ if $Z_1 = 1$ or $\hat{Y}$ was generated from Uniform distribution on 0 to $\max(YI(Z_1 = 1))$. We also fit our proposed model with the stopping criteria of $\xi = 0.01$. The scale parameter of $s = 0.01/\bar{\Phi}(\exp(-n^{-2}))$ was set by Theorem 1, and the initial value was chosen using the method described in Section 2.4 by generating U and e from $N(0, 1)$ independently and setting $\gamma_0 = \gamma_1 = 10$. We repeated the simulations 100 times with sample sizes of 200, 500, 1000. We evaluated each model by computing the bias, root mean squared error (RMSE), and coverage probability of nominal 95% confidence intervals over the 100 replicates. We also computed CPU time over 100 simulations using 2X AMD EPYC 7713 Zen3 processor.

Table 1 contains the simulation results. Our proposed method performed consistently well across all scenarios, demonstrating a clear reduction in bias and RMSE as the sample size increased, along with coverage probabilities close to the nominal 95%. On the other hand, as expected, the complete-case Cox model was unable to estimate β_1 because it disregarded all information from the control samples. The naive Cox model was able to estimate β_1, but the resulting estimate was biased due to the arbitrary imputation of $\hat{Y}$ from the Uniform distribution without theoretical justification. For both the complete-case and naive Cox models, estimating β_2 was feasible in Scenario 2 because the simulated data was generated under the missing completely at random assumption. However, neither model was appropriate for our main goal of estimating β_1 accurately, controlling for Z_2.

The average CPU time over 100 repetitions was less than one second for the two competitors because the *glm* and *coxph* functions from the R *stats* and *survival* packages (Therneau, 2024) efficiently maximized the corresponding likelihood functions with at most two regression parameters β_1 and β_2. In contrast, the proposed method required more computational resources to

Table 1: Simulation results: bias/RMSE/coverage probability, multiplied by 100. Cens: censoring rate, CC: complete-case, naive Cox: Cox model with random time origin; pseudo Cox: proposed pseudo Cox model. CPU time in sections for pseudo Cox.

Scenario	Cens	n	Parameter	CC Cox	Naive Cox	Pseudo Cox	CPU
1	80%	250	β_1	-	75/99/89	−19/63/91	0
		500	β_1	-	101/113/42	−5/49/96	2
		1000	β_1	-	109/113/2	−6/29/95	12
	60%	250	β_1	-	21/49/97	−5/37/96	1
		500	β_1	-	39/50/73	−7/30/94	5
		1000	β_1	-	48/54/37	−7/20/96	30
	40%	250	β_1	-	−23/43/82	−7/32/92	2
		500	β_1	-	−4/29/85	−9/25/93	8
		1000	β_1	-	10/23/86	−9/17/90	51
	20%	250	β_1	-	−47/58/53	−8/29/91	2
		500	β_1	-	−31/39/53	−11/22/88	12
		1000	β_1	-	−19/26/59	−12/17/87	79
2	80%	250	β_1	-	105/129/66	−2/64/93	1
			β_2	−1/26/93	−6/26/92	0/25/91	
		500	β_1	-	135/145/11	−2/45/93	5
			β_2	1/18/93	−4/16/92	2/16/97	
		1000	β_1	-	150/155/1	2/30/98	17
			β_2	2/11/97	−3/11/96	4/11/94	
	60%	250	β_1	-	59/78/69	0/41/96	2
			β_2	0/19/93	−7/19/93	1/18/95	
		500	β_1	-	84/93/26	−2/27/93	8
			β_2	1/11/97	−8/13/93	2/10/99	
		1000	β_1	-	94/98/2	1/18/96	45
			β_2	1/8/96	−9/12/81	3/8/96	
	40%	250	β_1	-	21/50/82	−2/31/94	2
			β_2	0/16/94	−11/18/87	1/14/94	
		500	β_1	-	45/57/53	−2/23/96	12
			β_2	2/10/98	−10/14/82	3/10/98	
		1000	β_1	-	61/66/12	−1/14/98	77
			β_2	0/8/95	−13/15/56	2/7/94	
	20%	250	β_1	-	−1/41/79	−4/26/93	4
			β_2	0/15/93	−13/19/76	1/13/95	
		500	β_1	-	19/35/74	−4/20/92	17
			β_2	1/10/94	−14/17/62	2/9/97	
		1000	β_1	-	33/41/44	−2/12/94	106
			β_2	0/7/94	−16/18/29	1/6/94	

