Journal of the Korean Statistical Society
https://doi.org/10.1007/s42952-025-00359-7

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Analysis of randomized response survival data in discrete-time

Jian-Hong Wu-Lin[1] · **Chi-Chung Wen**[2]

Received: 9 May 2025 / Accepted: 30 November 2025

Abstract

Event time data collected in sensitive survey studies are often subject to case-I interval censoring and response misreporting. To mitigate these issues, such data are frequently gathered as current status data using the randomized response technique (RRT). When event times are recorded on a discrete time scale, applying continuous-time methods to RRT data may result in substantial bias, underscoring the need for methodologies specifically tailored to discrete-time analysis. In this study, we propose a discrete time-to-event analysis framework for current status data obtained via the unrelated-question RRT. The event time is modeled using a general discrete-time transformation model, which encompasses widely used formulations such as the proportional continuation ratio and grouped proportional hazards models, with the baseline hazard specified in either a discrete or smooth form. Recognizing that the exact failure time and the indicator of whether a respondent answered the sensitive question are latent under the current status censoring and RRT design, we develop a novel, tailored expectation–maximization algorithm for efficient computation. We establish the asymptotic properties of the proposed estimators and demonstrate the utility of the method through comprehensive simulation studies and applications to two real-world datasets.

Keywords Discrete-time hazard · Expectation–maximization · Maximum likelihood · Sensitive issue

✉ Chi-Chung Wen
ccwen@nycu.edu.tw

[1] Department of Mathematics, Tamkang University, No. 151, Yingzhuan Rd., Tamsui Dist., New Taipei City 251301, Taiwan

[2] Institute of Public Health, National Yang Ming Chiao Tung University, No. 155, Sec. 2, Linong St., Beitou Dist., Taipei City 112, Taiwan

Published online: 06 January 2026

1 Introduction

For many measurements, such as surveys, event time processes are subject to interval censoring because observations are not made continuously but periodically or intermittently. In particular, if only one measurement occurs during the observation period, the studied event time process is said to have "case-I" interval censoring (Sun 2006; Chen et al. 2012), and the collected event time data are reffered to as "current status data" (Huang 1996; Jewell and van der Laan 2003). These data include only the survey time and whether the event has occurred by that time. For example, current status data for the event of extramarital sex in married individuals could be collected in a survey by asking the respondents the following questions

C: *How many years have you been married?*
A: *Have you ever had sex with someone other than your spouse?*

In addition to censoring, event process data collected from survey studies may be subject to response bias, especially if the questions pertain to private matters in one's personal life. Respondents often provide untruthful answers to questions regarding moral, legal, or other sensitive issues, such as extramarital sex. Hence, response data from direct questions on sensitive issues can be seriously biased. To mitigate untruthful responses to sensitive questions, various randomized response techniques (RRTs) have been developed (e.g. Warner 1965; Greenberg et al. 1969; Horvitz et al. 1967; Kuk 1990; Singh et al. 2000; Gjestvang and Singh 2006). One popular and efficient method is the unrelated-question RRT of Greenberg et al. (1969), where questionnaires pose a sensitive question A and an innocuous question B; for example:

A: *Have you ever had sex with someone other than your spouse?*
B: *Were you born in January, February, or March?*

Whether the respondent answers the sensitive or innocuous question is determined by using a random device, such as a coin or a deck of cards. The respondent only answers the selected question without revealing to the interviewer which question was selected, ensuring confidentiality and respondent willingness to answer truthfully. Event times in survey studies are usually observed as current status data, and if the event of interest is sensitive, an RRT is often used to collect these data. Although RRT has been widely applied to increase the accuracy of estimates of the prevalence (Warner 1965; Greenberg et al. 1969) and covariate-adjusted prevalence (Scheers and Dayton 1988; Hsieh et al. 2016) of sensitive attributes, it has rarely been applied to estimating an event time distribution. Recently, Wen and Chen (2025) developed a regression approach to analyze current status event time data collected by the RRT; however, their method has only been explored under a continuous event time framework.

In survey studies, the scale of event time data is also an important consideration. Time-to-event analysis is conventionally performed on a continuous-time scale, where the event time is treated as a continuous random variable. However, in practice, the event time is often measured on a discrete-time scale. For example, event times are discrete if they are grouped or rounded in the survey for convenience or, as is the case in the question“ How many years have you been married?”, observations of event times are made only at discrete time points; hence, the underlying event times are measured discretely. Simulations (Table 3) have demonstrated that a naive application of the continuous-time method of Wen and Chen (2025) to grouped continuous-time RRT data can lead to substantial bias if tied or grouped observations are not properly handled. Therefore, a methodology for discrete-time data must be developed in addition to the numerous methodologies for continuous-time data in the literature (Tutz and Schimid 2016).

In this work, we formulate a method for analyzing event time data that addresses these aforementioned challenges. Specifically, for a discrete-time setting, we propose a maximum likelihood method for the analysis of current status event time data collected using the unrelated-question RRT in survey studies. The discrete-time hazard rate for an event is given by a general discrete-time transformation regression model, which includes the logistic discrete hazard model or the proportional continuation ratio (PCR) model (Thompson 1977), and the grouped proportional hazards (GPH) model (Prentice and Gloeckler 1978) as special cases. The baseline hazard in the proposed discrete-time model is specified by a unique intercept for each time point. As mentioned by Berger and Schmid (2018), if the number of time points (i.e., the number of intercept parameters) is large relative to the sample size, the event counts at some time points may be small, resulting in unstable estimation. To address this problem, the method was further developed for models with a baseline hazard that is approximated (or ‘specified’) by smooth functions with fewer parameters. The use of smooth approximations effectively reduces the dimensionality of the parameters, stabilizing the estimation. An asymptotic theory (along with explicit variance estimators) of the proposed estimators was established for transformation models with a specified discrete and smooth baseline hazard. In contrast to discrete right-censored data for which general discrete-time models can be easily proposed using open-source statistical software (Berger and Schmid 2018), no computational software is available for analyzing discrete current-status RRT data without user configuration. Hence, we developed a novel expectation–maximization (EM) computation procedure comprising the self-consistency and Newton-Raphson algorithms for the proposed estimators. The finite sample properties of the proposed estimators were examined in simulation studies. Finally, we illustrate practical applications of our method through examples from an extramarital sex study and a militant connection study. In both studies, current status data for the events of interest were collected using the RRT under a discrete-time scale.

This paper is organized as follows. Section 2 introduces the data and the model. Sections 3 and 4 describe the estimation procedures for models with the baseline hazard specified by discrete intercepts and smooth functions, respectively. Section 5 evaluates the proposed method through simulation studies, and Sect. 6 presents the practical applications using extramarital sex data and militant connection data. Section 7 contains some concluding remarks and the “Appendix” provides the variance estimation for the proposed estimators.

2 Data and model descriptions

Suppose that current status data for a sensitive event are collected using unrelated-question RRT (Greenberg et al. 1969). In this method, either a sensitive question A regarding the event or an innocuous question B, determined randomly, is posed to the respondent.

For respondent $i, i = 1, \ldots, n$, let T_i, C_i, and Z_i be the sensitive event time of interest, the survey time, and the covariate vector, respectively. We assume that T_i takes discrete values in $\{1, \ldots, J+1\}$ and C_i takes values in $\{1, \ldots, J\}$, where J is a positive integer. The variable $\Delta_i \equiv I(T_i \le C_i)$ indicates whether the sensitive event has occurred by the survey time C_i; this Δ_i represents the response to the sensitive question A at C_i, where $I(\cdot)$ is the indicator function. Let W_i be the 0–1 binary response to the innocuous question B with probability $\Pr(W_i = 1) = c$. For the unrelated-question RRT, whether sensitive question A or innocuous question B is posed to each participant is determined based on a latent binary variable Q_i; the probability that question A is posed to any respondent is $\Pr(Q_i = 1) = q$. Instead of observing T_i, we can only observe $Y_i = Q_i\Delta_i + (1 - Q_i)W_i$. The observed survey data comprise $\{O_i = (Y_i, C_i, Z_i) | i = 1, \ldots, n\}$.

Given the covariate Z_i, the sensitive event time T_i is assumed to follow a general discrete-time transformation model. Specifically, at time t, the discrete hazard rate of T_i given Z_i is of the form

$$\Pr(T_i = t | T_i \ge t, Z_i = z) = \frac{G(\gamma_t + \beta' z)}{1 + G(\gamma_t + \beta' z)}, \quad t = 1, \ldots, J, \tag{1}$$

where G is a known increasing function, β is an unknown vector of regression parameters, and $\gamma = (\gamma_1, \ldots, \gamma_J)'$ is an unknown vector of baseline coefficients. By definition, the hazard of T_i at time $J+1$ is 1. The choice of $G(x) = \exp(x)$ yields the PCR model and $G(x) = \exp(e^x) - 1$ yields the GPH model. Both models are convenient for interpreting the effect of a given covariate on event time. For PCR, the continuation ratio

$$\frac{\Pr(T_i = t \mid Z_i)}{\Pr(T_i > t \mid Z_i)}$$

is proportional across covariate groups, whereas for GPH, the log-survival probability

$$\log \Pr(T_i > t \mid Z_i)$$

is proportional (Tutz and Schimid 2016). Thus, the choice between PCR and GPH depends on whether the covariate effects are more plausibly represented through continuation ratios or through log-survival ratios. In practice, this decision can be guided by both substantive considerations (e.g., which interpretation is more meaningful for the research question) and empirical model diagnostics (e.g., AIC or BIC comparisons).

Throughout the study, we assume that T_i and C_i are conditionally independent given Z_i (non-informative censoring), W_i, Q_i and (T_i, C_i, Z_i) are independent, and that the respondents answer the sensitive RRT question honestly. We further assume that the probability q of the respondent selecting the sensitive question from the question set and the proportion c of the respondent answering "yes" to the innocuous question are known and can be handled in the survey design. The probability q can be determined through a deliberately designed game of chance, and the proportion c can be obtained from available group-level data before the survey.

