Revisiting the Cumulative Incidence Function With Competing Risks Data

David M. Zucker

Department of Statistics and Data Science, The Hebrew University of Jerusalem, Jerusalem, Israel.

E-mail: david.zucker@mail.huji.ac.il

Malka Gorfine

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel.

Summary. We consider estimation of the cumulative incidence function (CIF) in the competing risks Cox model. We study three methods. Methods 1 and 2 are existing methods while Method 3 is a newly-proposed method. Method 3 is constructed so that the sum of the CIF's across all event types at the last observed event time is guaranteed, assuming no ties, to be equal to 1. The performance of the methods is examined in a simulation study, and the methods are illustrated on a data example from the field of computer code comprehension. The newly-proposed Method 3 exhibits performance comparable to that of Methods 1 and 2 in terms of bias and variance, and better than that of Methods 1 and 2 in terms of confidence interval coverage rates.

Keywords: Competing events; Computer program comprehension; Cox regression; Prediction; Survival analysis.

1. Introduction

Competing risks arise when individuals are susceptible to several types of event and can experience at most one event. Analysis of time to event without distinguishing between the different event types often yields an inadequate picture of the data (Kalbfleisch and Prentice, 2002, page 249). There is a vast literature on competing risks, and the topic remains an active area of research.

This paper is motivated by the recently conducted experiment of Ajami et al. (2019) in the area of computer program comprehension. Their goal was to measure how different syntactic and other factors influence code complexity and comprehension. To reach many subjects and obtain accurate measurements, they implemented a website for the experiment designed based on some gamification principles, for details see Ajami et al. (2019). The design consists of 40 code snippets, with each participant asked to interpret a subset of 11–14 snippets, presented in random order. The outcomes were time to answer and the accuracy of the snippet interpretation, i.e. correct or incorrect. Thus, correct response and incorrect response were competing events. Out of the 2761 recorded trials, only 27 (0.98%) of them ended in right censoring (i.e. no answer was provided after a certain period of time and the participant gave up).

Modeling based on *cause-specific hazard* functions is a popular and useful approach for handling competing events (Putter et al., 2007, Section 3.2). If we let T denote the time to event and D denote the type of event, the cause-specific hazard $\lambda_j(t|\mathbf{z})$ for event type j, $j = 1, \ldots, J$, for an individual with covariate vector $\mathbf{z}$ is defined as

$$\lambda_j(t|\mathbf{z}) = \lim_{\epsilon \downarrow 0} \epsilon^{-1} P(T \in [t, t+\epsilon), D = j | \mathbf{z}, T \geq t) \,.$$

This quantity represents the instantaneous incidence rate of cause j, given that the individual was free of any event up to time t. Useful functions for prediction are the cause-specific cumulative incidence functions (CIFs), defined as

$$F_j(t|\mathbf{z}) = P(T \leq t, D = j|\mathbf{z}) = \int_0^t S(u - |\mathbf{z}) \lambda_j(u|\mathbf{z}) \, du \,, \tag{1}$$

where

$$S(t|\mathbf{z}) = P(T > t|\mathbf{z}) = \exp\left\{ - \sum_{m=1}^{J} \Lambda_m(t|\mathbf{z}) \right\}$$

with

$$\Lambda_j(t|\mathbf{z}) = \int_0^t \lambda_j(u|\mathbf{z}) \, du \,.$$

In the case of competing risks with no covariates, the CIF of each event type is usually estimated by the Aalen-Johansen estimator (Aalen and Johansen, 1978). It can be shown that when the last observed time is an event time and there are no ties, the sum of the Aalen-Johansen estimators of the CIFs over all event types evaluated at the last event time is equal to 1. In the presence of covariates, however, the situation is more complicated.

In this paper, we study three methods for estimating the CIFs with covariates under a Cox-type model for the cause-specific hazards. We refer to these methods as Method 1, Method 2, and Method 3. Methods 1 and 2 are existing estimators, while Method 3 is a newly-proposed estimator. The new method is constructed so as to guarantee, in the absence of ties, that the sum of the CIF's across all event types at the last observed event time is equal to 1. Section 2 presents the methods, Section 3 presents a simulation study comparing the methods, Section 4 presents an application of the methods to the Ajami et al. program comprehension study, and Section 5 presents a short discussion.