deal with the additional m dimensional nuisance parameter η; however, this did not not impose a significant computational burden, as demonstrated in Table 1. In the Supplementary Material, we conducted additional simulations with $p = 0.7$ and got slightly improved results because the numbers of subjects in the control and treatment groups were more balanced.

Remark 4. In addition to the CC and naive Cox models, the Weibull proportional hazard method (Chen et al., 2024) could also be considered for comparison in this simulation. However, a fair comparison may not be feasible because the Weibull model assumes access to the true value for initializing the algorithm.

4 ASU SARS-CoV-2 Serological Prevalence Survey Data

Arizona State University conducted a SARS-CoV-2 serosurvey study from September 13th to September 17th, 2021, to determine the seroprevalence of SARS-CoV-2 antibodies in the ASU community (Hou et al., 2023). The study recruited 1064 healthy students and employees aged 18 to 72 years across four ASU campuses (Downtown Phoenix, Polytechnic, Tempe, and West) and collected their saliva and blood samples for SARS-CoV-2 antibody testing. Additionally demographic and self-reported vaccination variables were collected for each participant. Approximately 52% of the participants were female; 71% were aged 18 to 25, 18% were aged 26 to 40, and 8% were 40 years and older; 50% were White, 27% were Asian, 3% were Black, 1.3% were Native, and 15% identified as Other; and 92% were fully vaccinated with Pfizer (48%), Moderna (29%), Janssen (9%), AstraZeneca (4%), Covaxin (0.9%), Sinopharm (0.2%), or Sinovac (0.1%) COVID-19 vaccines. Among these vaccines, the first three were approved for Emergency Use Authorization by the US Food and Drug Administration. After excluding non-responders, we used a subset of the data consisting of 747 participants, of whom 656 received full dose(s) of Pfizer, Moderna or Janssen and 81 did not receive any vaccines.

Of the 747 participants, 112 were infected with 71/656=0.11 (95% CI: (0.09,0.13)) and 41/81=0.51 (95% CI: (0.40,0.61)) infection rates for the vaccinated and unvaccinated groups, respectively. We used a logistic regression to model

$$\text{logit}(\Pr(\delta = 1|Z)) = \alpha_0 + \alpha_1 \text{Vac} + \alpha_2 \text{Age1} + \alpha_3 \text{Age2} + \alpha_4 \text{Sex} + \alpha_5 \text{Race1} + \alpha_6 \text{Race2},$$

where δ is an infection status (1 for infected, 0 for not infected), Vac is the vaccination status (1 for vaccinated, 0 for unvaccinated), Age1 and Age2 are dummy variables representing the 26-40 and 40 and above age categories, respectively (reference group is age 18 to 25), Sex=0 for male and Sex=1 for female, Race1 and Race2 are dummy variables representing the White and Asian categories (reference group is Black, Native, or Other). The results showed that vaccination significantly reduced the rate of infection, after controlling for the other variables ($\hat{\alpha}_1 = -2.16$, p-value <0.01). The other predictors were not significant, as summarized in Table 2.

The cross-sectional analysis above is useful but limited because it completely ignores the survival time, as discussed in Section 3 and Subsection 2.2. Evaluation of vaccine efficacy over time provides useful information. For example, this supports decision-making regarding when to recommend additional doses since the study was conducted prior to the official availability of booster doses. We considered a time-to-infection endpoint, defined as the number of days from vaccine completion to infection. Naturally, the corresponding time origin does not exist for unvaccinated individuals because they did not receive any vaccinations. There is no additional information available regarding these missing time origins. By treating the missing time origins

Table 2: Logistic regression versus pseudo Cox proportional models with main effects.