3 Parameter estimation

In this section, we discuss the estimation for the discrete hazard model (1) based on unrelated-question RRT data $\{O_i = (Y_i, C_i, Z_i) | i = 1, \ldots, n\}$. Under non-informative censoring and the assumptions that $\Pr(W_i = 1 | T_i, C_i, Z_i) = c$ and $\Pr(Q_i = 1 | C_i, Z_i) = q$, we have

$$\begin{aligned}\Pr(Y_i = 1 | C_i, Z_i) =& \Pr(Y_i = 1 | C_i, Z_i, Q_i = 1)q + \Pr(Y_i = 1 | C_i, Z_i, Q_i = 0)(1 - q) \\ =& \Pr(\Delta_i = 1 | C_i, Z_i)q + \Pr(W_i = 1 | C_i, Z_i)(1 - q) \\ =& q\left[1 - \prod_{j \le C_i} \{1 + G(\gamma_j + \beta' Z_i)\}^{-1}\right] + (1 - q)c.\end{aligned}$$

Let $G_{i,j} = G(\gamma_j + \beta' Z_i)$ and $S_{i,t} = \prod_{j \le t} \{1 + G_{i,j}\}^{-1}$. The likelihood function of $(O_1, \ldots, O_n)$ takes the form

$$L_n(\theta) = \prod_{i=1}^{n} L(\theta)(O_i), \tag{2}$$

where $L(\theta)(O_i) = \{q(1 - S_{i,C_i}) + (1 - q)c\}^{Y_i} \{1 - q(1 - S_{i,C_i}) - (1 - q)c\}^{1 - Y_i}$ and $\theta = (\beta', \gamma')'$. The maximum likelihood estimator (MLE) $\widehat{\theta} = (\widehat{\beta}', \widehat{\gamma}')'$ of θ maximizes the likelihood in (2).

To compute the MLE, we propose the EM method. In the absence of randomized response sampling, the survey data set $\{Y_i, C_i, Z_i | i = 1, \ldots, n\}$ reduces to a current status data set $\{\Delta_i, C_i, Z_i | i = 1, \ldots, n\}$ (Huang 1996). For continuous-time data, Turnbull (1976) proposed a self-consistency formula for computing the MLE, which is essential for an EM algorithm based on current status data without covariates. Herein we propose a novel EM algorithm that extends Turnbull's method to regression analysis with our randomized response survival data but under a discrete-time setting. Let $N_{i,j} = I(T_i = j), j = 1, \ldots, J$. We treat the failure time indicator $N_{i,j}$ and the latent indicator Q_i of selecting the sensitive question in the unrelated-question set as missing values in the EM method. The complete-data likelihood of $\{(Y_i, C_i, Z_i, N_{i,j}, Q_i), i = 1, \ldots, n, j = 1, \ldots, J\}$ takes the form

$$L_n^c(\theta) = \prod_{i=1}^{n} \left\{ \left[q \prod_{j \le C_i} (G_{i,j} S_{i,j})^{N_{i,j}} \right]^{Q_i Y_i} [q S_{i,C_i}]^{Q_i(1-Y_i)} \times [(1-q)c]^{(1-Q_i)Y_i} [(1-q)(1-c)]^{(1-Q_i)(1-Y_i)} \right\}.$$

In the maximization step (M-step), we maximize

$$\sum_{i=1}^{n} \left\{ Y_i \sum_{t_j \le C_i} \widehat{N_{i,j} Q_i} \log(G_{i,j} S_{i,j}) + (1-Y_i) \widehat{Q}_i \log(S_{i,C_i}) + \widehat{Q}_i \log q + (1-\widehat{Q}_i) Y_i \log[(1-q)c] + (1-\widehat{Q}_i)(1-Y_i) \log[(1-q)(1-c)] \right\}, \quad (3)$$

where $\widehat{Q_i} = E\{Q_i | Y_i, C_i, Z_i\}$ and $\widehat{N_{i,j} Q_i} = E\{N_{i,j} Q_i | Y_i, C_i, Z_i\}$. If $Y_i = 1$, then

$$\widehat{Q_i} = \frac{\Pr(Y_i = 1 | Q_i = 1, C_i, Z_i) \Pr(Q_i = 1 | C_i, Z_i)}{\sum_a \Pr(Y_i = 1 | Q_i = a, C_i, Z_i) \Pr(Q_i = a | C_i, Z_i)} = \frac{q(1 - S_{i,C_i})}{q(1 - S_{i,C_i}) + (1-q)c},$$

$$\begin{aligned}
\widehat{N_{i,j} Q_i} &= \Pr(N_{i,j} = 1 | Q_i = 1, Y_i = 1, C_i, Z_i) \Pr(Q_i = 1 | Y_i = 1, C_i, Z_i) \\
&= \Pr(N_{i,j} = 1 | \Delta_i = 1, C_i, Z_i) \Pr(Q_i = 1 | Y_i = 1, C_i, Z_i) \\
&= \left[\frac{G_{i,j} S_{i,j}}{(1 - S_{i,C_i})} I(j \le C_i) \right] \left[\frac{q(1 - S_{i,C_i})}{q(1 - S_{i,C_i}) + (1-q)c} \right] \\
&= \frac{q G_{i,j} S_{i,j}}{q(1 - S_{i,C_i}) + (1-q)c} I(j \le C_i).
\end{aligned}$$

If $Y_i = 0$, then

$$\widehat{Q_i} = \frac{q S_{i,C_i}}{q S_{i,C_i} + (1-q)(1-c)},$$

and $\widehat{N_{i,j} Q_i} = 0$. To maximize (3), we set the first derivatives of (3) relative to β and γ to 0 to obtain the following estimating equations for β and γ, respectively.

$$\sum_{i: Y_i = 1} Z_i \left\{ \sum_{j: j \le C_i} \widehat{N_{i,k} Q_i} W_{i,j} - \sum_{j: j \le C_i} \widehat{N_{i,j} Q_i} \sum_{l \le j} w_{i,l} \right\} - \sum_{i: Y_i = 0} Z_i \widehat{Q_i} \sum_{j: j \le C_i} w_{i,j} = 0, \quad (4)$$

$$\sum_{i:Y_i=1}\left\{\widehat{N_{i,k}Q_i}W_{i,k}-\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}I(k\le j)w_{i,k}\right\} \\ -\sum_{i:Y_i=0}\widehat{Q}_iI(k\le C_i)w_{i,k}=0, \tag{5}$$

for $k=1,\ldots,J$, where $W_{i,k}=(\dot{G}/G)(\gamma_k+\beta'Z_i), w_{i,k}=(\dot{G}/(1+G))(\gamma_k+\beta'Z_i)$, and $\dot{G}(x)=dG(x)/dx$. To update β, we apply the one-step Newton-Raphson algorithm based on (4). To update $\gamma_k, k=1,\ldots,J$, we use the one-step self-consistency algorithm based on (5), that is,

$$\gamma_k=\log\left[e^{\gamma_k}\left\{\sum_{i:Y_i=1}\widehat{N_{i,k}Q_i}W_{i,k}\right\}\left\{\sum_{i:Y_i=1}\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}I(k\le j)w_{i,k} \\ +\sum_{i:Y_i=0}\widehat{Q}_iI(k\le C_i)w_{i,k}\right\}^{-1}\right],$$

The initial value of β is set to 0 and the initial value of $\gamma_k, k=1,\ldots,J$, is set to $\log(1/J)$. The expectation steps (E-steps) and M-steps are iterated until the change in all parameter estimates between two successive iterations is less than 10^{-4}.

The distribution of the estimator $\widehat{\theta}=(\widehat{\beta}',\widehat{\gamma}')'$ can be approximated by a multivariate normal distribution in large samples; this can be straightforwardly demonstrated using standard maximum likelihood theory or Z-estimation theory (see, e.g., A.6 of Carroll et al. (2006)). Moreover, the asymptotic variance-covariance matrix of the estimators can be estimated by the sandwich estimator as described in the "Appendix".

4 Smooth baseline hazard

If the number of discrete time points is large, then the number of the parameters in θ is also large. In this case, the parameter estimation in the previous section may become unstable as discussed in Sect. 1.

According to Berger and Schmid (2018), one approach for handling this difficulty is to approximate the baseline coefficients in the model (1) by using a smooth function with fewer parameters as follows. We expand the baseline coefficients γ_t to a smooth function in time $\gamma(t)$, pardon the abuse of notation. A common method of approximating a smooth function is to use a weighted sum of *m* basis functions. Let $\gamma(1),\ldots,\gamma(J)$ be approximated by

$$\gamma_t=\alpha'\psi(t), \tag{6}$$

where ψ is a set of m basis functions. The general choice of ψ can be based on a polynomial or spline basis. Specifically, we consider a Bernstein polynomial model (Lorentz 1986) for (6), where $\alpha = (\alpha_0, \dots, \alpha_K)'$, $\psi = (\psi_0, \dots, \psi_K)'$ with $\psi_k(t) = C_k^K (t/J)^k (1 - t/J)^{K-k}$ and K is the order of the Bernstein polynomial. The number of basis functions $m(= K + 1)$ can typically be chosen to be much smaller than the number of discrete time points J without substantially reducing the accuracy of fit.

Let $\theta_s = (\beta', \alpha')'$ and $\widetilde{\theta_s} = (\widetilde{\beta}', \widetilde{\alpha}')'$ denote the maximum likelihood estimate for θ_s. To compute $\widetilde{\theta_s}$, we employ the EM method. The computation algorithm is identical to that in Sect. 3 except for the method of updating α in the M-step. Reparameterizing the likelihood function in (2) by the parameter θ_s and differentiating the corresponding log complete-data likelihood relative to β, we obtain an estimating equation for β with the same form as that in (4); thus, the Newton–Raphson updating formula for β is the same as that in Sect. 3. Differentiating the log complete-data likelihood relative to α, we obtain the following estimating equation for α

$$\begin{aligned}\sum_{i:Y_i=1}\left\{\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}W_{i,j}\psi_k(j)-\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}\sum_{l\le j}w_{i,l}\psi_k(l)\right\}\\ -\sum_{i:Y_i=0}\left\{\widehat{Q_i}\sum_{j\le C_i}w_{i,j}\psi_k(j)\right\}=0,\quad k=0,\dots,K.\end{aligned} \tag{7}$$

The self-consistency updating formula for α derived from (7) takes the following form.

$$\begin{aligned}\alpha_k=&\log\left[e^{\alpha_k}\left\{\sum_{i:Y_i=1}\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}W_{i,j}\psi_k(j)\right\}\left\{\sum_{i:Y_i=1}\sum_{j:j\le C_i}\widehat{N_{i,j}Q_i}\sum_{l\le j}w_{i,l}\psi_k(l)\right.\right.\\ &\left.\left.+\sum_{i:Y_i=0}\widehat{Q_i}\sum_{j\le C_i}w_{i,j}\psi_k(j)\right\}^{-1}\right],\quad k=0,\dots,K.\end{aligned}$$

Given an initial value for θ_s, the estimator $\widetilde{\theta_s}$ can be obtained by iterating the E-steps and M-steps of the EM method using these updating formulas until convergence is reached. In addition, the asymptotic normality of the estimator $\widetilde{\theta_s}$ can be established, with a sandwich estimator for the asymptotic variance given in the "Appendix".