2. Methods Considered

To set the stage, we first consider the standard setup of ordinary survival data analyzed using the Cox model (Cox, 1972). For each individual i, we denote by X_i the observed follow-up time on individual i until the occurrence of an event or censoring, and we set D_i equal to 1 if individual i experienced an event and equal to 0 if individual i was censored. We define $N_i(t) = D_i I(X_i \leq t)$ and $Y_i(t) = I(X_i \geq t)$. We assume that the event time has a continuous distribution, so that the probability of tied event times is 0.

A common estimator of the survival function $S(t|\mathbf{z})$ is given, as in Section 8.8 of Klein and Moeschberger (2003), by

$$\hat{S}^{(1)}(t|\mathbf{z}) = \exp\left\{ -\hat{\theta}(\mathbf{z}) \hat{\Lambda}_0(t) \right\} \,, \tag{2}$$

where $\hat{\theta}(\mathbf{z}) = \exp\left\{\hat{\boldsymbol{\beta}}^T \mathbf{z}\right\}$ and

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \sum_{r=1}^{n} Y_r(t)\hat{\theta}(\mathbf{Z}_r) \right\}^{-1} dN_i(t)$$

is the Breslow estimator of the cumulative baseline hazard function. An alternate estimator, given by Eqn. (7.2.34) of Andersen et al. (1993), is

$$\hat{S}^{(2)}(t|\mathbf{z}) = \mathcal{P}_0^t \left\{1 - \hat{\theta}(\mathbf{z}) d\hat{\Lambda}_0(s)\right\} = \prod_{k=1}^{K(t)} \left\{1 - \hat{\theta}(\mathbf{z}) \Delta\hat{\Lambda}_0\left(T_{(k)}\right)\right\}, \tag{3}$$

where $\mathcal{P}_0^t$ denotes the product integral, $T_{(k)}$ denotes the k-th ordered event time, $K(t)$ denotes the number of event times in the interval $[0, t]$, and $\Delta\hat{\Lambda}_0(s) = \hat{\Lambda}_0(s) - \hat{\Lambda}_0(s-)$. Kalbfleisch and Prentice (2002), in Section 4.3, present another possible estimator,

$$\hat{S}^{(3)}(t|\mathbf{z}) = \left(\prod_{k=1}^{K(t)} \hat{\alpha}_k \right)^{\hat{\theta}(\mathbf{z})}, \tag{4}$$

where, in the absence of tied survival times, $\hat{\alpha}_k$ is given by

$$\hat{\alpha}_k = \left\{ 1 - \frac{\hat{\theta}\left(\mathbf{Z}_{I(k)}\right)}{\sum_{r=1}^{n} Y_r(T_{(k)})\hat{\theta}(\mathbf{Z}_r)} \right\}^{1/\hat{\theta}(\mathbf{Z}_{I(k)})}.$$

Here, $I(k)$ is the index of the individual who had the event at time $T_{(k)}$. Kalbfleisch and Prentice derived this estimator using a nonparametric maximum likelihood argument.

All three of the above estimators are presented in Section VII.2.3 of Andersen et al. (1993). If the end-of-study risk set is large, these three estimators are nearly identical, whereas if the end-of-study risk set is small they can differ substantially. With $\hat{S}^{(2)}(t|\mathbf{z})$, the quantity $\{1-\hat{\theta}(\mathbf{z})\Delta\hat{\Lambda}_0(T_{(k)})\}$ can go negative; a simple fix is to set $\hat{S}^{(2)}(t|\mathbf{z})$ to 0 when this occurs. The estimator $\hat{S}^{(3)}(t|\mathbf{z})$, like the univariate Kaplan-Meier survival function estimator, drops to 0 if the last observation time is as an event time. Our proposed analogue to $\hat{S}^{(3)}(t|\mathbf{z})$ in the competing risk case, presented below, is constructed so as to achieve an analogous property: that the estimated total cumulative distribution function of the time to event, taken over all event types, is equal to 1 if the last observation time is an event time.