	Logistic regression			Pseudo Cox proportional model		
Variable	Estimate	95% CI	P-value	Estimate	95% CI	P-value
Intercept	0.17	(−0.37, 0.72)	0.53	-	-	-
Vac	−2.17	(−2.71, −1.65)	<0.01	−2.14	(−2.59, −1.69)	<0.01
Age1	−0.10	(−0.65, 0.42)	0.72	−0.80	(−1.29, −0.30)	<0.01
Age2	−0.47	(−1.38, 0.32)	0.27	−1.51	(−2.27, −0.75)	<0.01
Sex	−0.21	(−0.65, 0.23)	0.35	−0.23	(−0.60, 0.15)	0.24
Race1	0.08	(−0.49, 0.63)	0.78	0.08	(−0.43, 0.58)	0.77
Race2	0.19	(−0.37, 0.74)	0.49	0.32	(−0.16, 0.79)	0.19

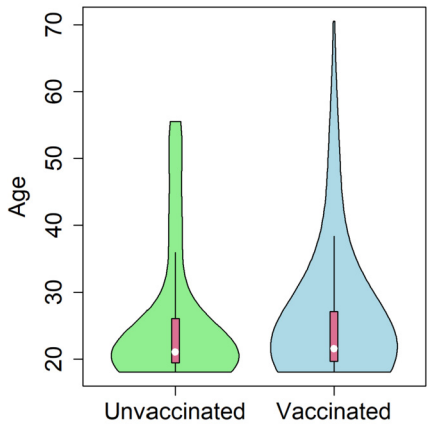

Figure 3: Violin plot of age between the unvaccinated and vaccinated groups.

as nuisance parameters, we used the proposed pseudo-Cox proportional hazards model with the same predictors listed above. The initial value was determined using the method described in Section 2.4. Specifically, we set $\beta^{(0)} = \hat{\alpha}$, assuming a relationship between the logistic regression and the Cox model. For the initial value of $\eta^{(0)}$, we set A^O as December 11, 2020 when the Pfizer COVID-19 vaccine received Emergency Use Authorization (EUA) from the U.S. Food and Drug Administration (FDA). We then fit the linear regression model (3) to estimate the time from the availability of first vaccine to infection using the age predictor variable This approach estimated $\hat{\gamma}_0 = 170.97$ and $\hat{\gamma}_1 = -1.73$ because Pfizer as well as other COVID-19 vaccines were distributed to older adults prior to younger adults. Finally, we set $\eta^{(0)}$ using the relationship (4). Table 1 summarized the results of the pseudo-Cox model, where the variables Vac, Age1, and Age2 were significant. However, both Age1 and Age2 had negative coefficients, implying older people were less infected than younger people. The logistic regression we used above also showed negative coefficients for Age1 and Age2, but they were not significant. These results could be misleading because older people received more vaccinations than younger people did, as shown in Figure 3. Thus, it was vaccines, not age, that protected people from being infected with the SARS-CoV-2 virus.

To avoid Simpson's paradox, we alternatively fit the pseudo Cox model for each age group,

Table 3: Subgroup analysis by age: logistic regression versus pseudo Cox proportional models.

		Logistic regression			Pseudo Cox proportional model		
Age	Variable	Estimate	95% CI	P-value	Estimate	95% CI	P-value
18-25	Intercept	0.27	(−0.24, 0.79)	0.30	-	-	-
	Vac	−2.37	(−2.97, −1.79)	<0.01	−2.36	(−2.89, −1.84)	<0.01
26-40	Intercept	−0.98	(−2.50, 0.26)	0.15	-	-	-
	Vac	−0.99	(−2.33, 0.59)	0.17	−1.24	(−2.51, 0.04)	0.06
41-70	Intercept	−0.41	(−1.77, 0.85)	0.53	-	-	-
	Vac	−2.20	(−3.88, −0.56)	0.01	−3.36	(−5.57, −1.15)	<0.01

separately. This subgroup analysis consistently demonstrated that vaccination was effective against to infection across all age groups, with statistically significant p-values except for the age group of 26-40 years old. For comparison, we also conducted a subgroup analysis using logistic regression, as summarized in Table 3. The two models yield consistent results in terms of significant p-values; however, the regression coefficients from the logistic regression may be biased due to the omission of survival time. For this comparison, given the reduced sample size, each subgroup model included only the vaccination variable.