5 Simulation studies

In this section, we describe simulation studies for assessing the performance of the proposed estimation methods and examining the adequacy of the associated normal approximations. The final simulation study is conducted to illustrate the necessity of establishing the proposed discrete time-to-event method. Specifically, we demonstrate that applying the existing continuous-time method of Wen and Chen (2025), without appropriately addressing with tied or grouped observations, to discrete-time RRT data can lead to substantial bias. Each study comprises 1,000 simulation replications. The sample size $n = 800$ for each simulated data.

In the first simulation study, we consider the covariate vector $Z_i = (Z_{1i}, Z_{2i})$, where Z_{1i} is generated from Bernoulli(0.5) and Z_{2i} is generated from N(0, 1). Given the covariate Z_i, a sensitive event time T_i is generated from model (1) with parameters $\beta = (0.5, -0.5)'$ and $\gamma_j = \log\{0.05((J - j + 1)/J)^{0.5}\}, j = 1, \ldots, J$, $J = 10, 20$, or 40. The transformation $G(x) = \exp(e^x) - 1$ which yields the GPH model. The survey time C_i is simulated from Uniform$\{1, \ldots, J\}$. The answer to the sensitive question is given by $\Delta_i = I(T_i \leq C_i)$, whereas the answer to innocuous question W_i is generated from a Bernoulli(c) and $c = 0.25$ or 0.5. Given Δ_i and W_i, the observed response Y_i of the unrelated-question RRT survey is given by $Y_i = Q_i\Delta_i + (1 - Q_i)W_i$, where Q_i is Bernoulli(q) and $q = 0.5$ or 0.7. The set of observed data is $\{O_i = (Y_i, C_i, Z_i) | i = 1, \ldots, n\}$.

For each simulated dataset, we apply (i) the discrete-baseline-hazard (DBH) method where the baseline coefficients $\gamma_t, t = 1, \ldots, J$, are the parameters themselves and the parameter estimation method is given in Sect. 3; (ii) the smooth-baseline-hazard (SBH) method where the baseline coefficients are approximated by the smooth Bernstein polynomials $\gamma(t)$ in (6) and the parameter estimation method is given in Sect. 3. The general suggestion of the number of basis functions to ensure numerically stable estimates exist is 4 or 5. (see, e.g., Tutz and Schimid 2016, p. 108). We set the number of Bernstein basis functions to 5 for convenience. A sensitivity analysis to examine the effect of different numbers of basis functions on the inference and the determination of the number of basis functions in practice is discussed in the subsequent section. For comparison, we also consider an analysis of the full data, namely the data obtained when all respondents are surveyed using sensitive questions (i.e., $q = 1$) such that the full data $\{(\Delta_i, C_i, Z_i) | i = 1, \ldots, n\}$ are available. In the second simulation study, data are simulated with the same parameters except that $G(x) = \exp(x)$, which corresponds to the PCR model.

The results of the first and second simulation studies are presented in Tables 1 and 2, respectively, where "Bias" refers to the simulation bias of the estimates; "SD" the simulation standard deviation of the estimates; "ASE" the average of the standard error estimates over the simulations; and "CP" the coverage probability of the 95% Wald-type confidence intervals obtained by asymptotic normality. The tables also include the results of the relative efficiency (RE) in comparisons with the full-data analysis, which is defined as MSE($\check{\beta}_f$)/MSE($\check{\beta}$), where MSE is the mean squared error over the simulation replicates and $\check{\beta}_f$ is the estimate of the full-data analysis, that is, the estimate from (2) with $q = 1$.

As we can see from Tables 1 and 2, except in some simulation scenarios with a larger number of discrete time points ($J = 40$ in Table 2), the proposed DBH estimator $\widehat{\beta}$ and SBH estimator $\widetilde{\beta}$ perform quite well: the estimation bias is close to zero, the average of the standard error estimates is close to the simulation standard deviation, and the coverage probability of the 95% confidence intervals based on the asymptotic

Table 1 Simulation results for the GPH model

J	q	c		DBH analysis ($\widehat{\beta}$)					SBH analysis ($\widetilde{\beta}$)				
				Bias	SD	ASE	CP	RE	Bias	SD	ASE	CP	RE
10	0.5	0.25	β_1	0.033	0.309	0.305	96.1	22.0	0.028	0.307	0.303	96.1	21.9
			β_2	− 0.029	0.176	0.166	95.6	17.8	− 0.023	0.171	0.164	95.4	18.1
		0.5	β_1	0.044	0.360	0.332	94.9	16.2	0.037	0.355	0.331	95.0	16.4
			β_2	− 0.031	0.201	0.180	94.1	13.5	− 0.023	0.194	0.178	93.9	14.2
	0.7	0.25	β_1	0.016	0.221	0.213	95.4	43.1	0.013	0.219	0.212	95.6	43.1
			β_2	− 0.012	0.116	0.113	95.6	40.8	− 0.009	0.115	0.112	95.3	40.9
		0.5	β_1	0.016	0.240	0.225	94.4	36.4	0.013	0.238	0.224	94.5	36.8
			β_2	− 0.013	0.129	0.119	93.9	33.1	− 0.009	0.126	0.119	94.4	33.7
	1	–	β_1	0.006	0.145	0.144	94.5	100.0	0.004	0.144	0.143	94.5	100.0
			β_2	− 0.006	0.074	0.073	94.5	100.0	− 0.004	0.074	0.073	94.8	100.0
20	0.5	0.25	β_1	0.030	0.268	0.257	95.1	19.0	0.015	0.257	0.250	94.6	20.2
			β_2	− 0.034	0.156	0.147	96.1	15.7	− 0.019	0.146	0.142	95.5	17.1
		0.5	β_1	0.040	0.286	0.268	94.9	16.7	0.024	0.275	0.261	95.4	17.6
			β_2	− 0.037	0.172	0.155	93.3	12.8	− 0.021	0.162	0.150	93.7	14.0
	0.7	0.25	β_1	0.015	0.183	0.179	95.5	40.6	0.008	0.180	0.176	95.6	41.4
			β_2	− 0.017	0.100	0.100	95.8	38.1	− 0.009	0.098	0.098	95.5	38.8
		0.5	β_1	0.015	0.189	0.182	94.9	38.2	0.007	0.185	0.179	94.6	39.0
			β_2	− 0.017	0.106	0.102	94.4	34.1	− 0.009	0.102	0.100	95.3	35.2
	1	–	β_1	0.007	0.117	0.118	95.7	100.0	0.004	0.116	0.117	95.2	100.0
			β_2	− 0.009	0.062	0.062	95.9	100.0	− 0.005	0.061	0.061	95.7	100.0
40	0.5	0.25	β_1	0.047	0.283	0.284	92.9	15.0	0.013	0.256	0.246	93.6	17.6
			β_2	− 0.057	0.180	0.165	92.6	11.6	− 0.018	0.153	0.145	93.6	15.2
		0.5	β_1	0.060	0.298	0.276	92.2	13.6	0.024	0.274	0.251	93.0	15.3
			β_2	− 0.064	0.199	0.164	90.5	9.5	− 0.025	0.169	0.149	92.0	12.3
	0.7	0.25	β_1	0.023	0.191	0.185	94.8	33.0	0.005	0.182	0.174	94.7	34.8
			β_2	− 0.024	0.109	0.107	95.7	32.0	− 0.006	0.102	0.101	95.8	34.8

Table 1 (continued)

J	q	c		DBH analysis ($\widehat{\beta}$)					SBH analysis ($\widetilde{\beta}$)				
				Bias	SD	ASE	CP	RE	Bias	SD	ASE	CP	RE
		0.5	β_1	0.022	0.190	0.182	94.2	33.3	0.005	0.181	0.171	93.8	35.3
			β_2	− 0.025	0.112	0.107	95.1	30.0	− 0.008	0.105	0.100	94.0	32.6
	1	–	β_1	0.011	0.110	0.108	93.3	100.0	0.002	0.108	0.105	93.5	100.0
			β_2	− 0.014	0.061	0.059	94.0	100.0	− 0.005	0.060	0.057	93.8	100.0

$\widehat{\beta}$, estimator assuming DBH; $\widetilde{\beta}$, estimator assuming SBH; J, number of discrete time points; q, probability of selecting the sensitive question; c, proportion answering "yes" to the innocuous question

ASE: average of standard error estimates; CP(%): coverage probability of the 95% confidence interval; RE(%): ratio of the MSEs of the full-data estimate to that of the considered estimate