If we define

$$\hat{\gamma}_k(\mathbf{z}) = 1 - \hat{\alpha}_k^{\hat{\theta}(\mathbf{z})} = 1 - \left\{ 1 - \frac{\hat{\theta}(\mathbf{Z}_{I(k)})}{\sum_{r=1}^{n} Y_r(T_{(k)})\hat{\theta}(\mathbf{Z}_r)} \right\}^{\hat{\theta}(\mathbf{z})/\hat{\theta}(\mathbf{Z}_{I(k)})} \tag{5}$$

and

$$\hat{\Gamma}(t|\mathbf{z}) = \int_0^t \left[1 - \left\{ 1 - \frac{\sum_{i=1}^{n} \hat{\theta}(\mathbf{Z}_i) dN_i(s)}{\sum_{i=1}^{n} Y_i(s)\hat{\theta}(\mathbf{Z}_i)} \right\}^{\hat{\theta}(\mathbf{z})/\sum_{i=1}^{n} \hat{\theta}(\mathbf{Z}_i) dN_i(s)} \right] \sum_{i=1}^{n} dN_i(s) \tag{6}$$

we can write

$$\hat{S}^{(3)}(t|\mathbf{z}) = \prod_{k=1}^{K(t)} \{1 - \hat{\gamma}_k(\mathbf{z})\} = \mathcal{P}_0^t \left\{1 - \Delta\hat{\Gamma}(t|\mathbf{z})\right\} . \tag{7}$$

Here, if we evaluate the expression

$$\left\{1 - \frac{\sum_{i=1}^n \hat{\theta}(\mathbf{Z}_i) dN_i(s)}{\sum_{i=1}^n Y_i(s)\hat{\theta}(\mathbf{Z}_i)}\right\}^{\hat{\theta}(\mathbf{z})/\sum_{i=1}^n \hat{\theta}(\mathbf{Z}_i) dN_i(s)}$$

at $s = T_{(k)}$, we get

$$\left\{1 - \frac{\hat{\theta}(\mathbf{Z}_{I(k)})}{\sum_{r=1}^n Y_r(T_{(k)})\hat{\theta}(\mathbf{Z}_r)}\right\}^{\hat{\theta}(\mathbf{z})/\hat{\theta}(\mathbf{Z}_{I(k)})}$$

in correspondence with (5), and if we evaluate it any other s value, we get $1^\infty = 1$, making the bracketed term in (6) equal to 0. Comparing (7) with (3), we can draw an association between $\Delta\hat{\Gamma}(t|\mathbf{z})$ and $\hat{\theta}(\mathbf{z})\Delta\hat{\Lambda}_0(T_{(k)})$. If $\sum_{i=1}^n Y_i(s)\hat{\theta}(\mathbf{Z}_i)$ is large, the two quantities are approximately equal, as may be seen using the approximations $\log(1-u) \doteq -u$ and $1 - e^{-v} \doteq v$.

We now move to the competing risk setting. We let $T, D, \lambda_j(t|\mathbf{z})$, and $F_j(t|\mathbf{z})$ be defined as in the introduction. We again denote the k-th ordered event time by $T_{(k)}$, and we denote the corresponding event type by $D_{(k)}$. A common approach to modeling competing risks data is to use a Cox-model form for the cause-specific hazard:

$$\lambda_j(t|\mathbf{z}) = \lambda_{0j}(t)\exp(\boldsymbol{\beta}_j^T\mathbf{z}) \quad j = 1, \ldots, J .$$

It is well-known that $\boldsymbol{\beta}_j$ can be consistently estimated using the Cox partial likelihood with event types other than j handled as censoring (Kalbfleisch and Prentice, 2002, Section 8.2.3).