5 Discussion

In this article, we proposed a novel pseudo-partial-likelihood approach for estimating treatment effects for right-censored time-to-event data when the time origin is undefinable for the control group. By treating these time origins as nuisance parameters η, we developed an efficient computational method that estimates both the regression coefficients and time origin without requiring simultaneous estimation of the baseline hazard functions. However, this flexibility may introduce a curse of dimensionality, particularly when the number of control samples is large. To address this issue, we limited the proportion of missing time origins to a modest, fixed level, e.g., less than or equal to 30% missing time origins, based on results from our simulation studies. Regardless of the missing rate, it is important to emphasize that while our method does not provide a consistent estimator for $\hat{\eta}$, it does yield consistent estimates of the regression coefficient $\hat{\beta}$, as shown in Theorem 2, demonstrating the robustness of the approach. In our application to data from the ASU SARS-CoV-2 serosurvey study, both the standard logistic regression and the proposed models produced consistent conclusions. Nevertheless, such agreement may not always be observed. In those instances, careful consideration is warranted, and model selection should be guided by the scientific question of interest rather than statistical significance.

Several alternatives are available for analyzing the SARS-CoV-2 serosurvey data. One such approach involves selecting an arbitrary time origin, such as January 20, 2020, the date on which the first COVID-19 case in the United States was confirmed by the Centers for Disease Control and Prevention (CDC). A corresponding survival time is well-defined survival time for all participants, and the standard time-dependent Cox model can be directly used to estimate time-varying vaccine effects. Although this approach is analytically appropriate in some contexts, its

results may be misleading if the primary outcome of interest is the time from vaccine completion to infection. In such cases, the issue of missing time origins reemerges, and extending this method to accommodate the relevant outcome is not straightforward.

For observational studies, where subject characteristics are expected to be imbalanced between the treatment and control groups, our method can be extended by maximizing a weighted pseudo-likelihood function,

$$\prod_{i=1}^{n}\left\{\frac{e^{\beta^\top Z_i}}{\sum_{j=1}^{n} I_{ij,s}(\eta)e^{\beta^\top Z_j}}\right\}^{w_i\delta_i},$$

where $w_i = Z_{1i}/\pi_i + (1-Z_{1i})/(1-\pi_i)$ is the inverse probability of treatment weighting (IPTW), and $\pi_i = \Pr(Z_{1i} = 1|Z_{i2}, Z_{i3}, \ldots, Z_{ip})$ is the propensity score, or the probability of receiving the treatment given the observed covariates $Z_{i2}, Z_{i3}, \ldots, Z_{ip}$. The propensity score can be estimated using logistic or probit regression models (Rosenbaum et al., 2010), even when time origins are missing. The IPTW can adjust for the imbalance of the observed variables between the treatment and control groups. However, a more rigorous approach would be needed to examine the validity of IPTW particularly in the absence of a well defined time origin.

Supplementary Material

Sections A and B of the Supplementary Material provide the proofs of Theorems 1–2 and additional simulation results, respectively. The SARS-CoV-2 serological prevalence data and corresponding R code used for analysis are also included in the Supplementary Material. The *coxphm* package (Chung, 2025), which implements the methods developed in this article, is publicly available on CRAN.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, and the two reviewers for their constructive comments, which have significantly enhanced the manuscript.

Funding

The work was supported by funding from Arizona State University Knowledge Enterprise.