Table 2 Simulation results for the PCR model

J	q	c		DBH analysis ($\widehat{\beta}$)					SBH analysis ($\widetilde{\beta}$)				
				Bias	SD	ASE	CP	RE	Bias	SD	ASE	CP	RE
10	0.5	0.25	β_1	0.064	0.345	0.341	96.4	20.2	0.037	0.328	0.322	96.3	21.7
			β_2	− 0.058	0.196	0.192	97.2	16.8	− 0.029	0.182	0.175	96.0	17.8
		0.5	β_1	0.081	0.411	0.378	95.4	16.2	0.047	0.380	0.354	95.2	16.2
			β_2	− 0.065	0.228	0.214	96.0	13.6	− 0.031	0.208	0.191	94.6	13.6
	0.7	0.25	β_1	0.037	0.241	0.233	95.4	43.6	0.018	0.232	0.223	95.1	43.7
			β_2	− 0.032	0.128	0.128	96.2	39.3	− 0.012	0.122	0.119	95.1	40.2
		0.5	β_1	0.039	0.265	0.249	94.4	38.2	0.018	0.252	0.238	93.9	37.1
			β_2	− 0.035	0.143	0.137	95.3	34.0	− 0.013	0.135	0.126	94.4	32.8
	1	–	β_1	0.017	0.157	0.155	93.9	100.0	0.006	0.154	0.151	94.4	100.0
			β_2	− 0.016	0.080	0.081	95.4	100.0	− 0.004	0.078	0.077	95.6	100.0
20	0.5	0.25	β_1	0.074	0.301	0.300	96.5	18.9	0.019	0.266	0.260	95.1	20.9
			β_2	− 0.079	0.179	0.185	97.6	15.9	− 0.022	0.153	0.148	96.0	17.3
		0.5	β_1	0.092	0.327	0.319	96.5	16.4	0.030	0.288	0.273	95.5	17.8
			β_2	− 0.088	0.199	0.199	96.7	13.5	− 0.025	0.169	0.156	93.6	14.2
	0.7	0.25	β_1	0.046	0.202	0.203	95.7	41.8	0.009	0.187	0.183	95.2	42.5
			β_2	− 0.049	0.111	0.120	95.9	38.0	− 0.011	0.102	0.102	95.5	39.6
		0.5	β_1	0.049	0.210	0.208	95.6	39.4	0.010	0.192	0.186	94.9	40.1
			β_2	− 0.052	0.119	0.125	96.5	34.9	− 0.011	0.107	0.104	95.2	35.5
	1	–	β_1	0.029	0.129	0.131	95.8	100.0	0.005	0.122	0.122	96.0	100.0
			β_2	− 0.032	0.068	0.073	96.4	100.0	− 0.007	0.064	0.064	95.6	100.0
40	0.5	0.25	β_1	0.112	0.323	0.346	98.0	14.2	0.016	0.262	0.257	95.6	17.4
			β_2	− 0.122	0.203	0.229	98.9	13.7	− 0.023	0.160	0.151	95.4	14.7
		0.5	β_1	0.135	0.349	0.356	98.1	13.0	0.029	0.281	0.262	94.8	15.0
			β_2	− 0.138	0.235	0.243	98.8	11.8	− 0.028	0.174	0.155	93.5	12.3
	0.7	0.25	β_1	0.071	0.213	0.221	97.2	35.0	0.006	0.185	0.178	95.0	35.2
			β_2	− 0.073	0.121	0.143	98.4	35.9	− 0.008	0.104	0.103	95.3	34.9

Table 2 (continued)

J	q	c		DBH analysis ($\widehat{\beta}$)					SBH analysis ($\widetilde{\beta}$)				
				Bias	SD	ASE	CP	RE	Bias	SD	ASE	CP	RE
		0.5	β_1	0.074	0.215	0.221	97.3	34.9	0.005	0.184	0.175	94.5	35.4
			β_2	− 0.078	0.125	0.145	98.0	33.9	− 0.009	0.107	0.102	94.1	33.2
	1	–	β_1	0.048	0.120	0.129	95.0	100.0	0.003	0.110	0.108	94.0	100.0
			β_2	− 0.052	0.069	0.081	96.3	100.0	− 0.006	0.061	0.059	94.6	100.0

$\widehat{\beta}$, estimator assuming DBH; $\widetilde{\beta}$, estimator assuming SBH; J, number of discrete time points; q, probability of selecting the sensitive question; c, proportion answering "yes" to the innocuous question

ASE: average of standard error estimates; CP(%): coverage probability of the 95% confidence interval; RE(%): ratio of the MSEs of the full-data estimate to that of the considered estimate normality is close to the nominal level. For larger numbers of discrete time points ($J = 40$ in Table 2), the DBH estimator $\widehat{\beta}$ performs less satisfactorily due to unstable estimation caused by the high dimension of the parameters. However, this is not the case for the SBH method, where the baseline coefficients are specified by a smooth function with far fewer parameters, thereby reducing the parameter dimensionality. All simulation results indicate that the SBH analysis provides more stable and efficient estimates than the DBH analysis. Moreover, the RE of the two estimators increase as the number of discrete time points J decreases; the probability of choosing the sensitive question q increases, and the probability of answering "yes" to the innocuous question c decreases.

The adequacy of the normal approximation theory for the proposed estimators can be further verified through the nearly linear patterns of Q–Q plots in Fig. 1, this figure illustrates the standardized estimates for β versus the standard normal quantile

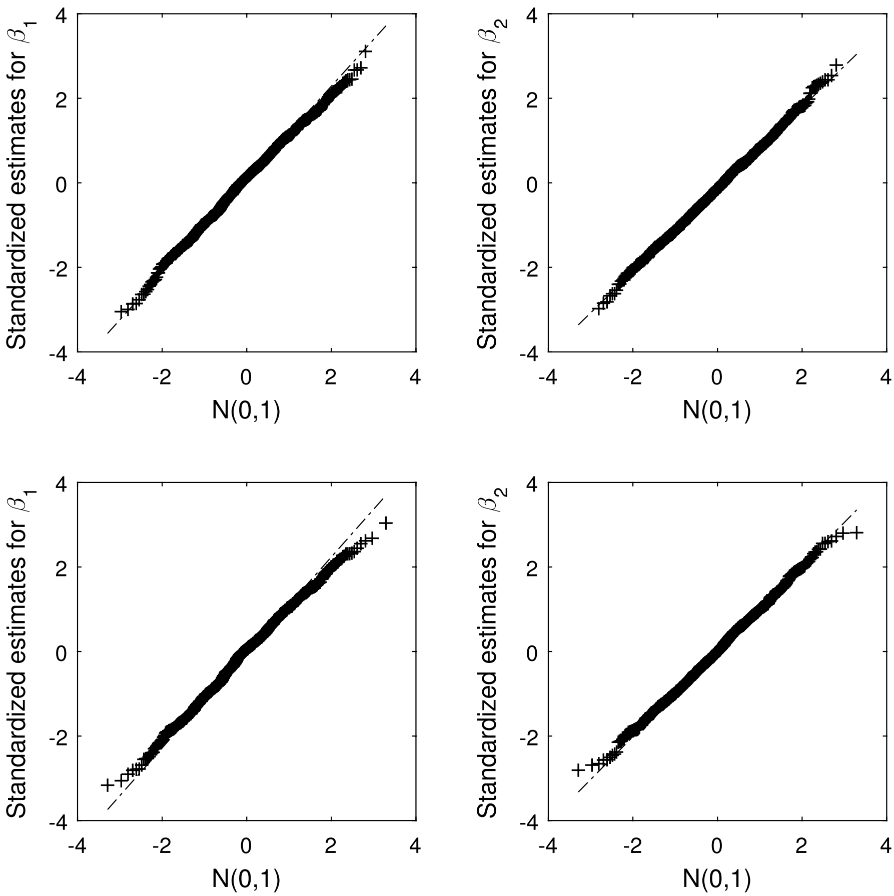

Fig. 1 Q–Q plots of standardized estimates versus the standard normal distribution for the simulation with $J = 40, q = 0.7$, and $c = 0.25$ for the GPH fitted model with discrete (upper panels) or smooth (lower panels) baseline hazards assumed

values, under the simulation scenario with $J = 40, q = 0.7, c = 0.25$ for the GPH model with discrete and smooth baseline hazards assumed.

To compare the existing continuous-time method for analyzing discrete-time RRT data, we conducted simulations with the same settings as those for the first scenario of Table 1; however, the sensitive event time T_i was simulated by grouping the continuous data from the Cox model with the cumulative hazard $e^{Z_{1i}-Z_{2i}}0.07t^{0.4}$ and the survey time C_i is simulated by grouping the continuous data from uniform(0,40) in the following manner. The continuous time was divided into $J+1$ time intervals [0, 40/*J*), $[40/J, 2\{40/J\}), \ldots, [J\{40/J\}, \infty)$, and T_i or $C_i = t$ if its original continuous time lies in the *t*-th interval, with $t = 1, \ldots, J+1$. Specifically, we analyzed the grouped RRT data (discrete-time data) from this setting using the proposed methods and the method of Wen and Chen (2025). The method of Wen and Chen (2025) was based on the Cox model aimed for RRT data measured on a continuous-time scale. Table 3 reveals that the proposed DBH and SBH estimators, using a GPH-fitted model, perform favorably. The result is expected because the GPH model, rather than the Cox model, is the true (correct) model when continuous time-to-event data generated from a Cox model are grouped (see, e.g., Tutz and Schimid 2016). On the other hand, substantial bias may arise from the direct application of a continuous-time method to discrete-time RRT data, especially when there is no suitable handling of grouped or tied observations and when the number of time intervals (i.e., the number

Table 3 Results of the proposed discrete-time methods (DBH or SBH) and the continuous-time method (Naive) for analyzing discrete-time RRT data

J	Method		$n = 800$				$n = 1200$			
			Bias	SD	ASE	CP	Bias	SD	ASE	CP
10	Naive	β_1	− 0.252	0.222	0.291	91.8	− 0.244	0.173	0.237	88.6
		β_2	0.266	0.109	0.172	68.8	0.268	0.086	0.140	51.0
	DBH	β_1	0.041	0.337	0.314	95.3	0.027	0.253	0.252	96.3
		β_2	− 0.070	0.243	0.217	94.8	− 0.042	0.180	0.172	95.7
	SBH	β_1	0.028	0.326	0.310	95.6	0.020	0.248	0.249	96.0
		β_2	− 0.056	0.235	0.214	95.1	− 0.034	0.175	0.170	95.9
20	Naive	β_1	− 0.182	0.243	0.295	94.0	− 0.175	0.191	0.240	92.8
		β_2	0.188	0.129	0.182	85.3	0.192	0.101	0.147	78.2
	DBH	β_1	0.054	0.346	0.320	95.4	0.039	0.261	0.257	95.9
		β_2	− 0.085	0.255	0.223	94.8	− 0.052	0.184	0.176	96.2
	SBH	β_1	0.031	0.331	0.312	95.4	0.025	0.253	0.252	96.0
		β_2	− 0.059	0.238	0.215	94.7	− 0.037	0.176	0.172	96.1
40	Naive	β_1	− 0.124	0.263	0.300	95.7	− 0.119	0.206	0.243	95.2
		β_2	0.122	0.149	0.191	90.4	0.128	0.117	0.154	88.8
	DBH	β_1	0.066	0.357	0.334	93.6	0.048	0.268	0.269	96.2
		β_2	− 0.099	0.264	0.232	94.8	− 0.062	0.189	0.181	95.7
	SBH	β_1	0.032	0.334	0.313	95.2	0.027	0.255	0.253	96.1
		β_2	− 0.060	0.239	0.215	95.2	− 0.039	0.178	0.172	95.5

The discrete-time methods are based on the GPH model, while the continuous-time method is based on the Cox model. The simulated data are obtained by grouping continuous event time data from the Cox model into $J+1$ mutually exclusive time intervals

ASE: average of standard error estimates; CP(%): coverage probability of the 95% confidence interval of discrete time points) J is small. The performance of the naive method does not improve even with an increase in sample size. This simulation study highlights the necessity of developing the proposed discrete time-to-event method.