Corresponding to the three survival function estimators presented above for ordinary survival data, we can define three estimators of the CIF. Define $\hat{\theta}_j(\mathbf{z}) = \exp\left(\hat{\boldsymbol{\beta}}_j^T\mathbf{z}\right)$ and $\hat{\theta}_{ij} = \hat{\theta}_j(\mathbf{Z}_i)$. The analogue of (2) is then

$$\hat{F}_j^{(1)}(t|\mathbf{z}) = \int_0^t \exp\left\{-\sum_{m=1}^J \hat{\Lambda}_m(s-|\mathbf{z})\right\} d\hat{\Lambda}_j(s|\mathbf{z}) ,$$

where

$$\hat{\Lambda}_j(s|\mathbf{z}) = \hat{\theta}_j(\mathbf{z}) \int_0^s A_j(u)^{-1} \sum_{i=1}^n dN_{ij}(u)$$

with

$$A_j(u) = \sum_{i=1}^n Y_i(u)\hat{\theta}_{ij} .$$

The analogue of (3), as given by Section 8.5.1 of Beyersmann and Scheike (2014), is

$$\hat{F}_j^{(2)}(t|\mathbf{z}) = \int_0^t \hat{P}(s-|\mathbf{z}) d\hat{\Lambda}_j(s|\mathbf{z})$$

with

$$\hat{P}(s|\mathbf{z}) = \mathcal{P}_0^s \left\{ 1 - \sum_{j=1}^{J} d\hat{\Lambda}_j(u|\mathbf{z}) \right\}_+ = \prod_{k=1}^{K(s)} \left\{ 1 - \sum_{j=1}^{J} \Delta\hat{\Lambda}_j(T_{(k)}|\mathbf{z}) \right\}_+$$

where $a_+ = \max(a, 0)$.

Our proposed analogue of (7) is

$$\hat{F}_j^{(3)}(t|\mathbf{z}) = \sum_{k=1}^{K(t)} \left[\prod_{r=1}^{k-1} \{1 - \hat{\gamma}_{r\bullet}(\mathbf{z})\} \right] \hat{\gamma}_{kj}(\mathbf{z})$$

where, analogously to (5), we define

$$\hat{\gamma}_{kj}(\mathbf{z}) = 1 - \left\{ 1 - \frac{\theta_j\left(\mathbf{Z}_{I(k)}\right) I(D_{(k)} = j)}{A_j(T_{(k)})} \right\}^{\hat{\theta}_j(\mathbf{z})/\theta_j(\mathbf{Z}_{I(k)})}$$

and we set $\hat{\gamma}_{k\bullet}(\mathbf{z}) = \sum_{j=1}^{J} \gamma_{kj}(\mathbf{z})$.

As in the ordinary survival case, if the end-of-study risk set is large, the three estimators, $\hat{F}_j^{(m)}$, $m = 1, 2, 3$, are nearly identical, whereas if the end-of-study risk set is small they can differ substantially. Our proposed analogue of (7) does not have the nonparametric maximum likelihood interpretation that (7) has, but it is still a plausible estimator.

We define $F_\bullet(t|\mathbf{z}) = \sum_{j=1}^{J} F_j(t|\mathbf{z})$, which is the probability that an individual with covariate vector $\mathbf{z}$ experiences an event of some type during the interval $[0, t]$. Correspondingly, for $m = 1, 2$, and 3, we define $\hat{F}_\bullet^{(m)}(t|\mathbf{z}) = \sum_{j=1}^{J} \hat{F}_j^{(m)}(t|\mathbf{z})$.