References

Anand S, Montez-Rath M, Han J, Bozeman J, Kerschmann R, Beyer P, et al. (2020). Prevalence of SARS-CoV-2 antibodies in a large nationwide sample of patients on dialysis in the USA: a cross-sectional study. *The Lancet*, 396(10259): 1335–1344. https://doi.org/10.1016/S0140-6736(20)32009-2

Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al. (2021). Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *New England Journal of Medicine*, 384(5): 403–416. https://doi.org/10.1056/NEJMoa2035389

Chen DG, Chung Y, Beyene KM (2024). Estimate time-to-infection (TTI) vaccination effect when TTI for unvaccinated group is unknown. *Statistics in Biosciences*, 16(3): 723–741. https://doi.org/10.1007/s12561-024-09417-w

Chung Y (2025). coxphm: Time-to-Event Data Analysis with Missing Survival Times. R package version 0.2.1.

Cox DR (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B*, 34(2): 187–220. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x

Efron B (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402): 414–425. https://doi.org/10.1080/01621459.1988.10478612

Fleming TR, Harrington D (2013). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Havers FP, Reed C, Lim T, Montgomery JM, Klena JD, Hall AJ, et al. (2020). Seroprevalence of antibodies to SARS-CoV-2 in 10 sites in the United States, March 23-May 12, 2020. *JAMA Internal Medicine*, 180(12): 1576–1586. https://doi.org/10.1001/jamainternmed.2020.4130

Hou CW, Williams S, Taylor K, Boyle V, Bobbett B, Kouvetakis J, et al. (2023). Serological survey to estimate SARS-CoV-2 infection and antibody seroprevalence at a large public university: a cross-sectional study. *BMJ Open*, 13(8): e072627. https://doi.org/10.1136/bmjopen-2023-072627

Lombardi A, Mangioni D, Consonni D, Cariani L, Bono P, Cantù AP, et al. (2021). Seroprevalence of anti-SARS-CoV-2 IgG among healthcare workers of a large university hospital in Milan, Lombardy, Italy: a cross-sectional study. *BMJ Open*, 11(2): e047216. https://doi.org/10.1136/bmjopen-2020-047216

Mercado-Reyes M, Malagón-Rojas J, Rodríguez-Barraquer I, Zapata-Bedoya S, Wiesner M, Cucunubá Z, et al. (2022). Seroprevalence of anti-SARS-CoV-2 antibodies in Colombia, 2020: a population-based study. *The Lancet Regional Health–Americas*, 9: 100195. https://doi.org/10.1016/j.lana.2022.100195

Moreira-Soto A, Pachamora Diaz JM, González-Auza L, Merino Merino XJ, Schwalb A, Drosten C, et al. (2021). High SARS-CoV-2 seroprevalence in rural Peru, 2021: a cross-sectional population-based study. *Msphere*, 6(6): e00685-21.

Nah EH, Cho S, Park H, Hwang I, Cho HI (2021). Nationwide seroprevalence of antibodies to SARS-CoV-2 in asymptomatic population in South Korea: a cross-sectional study. *BMJ Open*, 11(4): e049837. https://doi.org/10.1136/bmjopen-2021-049837

Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. (2020). Safety and efficacy of the BNT162b2 mRNA COVID-19 vaccine. *New England Journal of Medicine*, 383(27): 2603–2615. https://doi.org/10.1056/NEJMoa2034577

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rosenbaum PR, Rosenbaum P, Briskman (2010). *Design of Observational Studies*. Springer.

Therneau TM (2024). survival: Survival Analysis. R package version 3.8-3.

Venugopal U, Jilani N, Rabah S, Shariff MA, Jawed M, Batres AM, et al. (2021). SARS-CoV-2 seroprevalence among health care workers in a New York City hospital: a cross-sectional analysis during the COVID-19 pandemic. *International Journal of Infectious Diseases*, 102: 63–69. https://doi.org/10.1016/j.ijid.2020.10.036

Vusirikala A, Whitaker H, Jones S, Tessier E, Borrow R, Linley E, et al. (2021). Seroprevalence of SARS-CoV-2 antibodies in university students: cross-sectional study, December 2020, England. *Journal of Infection*, 83(1): 104–111. https://doi.org/10.1016/j.jinf.2021.04.028

Xiong Y, Braun WJ, Hu XJ (2021). Estimating duration distribution aided by auxiliary longitudinal measures in presence of missing time origin. *Lifetime Data Analysis*, 27: 388–412. https://doi.org/10.1007/s10985-021-09520-w