To complement the main simulations, the "Appendix" reports additional investigations on computational time, the impact of small sample sizes and high missing rates, the effects of misspecified design parameters and working models, and scenarios involving noncompliance, skewed censoring.

6 Real data applications

We consider two real data applications. The first is the "extramarital sex" dataset from the RRT survey conducted in Taiwan to study the prevalence of extramarital sex among married individuals (Hsieh et al. 2016) and the second is the "militant connection" dataset from the RRT survey conducted in Nigeria to study the prevalence of civilian connections with militant groups (Blair et al. 2015). To illustrate the application of the proposed method, we re-analyzed these two real studies by converting prevalence estimation into time-to-event estimation.

6.1 Extramarital sex data

We applied the proposed method to a dataset from the 2012 Taiwan Social Change Survey (TSCS), consisting of $n = 805$ married Taiwan residents aged 21 or above. The outcome of interest was the time to extramarital sex since marriage. The TSCS study surveyed the incidence of extramarital sex on participants by using the unrelated-question RRT (Greenberg et al. 1969), wherein the sensitive and innocuous questions were as follows:

A: *Have you ever had sex with someone other than your spouse?*
B: *Were you born in January, February, or March?*

The question that the respondent was asked to answer was determined through the following prompt. Please pick one card from the deck of playing cards and do not tell the interviewer the number on the playing card. Then, please answer Question A or B according to the number on the playing card. If the number on the playing card is 1, 2, or 3, please answer Question A; if the number on the playing card is 4 or 5, please answer Question B.

Eight cards were numbered 1, four were numbered 2, and eight were numbered 3; moreover, sixteen were numbered 4, and four were numbered 5 (20 total); hence the probability of answering the sensitive question A was $\Pr(Q = 1) = q = 0.5$. The goal is to examine the relationship between a set of covariates Z_i and the time to extramarital sex T_i. Instead of observing T_i, we can only observe $Y_i = Q_i\Delta_i + (1 - Q_i)W_i$, where Δ_i and W_i are the responses to questions A and B, respectively. The response to the sensitive question $\Delta_i = I(T_i \leq C_i)$ can be treated as the current status indicator for T_i at the survey time C_i, indicating whether the sensitive event occurred before C_i. In the TSCS, the survey time C_i for the sensitive event (extramarital sex) is the answer to the question

"How many years have you been married?".

Due to this question design, the time points used for the examination of extramarital sex were discrete because they were measured in years. In the considered data, the maximum time on examining extramarital sex is 40. Hence, the study had 41 possible sensitive event times $t = 1, \ldots, 41$.

The covariates considered in the following analysis were "gender (Male, 1=Male)," "attitudes toward extramarital sex (Atti, 1=Yes)," "number of children (NChild, 0–5 children)," "years of education (EduYear, 1.5–27 years)," "age at marriage (MarryAge, 17–54 years)", and " income (Income, 1=Yes)." To identify the factors associated with extramarital sex, we assumed the discrete event time model (1) with $G(x) = \exp(e^x) - 1$ or $\exp(x)$. The transformation $G(x) = \exp(e^x) - 1$ corresponds to a GPH model, while $G(x) = \exp(x)$ corresponds to a PCR model. We then applied the proposed parameter estimations in Sects. 3 and 4 using the aforementioned models under either DBH or SBH specifications. In the SBH method, a Bernstien polynomial basis with various numbers of basis functions is considered for (6).

Table 4 Analysis results for extramarital sex data

Model	Method	*m*		Male	Atti	NChild	EduYear	MarryAge	Income	AIC
GPH	SBH	1	Est	1.718*	1.194*	0.072	− 0.017	0.026	− 0.048	941.697
			SE	0.448	0.414	0.179	0.051	0.038	0.505	
		2	Est	1.514*	0.808*	0.096	− 0.043	− 0.017	− 0.185	931.069
			SE	0.393	0.344	0.159	0.046	0.033	0.450	
		3	Est	1.498*	0.802*	0.103	− 0.044	− 0.016	− 0.187	932.627
			SE	0.391	0.345	0.159	0.046	0.033	0.448	
		4	Est	1.492*	0.799*	0.107	− 0.045	− 0.016	− 0.187	934.450
			SE	0.391	0.345	0.158	0.046	0.034	0.447	
		5	Est	1.488*	0.797*	0.109	− 0.045	− 0.016	− 0.188	936.354
			SE	0.391	0.345	0.159	0.046	0.034	0.447	
	DBH	–	Est	1.528*	0.814*	0.075	− 0.036	− 0.004	− 0.151	1000.434
			SE	0.409	0.361	0.171	0.049	0.037	0.484	
PCR	SBH	1	Est	1.736*	1.222*	0.073	− 0.017	0.027	− 0.048	941.705
			SE	0.452	0.427	0.182	0.052	0.039	0.511	
		2	Est	1.616*	0.920*	0.098	− 0.046	− 0.020	− 0.182	931.147
			SE	0.426	0.407	0.179	0.051	0.037	0.489	
		3	Est	1.628*	0.940*	0.105	− 0.049	− 0.020	− 0.182	932.747
			SE	0.442	0.435	0.185	0.053	0.038	0.497	
		4	Est	1.638*	0.952*	0.108	− 0.051	− 0.020	− 0.182	934.597
			SE	0.455	0.455	0.188	0.054	0.039	0.502	
		5	Est	1.646*	0.960*	0.111	− 0.052	− 0.021	− 0.182	936.521
			SE	0.470	0.471	0.192	0.055	0.040	0.506	
	DBH	–	Est	1.657*	0.969*	0.071	− 0.041	− 0.005	− 0.149	1000.344
			SE	0.486	0.491	0.200	0.057	0.043	0.536	

"$*$", significance; *m*, number of basis functions

Table 4 provides estimates of the regression parameters and standard errors based on asymptotic theory. For the GPH and PCR model, both DBH and SBH analyses indicate that males or individuals with a positive attitude toward extramarital sex were significantly more likely to have extramarital sex. In interpretation, for example, under the PCR model with DBH estimation, the continuation ratio for extramarital affairs among men is estimated to be $\exp(1.657)$ times that among women, while for respondents with permissive attitudes is $\exp(0.969)$ times that of non-permissive respondents. By contrast, under the GPH model, the corresponding log-survival ratios are $\exp(1.528)$ and $\exp(0.814)$, respectively. By contrast, number of children, years of education, the age at marriage, and income were not significantly associated with extramarital sex in these analyses. As expected, the SBH analysis, where the baseline hazard was specified with fewer parameters, demonstrated greater efficiency than the DBH analysis, except for the $m = 1$ case in which the baseline hazard may be approximated poorly.

The results of the proposed SBH analysis exhibited minimal variation with respect to the number of the basis functions *m* used, except for the case $m = 1$, and the results of the two model specifications, $G(x) = \exp(e^x) - 1$ or $\exp(x)$, were similarly consistent. To identify an optimal model among these analysis models, we considered the Akaike Information Criterion (AIC) given by

$$\text{AIC} = -2\widehat{\ell} + 2\{\text{number of parameters}\}, \tag{8}$$

where $\widehat{\ell}$ represents the maximum value of the log-likelihood under the considered model; the model minimizing this criterion is selected. The GPH model with the baseline hazard specified by Bernstein polynomials of order 1 (i.e., the SBH analysis with $m = 2$) produced the lowest AIC value of 931.0269.

Figure 2 shows the proportions of men and women who have had extramarital sex estimated using the GPH and PCR models with DBH or the suggested SBH with (*m*

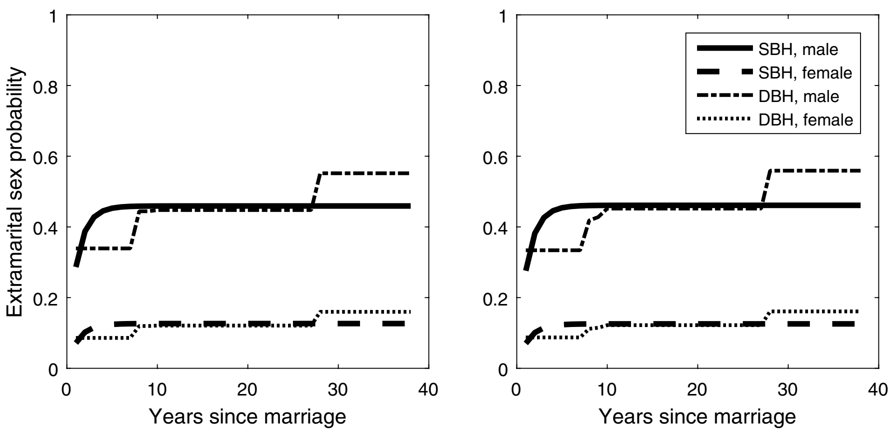

Fig. 2 Proportions of men and women having had extramarital sex estimated using the GPH model (left) and the PCR model (right) with DBH or SBH and other covariates fixed at their sample mean values. The SBH methods have 2 basis functions for the GPH and PCR models

= 2); other covariates (Atti, NChild, EduYear, MarryAge, Income) were fixed at their sample mean values. While the DBH and SBH analyses yielded different results, we posit that the SBH estimates might be more reliable.