It is an algebraic fact, which can be proved by induction, that for any $c_1, \ldots, c_K$ we have

$$1 - \sum_{k=1}^{K} \left\{ \prod_{r=1}^{k-1} (1 - c_r) \right\} c_k = \prod_{r=1}^{K} (1 - c_r)$$

Thus,

$$1 - \hat{F}_\bullet^{(3)}\left(T_{(K)}|\mathbf{z}\right) = \prod_{r=1}^{K} \{1 - \hat{\gamma}_{r\bullet}(\mathbf{z})\} \ .$$

Now, if $T_{(K)}$ is the last observed follow-up time (i.e., the last observed follow-up time was an event), then $A_j(T_{(K)}) = \theta_j(\mathbf{Z}_{I(K)})$, and so we have $\hat{\gamma}_{Kj}(\mathbf{z}) = I(D_{I(K)} = j)$ and $\hat{\gamma}_{K\bullet}(\mathbf{z}) = 1$. Thus, in this case, such as with uncensored data, we obtain $1 - \hat{F}_\bullet^{(3)}(T_{(K)}) = 0$ and $\hat{F}_\bullet^{(3)}(T_{(K)}) = 1$. The estimators $\hat{F}_j^{(1)}(t)$ and $\hat{F}_j^{(2)}(t)$ do not have this property. In fact, for these estimators, $\hat{F}_\bullet^{(m)}(T_{(K)})$ can exceed 1.

We also considered a 95% simultaneous confidence band of the form

$$\hat{F}_j^{(m)}(t|\mathbf{z}) \pm c_{jm,0.95}$$

(i.e., a fixed-width band) for $F_j(t|\mathbf{z})$ over $t \in [0, T_{(K)}]$. We computed the critical value $c_{jm,0.95}$ using the weighted bootstrap (Kosorok and Song, 2007, Section 8.2). For each

bootstrap replication b ($b = 1, \ldots, B$), a set of weights w_{bi}° is generated as random draws from the $Exp(1)$ distribution, and then normalized weights are computed as $w_{bi} = w_{bi}^{\circ}/(n^{-1}\sum_{r=1}^{n} w_{br}^{\circ})$. In running the Cox models using the function `coxph` in the R package `survival`, we incorporate these weights w_{bi} using the `weights` argument. In addition, we incorporate the weights in other expressions appearing in the formulas for the estimators under study. The estimators involve various quantities of the form

$$\sum_{i=1}^{n} term_i \,,$$

such as

$$A_j(u) = \sum_{i=1}^{n} Y_i(u)\hat{\theta}_{ij} \,.$$

In the bootstap estimates, the quantities of the form

$$\sum_{i=1}^{n} term_i$$

are replaced by

$$\sum_{i=1}^{n} w_{bi}\, term_i \,,$$

so that, for example, we replace $A_j(u)$ by

$$A_j^{(boot,b)}(u) = \sum_{i=1}^{n} w_{bi} Y_i(u)\hat{\theta}_{ij} \,.$$

Subsequently, for each bootstrap replication b we compute

$$\Delta_{bjm}(\mathbf{z}) = \max_{t\in[0,T_{(K)}]} |\hat{F}_j^{(m,boot,b)}(t|\mathbf{z}) - \hat{F}_j^{(m)}(t|\mathbf{z})|$$

The critical value $c_{jm,0.95}$ is then taken to be the 95th percentile of the set of values $\Delta_{bjm}(\mathbf{z}), b = 1, \ldots, B$.

3. Simulation Study

We conducted a simulation study to compare the estimates $\hat{F}_j^{(m)}(t|\mathbf{z}), m = 1, 2, 3$. We considered a setup with two competing risks, one with a high final CIF (65%) and one with a low final CIF (35%). In the first set of simulations, there was a single covariate Z, with distribution $U(-0.5, 0.5)$. We used a baseline hazard of the form

$$\lambda_0(t) = \frac{\sigma p(t+a)^{p-1}}{1 + b(t+a)^p}$$

with a corresponding cumulative baseline hazard of the form

$$\Lambda_0(t) = \sigma b^{-1} \log(1 + b(t+a)^p)$$

As b tends to 0, we get a Weibull-type model with $\Lambda_0(t) = \sigma(t+a)^p$ and $\lambda_0(t) = \sigma p(t+a)^{p-1}$. We considered three shapes for the baseline hazard function, increasing ($a = 0, b = 0, p = 3$), decreasing ($a = 0.4, b = 0, p = 0.5$) and up-and-down ($a = 0, b = 0.75, p = 3$). The parameter σ was set so as to achieve CIF values approaching the desired final values at about time $t = 5$. The survival distributions were truncated at time $t = 10$. Note that the above parametric model was used only in the data generation part of the simulations; the estimators under study make no parametric assumptions about the form of the baseline hazard. We ran simulations for a sample size of 75 with no censoring and for a sample size of 150 with 50% censoring. We took the regression coefficient to be either $\log 3$ or $\log 6$ for both event types, so that the relative risk associated with a 1 unit increase in the covariate value was either 3 or 6. The estimates of $F_j(t|z)$ were computed at $z^* = -0.4$, $z^* = 0$, and $z^* = 0.4$. In all simulations, 1,000 simulation replications were run, and for each simulation replication the bootstrap confidence band procedures were carried out using 1,000 bootstrap replications. Table 1 summarizes the configurations studied under a uniformly distributed covariate.

At the suggestion of one of the reviewers, we also considered the the time-dependent prediction error of the estimates, along the lines of Schoop et al. (2011). The time-dependent prediction error is defined as

$$PE_j(t) = E[(I(T \le t, D = j) - \hat{F}_j(t|\mathbf{Z}))^2] \,.$$

We approximate $PE_j(t)$ by generating $n_{out} = 1,000$ out-of-sample observations of $(T, D, \mathbf{Z})$ and computing

$$\widehat{PE}_j(t) = \frac{1}{n_{out}} \sum_{i=n+1}^{n+n_{out}} (I(T_i \le t, D_i = j) - \hat{F}_j(t|\mathbf{Z}_i))^2 \,.$$

As a benchmark, we use

$$\widetilde{PE}_j(t) = \frac{1}{n_{out}} \sum_{i=n+1}^{n+n_{out}} (I(T_i \le t, D = j) - F_j(X_i|\mathbf{Z}_i))^2$$

and we report results on the prediction error difference

$$PED_j(t) = \widehat{PE}_j(t) - \widetilde{PE}_j(t) \,.$$

The PE does not depend on the value of z^*, so that the number of distinct scenarios for which we have have PE results is 12 rather than 36.

We also conducted a supplemental set of simulations where the final CIF was again set to be about 65% for the high CIF event and 35% for the low CIF event, but now the covariate had distribution $N(0, 4)$ truncated at ± 5. We considered the case of uncensored data with sample size 75. We used the increasing hazard. The values for the regression coefficient were again either $\log 3$ or $\log 6$ for both event types. The estimates of $F_j(t|z)$ were computed at $z^* = -1.68$, $z^* = 0$, and $z^* = 1.68$ (-1.68 and 1.68 are, respectively, the 20th and 80th percentiles of the $N(0, 4)$ distribution).

In the second set of simulations there were two covariates, a continuous covariate Z_1 distributed $U(-0.5, 0.5)$ and a binary covariate Z_2 distributed $\text{Ber}(\frac{1}{2})$. We considered the

case of uncensored data with sample size 75 and the increasing hazard. The regression coefficients of the two covariates were taken to be equal, and set to be $\log 2$, $\log 3$, or $\log 4$. estimates of $F_j(t|z)$ were computed for $z^* = (-0.4, 0)$ and $z^* = (0.4, 1)$.

Figures 1–9 present the results and Tables S.1–S.15 in the Supporting Information present the same results in tabular form. The figures show the maximum mean bias of the CIF estimates up to the 90th percentile of the last observed event time, the standard error of the estimates at the 90th percentile of the last observed event time (the standard error generally increased over time), the empirical coverage rates of the 95% confidence bands, the half-widths of the 95% confidence bands, and the prediction error difference results. In computing the bias, the true CIF was computed using the formula (1), where the integral was evaluated numerically using Simpson's rule with the width of the interval between quadrature points equal to 0.005.