6.2 Militant connection data

We applied the proposed method to militant connection data. The dataset consists of $n = 2{,}423$ local civilians all with complete information on the relevant variables. The study aimed to identify the characteristics of civilians associated with cooperation with militant groups. The survey was conducted using a forced-response design (Boruch 1971) with summarized instructions as follows. Please throw the dice and do not disclose the number on your dice throw to the interviewer. If 1 shows on the dice, answer no; if 6 shows, answer yes. But, if 2 or 3 or 4 or 5 shows, please answer the following question,

A: *Whether you hold direct social connections with members of militant groups?*

The aforementioned forced-response design is equivalent to an unrelated-question RRT with the probability of answering the sensitive question A being $q = 2/3$ and the proportion of answering "yes" to the innocuous question being $c = 0.5$. To demonstrate the application of the method, we treated the age of respondents as the examination time of the event (connection with militant groups). In the dataset, the times for examining cooperation events were measured in years (discrete) with a range of 16–99. Hence, the study had possible sensitive event times $t = 1, \ldots, 85$. The covariates considered in the analysis were "gender (Female, 1 = Female)," "number of assets owned by the respondent from a list of nine assets including radio, TV, motorbike, car, mobile phone, refrigerator, goat, chicken, and cow (Asset, 0–9)," "marital status (Married, 1 = Married)," "level of education (Edu, from 1 = no school to 10 = graduate education)," and "member of a civic group in the communities (Civic, 1 = Yes)."

As in Sect. 6.1, we considered the GPH and PCR models with either discrete (DBH) or smooth (SBH) baseline hazards. Table 5 presents the analysis results. According to AIC, the SBH analysis with $m = 3$ for the GPH model and the SBH analysis with $m = 4$ for the PCR model are suggested. These two suggested SBH analyses indicated that males or unmarried civilians were significantly more likely to connect with militant groups. However, the number of assets, level of education, and members of civic groups were not significantly associated with the connection. In contrast, for the DBH analysis, both GPH and PCR models indicated that males were significantly more likely to connect with militant groups, while other covariate effects were non-significant. As expected, the SBH analysis, which used fewer parameters to specify the baseline hazard, was more efficient than the DBH analysis.

Figure 3 illustrates the proportions of men and women who connected with militant groups estimated using the the GPH and PCR models with DBH or the suggested SBH; other covariates were fixed at the sample average values.

Table 5 Analysis results for militant connection data

Model	Method	*m*		Female	Asset	Marriage	Edu	Civic	AIC
GPH	SBH	1	Est	− 0.448*	0.059	− 1.464*	− 0.054	0.348*	3183.031
			SE	0.155	0.038	0.150	0.040	0.153	
		2	Est	− 0.545*	0.056	− 0.429*	− 0.039	0.199	3101.398
			SE	0.142	0.035	0.152	0.038	0.136	
		3	Est	− 0.543*	0.057	− 0.395*	− 0.036	0.212	3098.582
			SE	0.143	0.035	0.162	0.038	0.136	
		4	Est	− 0.541*	0.058	− 0.381*	− 0.035	0.219	3098.663
			SE	0.142	0.034	0.172	0.038	0.136	
		5	Est	− 0.540*	0.059	− 0.372*	− 0.035	0.223	3099.628
			SE	0.142	0.034	0.185	0.038	0.136	
	DBH	-	Est	− 0.496*	0.056	− 0.457	− 0.029	0.281	3245.907
			SE	0.160	0.041	0.375	0.052	0.148	
PCR	SBH	1	Est	− 0.453*	0.061	− 1.479*	− 0.054	0.357*	3182.864
			SE	0.157	0.039	0.151	0.040	0.155	
		2	Est	− 0.564*	0.059	− 0.446*	− 0.040	0.211	3101.525
			SE	0.146	0.036	0.155	0.040	0.142	
		3	Est	− 0.567*	0.062	− 0.416*	− 0.038	0.226	3098.541
			SE	0.148	0.036	0.166	0.040	0.143	
		4	Est	− 0.569*	0.063	− 0.405*	− 0.038	0.234	3098.534
			SE	0.149	0.037	0.176	0.040	0.144	
		5	Est	− 0.569*	0.064	− 0.398*	− 0.037	0.240	3099.446
			SE	0.149	0.037	0.189	0.041	0.145	
	DBH	-	Est	− 0.538*	0.063	− 0.522	− 0.032	0.312*	3246.388
			SE	0.160	0.041	0.375	0.052	0.148	

"$*$", significance; *m*, number of basis functions

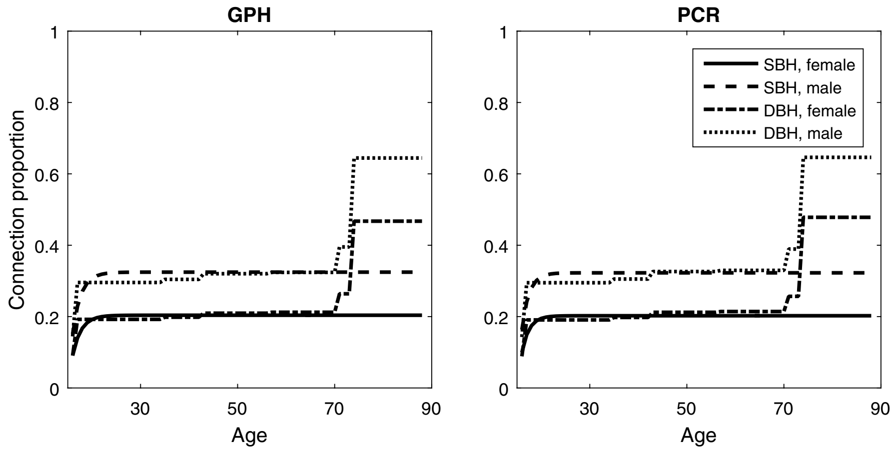

Fig. 3 Proportions of men and women connecting with militant groups estimated using the GPH (left) and the PCR (right) models with DBH or SBH and other covariates fixed at the sample average values. The SBHs have 3 and 4 basis functions for the GPH and PCR models, respectively

7 Concluding remarks

Current state event time data are commonly encountered in survey research; furthermore, when the events of interest involve sensitive issues, they are often collected via the Randomized Response Technique (RRT). In this study, we propose a general methodology for analyzing current status event time data collected by the unrelated-question RRT under the discrete-time framework. The event-time hazard rate of the event is modeled using a general discrete-time transformation model, including the proportional continuation ratio and grouped proportional hazards model; the baseline hazard is specified either through discrete intercepts at each time point or by smooth functions with reduced parameters. In general, the smooth baseline hazard based (SBH) analysis achieves more stable estimation and more efficient inference than the discrete baseline hazard based (DBH) analysis once the baseline hazard is well-approximated. When the number of discrete time points is large, SBH should be preferred, as it generally yields more stable and efficient estimation under such settings. As no formal model selection criterion exists for RRT survival analysis, we relied on AIC, with supplementary comparisons in the "Appendix". Formal model selection methods warrant further investigation in future work.

Many RRTs have been used to increase the validity and reliability of inferences for surveys on sensitive topics by enhancing privacy protection and reducing untruthful responses. Throughout this paper, we assume that event time data are collected by the unrelated-question RRT of Greenberg et al. (1969), a popular and efficient RRT, and that the respondents answer truthfully. However, this RRT may produce random errors or respondents may still be untruthful. Discrete time-to-event analysis of current status data collected using more efficient and convincing RRT procedures, such as a related- and unrelated-question combined RRT (Hsieh et al. 2016) or a two-stage RRT (Narjis and Shabbir 2023), could be explored in future research.

Appendix

Variance estimation

Variance estimation for the DBH analysis

Recall in Sect. 3 that $G_{i,j} = G(\gamma_j + \beta' Z_i)$, $S_{i,t} = \prod_{j \leq t} \{1 + G_{i,j}\}^{-1}$, and $w_{i,j} = (\dot{G}/(1+G))(\gamma_j + \beta' Z_i)$. The log-likelihood function parameterized by $\theta = (\beta', \gamma')'$ of individual i takes the form $\ell_i(\theta) = \{\mathcal{E}_i\}^{Y_i}\{1 - \mathcal{E}_i\}^{1-Y_i}$, where $\mathcal{E}_i = q(1 - S_{i,C_i}) + (1-q)c$. The scores for β and γ, defined by $\partial \ell_i(\theta)/\partial \beta$ and $\partial \ell_i(\theta)/\partial \gamma$, take the forms

$$m_{\beta,i} = Z_i' \frac{pS_{i,C_i} \sum_{j \le C_i} w_{i,j}}{\mathcal{E}_i(1-\mathcal{E}_i)}(Y_i - \mathcal{E}_i),$$

$$m_{\gamma_k,i} = \frac{pS_{i,C_i} I(k \le C_i) w_{i,k}}{\mathcal{E}_i(1-\mathcal{E}_i)}(Y_i - \mathcal{E}_i), \quad k = 1, \dots, J$$

respectively. Let $m_{\theta,i} = (m_{\beta,i}', m_{\gamma,i}')'$ with $m_{\gamma,i} = (m_{\gamma_1,i}, \dots, m_{\gamma_J,i})'$. The asymptotic variance matrix $\mathrm{var}(\widehat{\theta})$ of the MLE of θ can be estimated by the sandwich estimator $n^{-1}A^{-1}B(A^{-1})'$, where

$$B = \frac{1}{n}\sum_{i=1}^{n} m_{\theta,i} m_{\theta,i}', \quad A = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta} m_{\theta,i}\right\},$$

with both evaluated at $\theta = \widehat{\theta}$. We partition the matrix $n^{-1}A^{-1}B(A^{-1})'$ in accordance with the dimensions of β and γ; $\mathrm{var}(\widehat{\beta})$ and $\mathrm{var}(\widehat{\gamma})$ can then be estimated by the upper–left and lower–right submatrixes of $n^{-1}A^{-1}B(A^{-1})'$, respectively. The explicit forms for $\frac{\partial}{\partial\theta} m_{\theta,i}$ are presented below.

$$\frac{\partial}{\partial\beta} m_{\beta,i} = -Z_i' Z_i \frac{\left\{pS_{i,C_i} \sum_{j \le C_i} w_{i,j}\right\}^2}{\mathcal{E}_i(1-\mathcal{E}_i)},$$

$$\frac{\partial}{\partial\gamma_k} m_{\beta,i} = -Z_i' \frac{\{pS_{i,C_i}\}^2 \{\sum_{j \le C_i} w_{i,j}\}\{I(k \le C_i) w_{i,k}\}}{\mathcal{E}_i(1-\mathcal{E}_i)} = \left(\frac{\partial}{\partial\beta} m_{\gamma_k,i}\right)', \quad k = 1, \dots, J,$$

$$\frac{\partial}{\partial\gamma_r} m_{\gamma_k,i} = -\frac{\{pS_{i,C_i}\}^2 \{I(k \le C_i) w_{i,k}\}\{I(r \le C_i) w_{i,r}\}}{\mathcal{E}_i(1-\mathcal{E}_i)}, \quad k, r = 1, \dots, J.$$