Regarding the bias, Figure 1 depicts the results for a single uniformly distributed covariate, Figure 4 the results for a single normally distributed covariate, and Figure 7 the results for the case of two covariates. In the single covariate settings, Methods 2 and 3 generally yielded a maximum bias of less than 0.01 whereas the bias with Method 1 tended to be higher. In the two-variable simulations, for the estimates for Cause A, Method 3 was noticeably better than Methods 1 and 2. The three estimates were comparable in terms of maximum standard error. Regarding the confidence bands, Methods 1 and 2 often yielded low coverage rates, in some cases as low as 0.8. Method 3 tended to yield wider confidence bands but with proper coverage rates.

Figures 2, 5, and 8 (for the setting of a single uniformly distributed covariate, the setting of a single normally distributed covariate, and the setting of two covariates, respectively) show, for the configurations without censoring, various quantiles (1%, 10%, 50%, 90%, 99%) of the total CIF (i.e., the CIF summed over the two risks) at the last event time. Tables S.16–S.18 in the Supporting Information present the same results in tabular form. Since $\hat{F}_{\bullet}^{(3)}(T_{(K)}|\mathbf{z}) = 1$ by construction, this estimator is not included in these figures and tables. Although without censoring, the estimators should be exactly 1, we see that $\hat{F}_{\bullet}^{(1)}$ and $\hat{F}_{\bullet}^{(2)}$ could deviate from 1, sometimes substantially, and can be less than or greater than 1. The magnitude of the deviation from 1 was usually larger with $\hat{F}_{\bullet}^{(1)}$ than with $\hat{F}_{\bullet}^{(2)}$. Table 2 shows, for the case of a single uniformly distributed covariate, the percentage of scenarios for which the final total CIF exceeded 1. For Method 1, the exceedance percentage was greater than 90% in all scenarios and equal to 100% for most scenarios. For Method 2, the exceedence percentage was greater than 85% in all scenarios except those with covariate value equal to -0.4. Table 3 shows the corresponding results for the case of a single normally distributed covariate. For both methods, the exceedance percentage was greater than 90% for all scenarios and equal to 100% for most scenarios. Table S.19 in the Supporting Information show the results for the two-covariate setting, where the findings were similar.

Figures 3, 6, and 9 present the results for the prediction error difference (PED). Tables S.5, S.10, and S.15 in the Supporting Information present the same results in tabular form. Overall, the PED's were low with all three methods. For the case of a single uniformly distributed covariate, the PED's were comparable across the three methods. For the case of a single normally distributed covariate, the PED's with Methods 2 and 3 were noticably lower than with Method 1 and comparable to each other, with a slight

advantage for Method 3. For the two-covariate setting, the PED's were comparable, with a slight advantage of Methods 2 and 3 over Method 1.

4. Real Data Example – Program Comprehension

Most of a software engineer's time is spent reading codes. Sometimes this is their own code, while most of the time it is someone else's code. This reading is often named *program comprehension*. Being good at program comprehension is a critical skill but it is notoriously hard and time consuming. Surprisingly, there has been relatively little empirical work on how program structures effect comprehension. Recently, Ajami et al. (2019) used an experimental platform fashioned as an online game-like environment to measure how quickly and accurately 222 professional programmers can interpret code snippets with similar functionality but different structures. Their goal was to measure how different syntactic and other factors influence code complexity and comprehension. For example, what is the effect of control structures on code complexity? Is the complexity of an `if` the same as that of a `for`? For the complete list of their research questions, see Ajami et al. (2019).

The following summary of the design description is based on Ajami et al. (2019). The experiment was conducted by showing participants short code snippets which they needed to interpret. All code segments checked whether a number is in a set of non-overlapping ranges. The design consists of 40 code snippets: 12 each with 3-range and 4-range versions, 9 with 2-range versions, and 7 special loop cases. Table 4 provides a concise description of the snippets. In a pilot study they found that reading 40 snippets is too much for a single participant to perform, so a subset of snippets was selected to each participant. The selection was done efficiently in terms of including pairs or sets of snippets that are meaningful to compare to each other. The total number of snippets presented to each subject was between 11 and 14, presented in random order. The outcomes were the time to answer and accuracy of the response (correct/incorrect). To reach many subjects and achieve accurate measurements, the investigators implemented a website for the experiment, designed based on some gamification principles; for details see their paper. At the beginning of the experiment, a popup was opened with a demographic questionnaire and details on education and experience. The choice of a test plan did not depend on experience or any other demographic information of the participant. Afterward, an example screen was displayed, showing how the experimental screen looks and explaining the "game" rules. When the actual experiment started, a code snippet was presented and the subject were supposed to type in the snippet's output.