Variance estimation for the SBH analysis

Recall in Sect. 4 that $\gamma_t = \sum_{k=0}^{K} \alpha_k \psi_k(t)$ with $\psi_k(t) = C_k^K (t/J)^k (1-t/J)^{K-k}$. Use $\theta_s = (\beta', \alpha')'$ to reparameterize the log-likelihood function $\ell_i(\theta)$ in "Appendix A.1" as $\ell_i(\theta_s)$. The scores for β and α, defined by $\partial\ell_i(\theta_s)/\partial\beta$ and $\partial\ell_i(\theta_s)/\partial\alpha$, then take the forms

$$m_{\beta,i} = Z_i' \frac{pS_{i,C_i} \sum_{j \le C_i} w_{i,j}}{\mathcal{E}_i(1-\mathcal{E}_i)}(Y_i - \mathcal{E}_i),$$

$$m_{\alpha_k,i} = \frac{pS_{i,C_i} \sum_{j \le C_i} w_{i,j} \psi_k(j)}{\mathcal{E}_i(1-\mathcal{E}_i)}(Y_i - \mathcal{E}_i), \quad k = 0, \dots, K$$

respectively. Let $m_{\theta_s,i} = (m'_{\beta,i}, m'_{\alpha,i})'$ with $m_{\alpha,i} = (m_{\alpha_0,i}, \ldots, m_{\alpha_K,i})'$. The asymptotic variance matrix $\text{var}(\widetilde{\theta})$ of the MLE $\widetilde{\theta}$ of θ_s can be estimated by the sandwich estimator $n^{-1}A_s^{-1}B_s(A_s^{-1})'$, where

$$B_s = \frac{1}{n}\sum_{i=1}^{n} m_{\theta_s,i}m'_{\theta_s,i}, A_s = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta_s}m_{\theta_s,i},$$

with both evaluated at $\theta_s = \widetilde{\theta}(= (\widetilde{\beta}', \widetilde{\alpha}')')$. Likewise, partition the matrix $n^{-1}A_s^{-1}B_s(A_s^{-1})'$ in accordance with the dimensions of β and α; $\text{var}(\widetilde{\beta})$ and $\text{var}(\widetilde{\alpha})$ can be estimated by the upper–left and lower–right submatrixes of $n^{-1}A_s^{-1}B_s(A_s^{-1})'$, respectively. The explicit forms for $\frac{\partial}{\partial\theta_s}m_{\theta_s,i}$ are as follows.

$$\frac{\partial}{\partial\beta}m_{\beta,i} = -Z_i'Z_i\frac{\left\{pS_{i,C_i}\sum_{j\le C_i}w_{i,j}\right\}^2}{\mathcal{E}_i(1-\mathcal{E}_i)},$$

$$\frac{\partial}{\partial\alpha_k}m_{\beta,i} = -Z_i'\frac{\{pS_{i,C_i}\}^2\{\sum_{j\le C_i}w_{i,j}\}\{\sum_{j\le C_i}w_{i,j}\psi_k(j)\}}{\mathcal{E}_i(1-\mathcal{E}_i)} = \left(\frac{\partial}{\partial\beta}m_{\alpha_k,i}\right)', \quad k = 0, \ldots, K,$$

$$\frac{\partial}{\partial\alpha_r}m_{\alpha_k,i} = -\frac{\{pS_{i,C_i}\}^2\{\sum_{j\le C_i}w_{i,j}\psi_k(j)\}\{\sum_{j\le C_i}w_{i,j}\psi_r(j)\}}{\mathcal{E}_i(1-\mathcal{E}_i)}, \quad k, r = 0, \ldots\ldots, K.$$

More simulation and data analysis results

Computational time

Figure 4 reports illustrative computational times for the simulations in Tables 1 and 2 (with $q = 0.5$ and $c = 0.25$), to demonstrate feasibility. Specifically, it presents the average elapsed time per replication (in seconds), including both point and interval estimation, as a function of the number of time intervals *J*. The results show that DBH is faster than SBH, PCR is faster than GPH, and although computation time increases with *J*, the overall burden remains reasonably low.

Small sample size and high missing rates

To evaluate the impact of small sample sizes and high missing rates of Δ_i, we conducted an additional sensitivity analysis (Table 6). Specifically, we considered $n = 200$ and 400, missing rates $1 - q = 0.3$ and 0.7, with $J = 20$ (other settings same as in Table 1). The results show that small *n* or high missing rate markedly deteriorate estimator performance (e.g., CPs fall below nominal values and ASEs may become unstable). Nevertheless, performance improves as *n* or *q* increases. In practice, since *q* is under the investigator's control (e.g., via the TSCS card-drawing

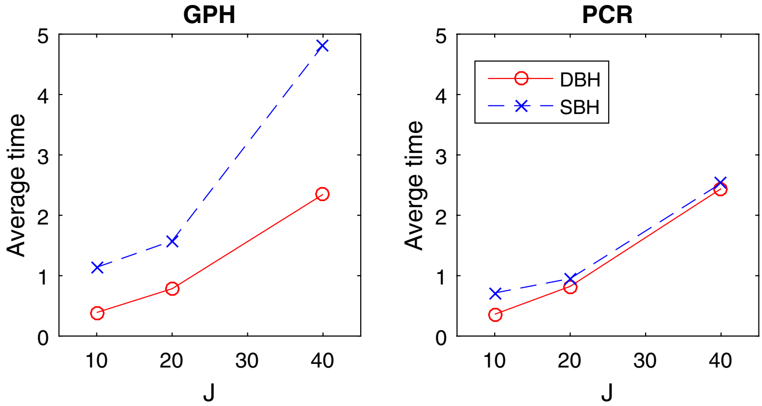

Fig. 4 Average elapsed time per replication (in seconds) for the DBH and SBH methods as a function of the number of time intervals J, under the simulation setting with $q = 0.5$ and $c = 0.25$ for the GPH model (Table 1) and the PCR model (Table 2)

Table 6 Sensitivity analysis for small sample sizes (n) and high missing rate $(1 - q)$

n	$1 - q$		DBH				SBH			
			Bias	SD	ASE	CP	Bias	SD	ASE	CP
200	0.7	β_1	1.071	4.694	460.119	47.5	0.297	1.605	16.640	56.8
		β_2	− 2.005	3.503	0.345	46.8	− 0.143	0.729	0.438	55.6
	0.3	β_1	0.048	0.420	0.398	95.2	0.017	0.376	0.368	92.8
		β_2	− 0.092	0.292	0.230	94.7	− 0.058	0.245	0.212	91.9
400	0.7	β_1	0.605	2.088	0.655	74.5	0.102	0.790	0.607	84.5
		β_2	− 0.839	1.725	0.365	73.7	− 0.105	0.474	0.356	82.7
	0.3	β_1	0.032	0.272	0.263	94.7	0.018	0.260	0.253	94.5
		β_2	− 0.040	0.156	0.149	95.0	− 0.025	0.147	0.143	95.7

mechanism), it is advisable to adopt a larger q for survey design when the sample size is limited. Because the investigator never observes which question is answered, privacy protection remains intact for a large q.

Misspecified design parameters and working models

To assess the impact of misspecifying the design parameters (q, c) and of fitting an incorrect working model, we conducted a sensitivity analysis (Table 7). The true data-generating model was GPH with $(q, c) = (0.5, 0.25)$, as in Table 1 with $J = 20$, but the analysis was performed under misspecified values of (q, c) or under a mismatched model (true GPH but fitted PCR). The results indicate that when the assumed (q, c) values are close to their true counterparts, the estimators remain reasonably robust, whereas larger deviations can induce noticeable bias. Moreover, model misspecification between GPH and PCR appears to have little practical impact on estimation performance in the scenarios considered.

Table 7 Sensitivity analysis results under misspecified design parameters (q, c) and mismatched working models

Assumed model		DBH				SBH			
		Bias	SD	ASE	CP	Bias	SD	ASE	CP
$q = 0.45$	β_1	0.116	0.337	0.295	93.9	0.082	0.296	0.285	94.6
	β_2	− 0.121	0.216	0.177	93.0	− 0.088	0.174	0.169	95.1
$q = 0.40$	β_1	0.249	0.519	0.342	89.5	0.171	0.354	0.335	93.1
	β_2	− 0.255	0.342	0.215	87.6	− 0.181	0.217	0.208	93.2
$c = 0.3$	β_1	0.075	0.294	0.275	94.6	0.055	0.279	0.269	94.8
	β_2	− 0.077	0.179	0.159	94.4	− 0.058	0.160	0.154	95.7
$c = 0.4$	β_1	0.183	0.372	0.334	93.0	0.158	0.348	0.329	94.1
	β_2	− 0.183	0.249	0.195	90.7	− 0.157	0.199	0.189	91.9
PCR	β_1	0.099	0.306	0.302	95.8	0.015	0.257	0.250	94.6
	β_2	− 0.102	0.185	0.188	97.4	− 0.019	0.146	0.142	95.5

The true data-generating model is a GPH with design parameters $(q, c) = (0.5, 0.25)$

Table 8 Sensitivity analysis results under noncompliance in the RRT design

π		DBH				SBH			
		Bias	SD	ASE	CP	Bias	SD	ASE	CP
0.1	β_1	− 0.106	0.242	0.238	91.7	− 0.113	0.236	0.233	90.9
	β_2	0.106	0.134	0.129	83.3	0.115	0.129	0.126	81.5
0.3	β_1	− 0.325	0.209	0.215	66.1	− 0.326	0.207	0.212	64.4
	β_2	0.328	0.110	0.110	17.2	0.330	0.108	0.108	15.6

A fraction π of respondents who were asked the sensitive question were assumed to provide the opposite response to their true Δ_i

Noncompliance and skewed censoring

Our analysis relies on the assumption that respondents provide truthful answers under the privacy protection offered by the RRT design; a violation of this assumption renders the model generally non-identifiable. To assess the potential impact of noncompliance, we conducted a sensitivity analysis (Table 8) under the same data-generating mechanism as in Table 1 with $J = 20$, where a fraction π of respondents asked the sensitive question were assumed to provide the opposite response to their true Δ_i. The results indicate that even moderate levels of noncompliance substantially distort estimation performance, underscoring the importance of ensuring respondent trust in the privacy protection mechanism.

To examine the impact of skewed censoring, we considered settings where the censoring time C_i follows $\min\{\lceil \text{Exp}(\mu) \rceil, 20\}$, with $\lceil \cdot \rceil$ denoting the ceiling function. When $\mu = 5$, C_i has mean 5.42, corresponding to a left-skewed distribution with $\Pr(\Delta_i = 1) = 0.26$, and when $\mu = 40$, C_i has mean 15.97, corresponding to a right-skewed distribution with $\Pr(\Delta_i = 1) = 0.51$. The results (Table 9) indicate that estimation performance is essentially unaffected by the shape of C_i.

Table 9 Simulation results under skewed censoring distributions

μ		DBH				SBH			
		Bias	SD	ASE	CP	Bias	SD	ASE	CP
5	β_1	0.066	0.380	0.327	94.4	0.046	0.344	0.312	92.2
	β_2	− 0.052	0.203	0.179	94.2	− 0.028	0.182	0.170	91.3
40	β_1	0.023	0.242	0.236	95.7	0.016	0.237	0.232	95.0
	β_2	− 0.025	0.143	0.139	95.8	− 0.016	0.138	0.136	94.8

The censoring time was generated as $C_i \sim \min\{\lceil \mathrm{Exp}(\mu) \rceil, 20\}$, where $\lceil \cdot \rceil$ denotes the ceiling function. For $\mu = 5$, C_i has mean 5.42 (left-skewed, $\Pr(\Delta_i = 1) = 0.26$); for $\mu = 40$, C_i has mean 15.97 (right-skewed, $\Pr(\Delta_i = 1) = 0.51$)

Table 10 Model selection results based on AIC, AICc, BIC, and 5-fold cross-validation (CV) for the extramarital sex data and the militant connection data

Method	m	GPH				PCR			
		AIC	AICc	BIC	CV	AIC	AICc	BIC	CV
Extramarital sex data									
SBH	1	941.70	941.84	974.53	− 94.68	941.70	941.85	974.54	− 94.69
	2	931.07*	931.25*	968.60*	− 93.63*	931.15*	931.33*	968.67*	− 93.63*
	3	932.63	932.85	974.84	− 93.64	932.75	932.97	974.96	− 93.64
	4	934.45	934.73	981.36	− 93.64	934.60	934.87	981.51	− 93.65
	5	936.35	936.69	987.95	− 93.64	936.52	936.85	988.12	− 93.65
DBH	–	1000.43	1005.64	1206.83	− 93.83	1000.35	1005.56	1206.74	− 93.68
Militant connection data									
SBH	1	3183.03	3183.07	3217.79	− 318.34	3182.86	3182.90	3217.62	− 318.32
	2	3101.40	3101.44	3141.95*	− 309.93	3101.53	3101.57	3142.07*	− 309.95
	3	3098.58*	3098.64*	3144.92	− 309.44	3098.54	3098.60*	3144.88	− 309.45
	4	3098.66	3098.74	3150.80	− 309.24*	3098.53*	3098.61	3150.67	− 309.24*
	5	3099.63	3099.72	3157.56	− 310.27	3099.45	3099.54	3157.37	− 310.25
DBH	–	3220.39	3225.51	3666.43	− 309.27	3219.91	3225.03	3665.95	− 309.22

An asterisk (*) denotes the number of basis functions (*m*) for SBH selected by each criterion

Model selections

To complement the AIC-based model selection presented in the main text, we additionally evaluated corrected-AIC (AICc), BIC, and 5-fold cross-validation (CV) using log-likelihood as the predictive measure (Table 10). For the extramarital sex data, all criteria consistently supported SBH with $m = 2$. For the militant connection data, AIC and AICc favored $m = 3$ or 4, BIC preferred $m = 2$, and CV suggested $m = 4$, with predictive performance comparable between SBH ($m = 4$) and DBH. Given the large number of discrete time points ($J = 83$) in the militant connection data, SBH remains preferable for estimation stability.

Figure 5 further shows estimated regression coefficients and the baseline survival probability at [*J*/2] as functions of the polynomial degree, with $[\cdot]$ denoting the rounding function. The results indicate that estimates with $m \geq 2$ are highly consistent, indicating that the values of *m* selected by the criteria in Table 10 are sufficiently large to approximate the baseline function well, thereby yielding stable estimates. Although different criteria occasionally selected slightly different models (e.g., in the

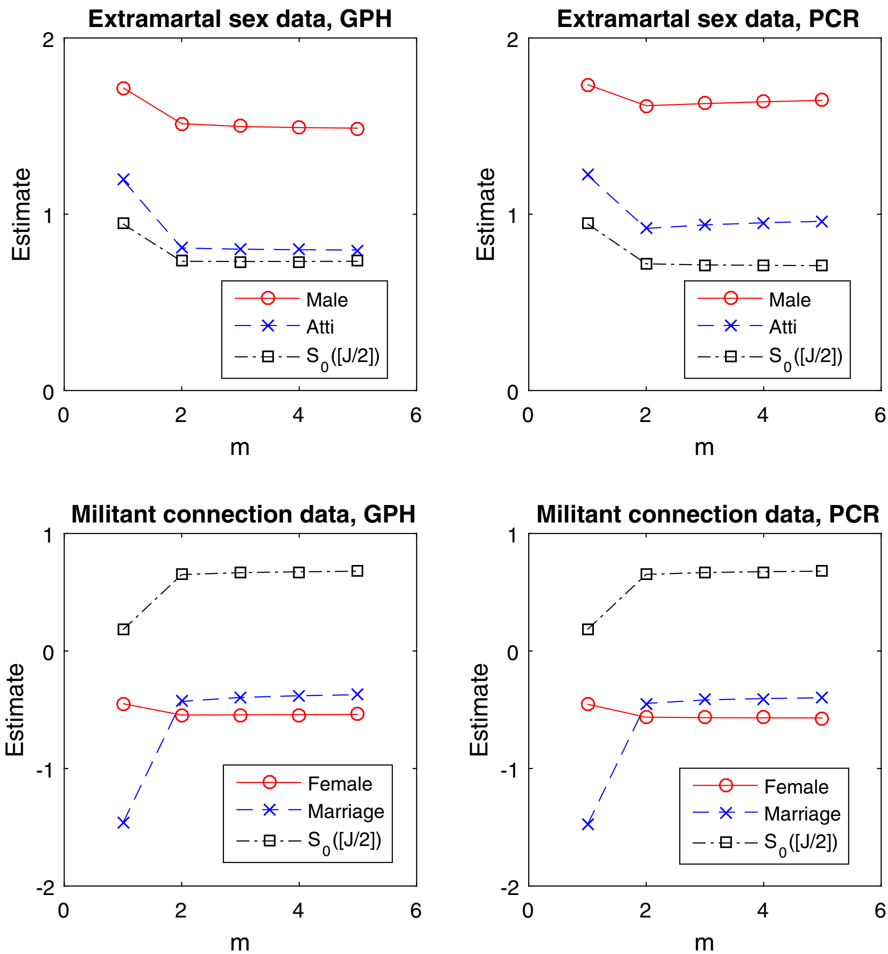

Fig. 5 Sensitivity analysis of the SBH method with respect to the polynomial degree m. The panels display the estimated regression coefficients and the baseline survival probability at $[J/2]$ under different choices of m

militant connection application), the estimated parameters remained largely consistent across model choices (see Tables 4, 5 and Fig. 5).

Funding Open Access funding enabled and organized by National Yang Ming Chiao Tung University

Data availability The extramarital sex data used in this study were obtained from the Center for Survey Research, Academia Sinica, with permission. Access to these data may be granted upon request and subject to approval by the Center for Survey Research, Academia Sinica. The militant connection data used in this paper are publicly available through the rr R package.

References

Berger, M., & Schmid, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling, 18*, 322–345.

Blair, G., Imai, K., & Schmid, M. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association, 110*(511), 1304–1319.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *American Sociologist, 6*, 308–311.

Carroll, R. J., Ruppert, D., Stefansk, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). Chapman & Hall/CRC Press.

Chen, D., Sun, J., & Peace, K. E. (2012). *Interval-censored time-to-event data: Methods and applications* (1st ed.). Chapman & Hall/CRC Press.

Gjestvang, C. R., & Singh, S. (2006). A new randomized response model. *Journal of the Royal Statistical Society: Series B, 68*, 523–530.

Greenberg, B. G., Abul-Ela, A. L., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association, 64*, 520–539.

Horvitz, D. G., Shah, B. V., & Simmons, W. R. (1967). The unrelated question randomized response model. In *Proceedings of the social statistics section* (pp. 65–72). American Statistical Association.

Hsieh, S. H., Lee, S. M., Li, C. S., & Tu, S. H. (2016). An alternative to unrelated randomized response techniques with logistic regression analysis. *Statistical Methods and Applications, 25*, 601–621.

Huang, J. (1996). Efficient estimation for the cox model with interval censoring. *The Annals of Statistics, 24*, 540–568.

Jewell, N. P., & van der Laan, M. (2003). Current status data: Review, recent developments and open problems. *Handbook in Statistics, 23*, 625–642.

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika, 77*, 436–438.

Lorentz, G. G. (1986). *Bernstein polynomials* (2nd ed.). Chelsea Publishing Company.

Narjis, G., & Shabbir, J. (2023). An improved two-stage randomized response model for estimating the proportion of sensitive attribute. *Sociological Methods and Research, 52*, 335–355.

Prentice, R. L., & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics, 34*, 57–67.

Scheers, N. J., & Dayton, C. M. (1988). Covariate randomized response models. *Journal of the American Statistical Association, 83*, 969–974.

Singh, S., Singh, R., & Mangat, N. S. (2000). Some alternative strategies to Moor's model in randomized response sampling. *Journal of the Statistical and Planning Inference, 83*, 243–255.

Sun, J. (2006). *The statistical analysis of interval-censored failure time data* (1st ed.). Springer.

Thompson, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics, 33*, 463–470.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B, 38*, 290–295.

Tutz, G., & Schimid, M. (2016). *Modeling discrete time-to-event data* (1st ed.). Springer.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 60*, 63–69.

Wen, C. C., & Chen, Y. H. (2025). Regression analysis of randomized response event time data. *Statistica Sinica*. https://doi.org/10.5705/ss.202022.0320