In the above setup, correct and incorrect response are competing events. Time was measured from displaying the code until the participant pressed the button to indicate he/she was done. Another outcome variable was the button the subject chose to click: either "I think I made it" or "skip", where skip indicates right censoring. However, skip was used only 27 times in total, out of 2761 recorded trials, such that among the 40 snippets, 12 were with a single right-censored response, 5 with two right-censored responses, 1 with three right-censored responses, and 22 with no right-censored response. Therefore, the practical effect of excluding right-censored responses is obviously negligible (if any). Thus, the dataset is essentially free of censoring. If a participant decided to quit the

experiment before completing the snippets set, the analysis consisted of the completed snippets. Another important issue is the order in which the snippets were presented to each participant, as the common framework behind the snippets may lead to learning effects. Therefore, one of the covariates in the following analysis is the snippet's place in the sequence of snippets shown to the examinee (snippet order).

In total there were 1,893 correct answers and 868 incorrect answers. On average, a question was answered by 58.15 participants. To demonstrate the differences between the three CIF estimators and the advantage of $\hat{F}_j^{(3)}$, $j = 1, 2$, we show here the result of the snippet `lp3` (a snippet of "special loop" type). The analysis is based on 69 players; 49 provided a correct answer and 20 provided an incorrect answer. For this particular snippet, there were no "skip" responses, so that there is no censoring at all. The covariates included in the Cox regression analysis were the participant's age, sex and years of experience (YoE), the snippet's order and the order squared. For this snippet, the proportional hazards assumption was tested (Grambsch and Therneau, 1994) and the global tests for the correct and incorrect answers indicated that there is no evidence of a violation of the assumption at the 0.1 significance level (for correct answers chi-square test statistic = 8.909, df=5, p-value=0.113; for incorrect answers chi-square test statistic = 3.843, df=5, p-value=0.570).

Figures 10 and 11 display $\hat{F}_j^{(m)}(\cdot|\mathbf{z})$ and $\hat{F}_\bullet^{(m)}(\cdot|\mathbf{z})$, $j = 1, 2$, $m = 1, 2, 3$, for various $\mathbf{z}$. The confidence bands were omitted for simplicity of presentation. The values of $\hat{F}_\bullet^{(m)}(T_{(K)}|\mathbf{z})$, for $m = 1, 2$, are presented with each plot; $\hat{F}_\bullet^{(3)}(T_{(K)}|\mathbf{z}) = 1$ in all cases. We see that the deviation of $\hat{F}_\bullet^{(m)}(T_{(K)}|\mathbf{z})$, $m = 1, 2$, from 1 can be substantial; in some cases the estimated end-of-study total CI was in the range 0.60-0.70. Figure 12 shows the CIFs of $\mathbf{z} = (\text{snippet order} = 2, \text{age} = 46, \text{female}, \text{YoE} = 0)$ of the three methods with 95% confidence bands.

5. Discussion

In this paper, we have studied three methods for estimating the cumulative incidence function (CIF) in the competing risks Cox model: two existing methods and a newly-proposed method. We focused on the Cox model because it is the most popular model, but we acknowlege that alternative models, such as the accelerated failure model and quantile regression models could also be considered.

Our new method is constructed so as to guarantee, in the absence of ties, that the sum of the CIF's across all event types at the last observed event time is equal to 1. By constrast, with the existing methods, the sum of the CIF's over all event types evaluated at the last event time can be less than or greater than 1. In an extensive simulation study, we showed that the deviation from 1 can be substantial under small sample size. The phenomenon tends not to appear when there is a high percentage of censoring, but here we have seen that for uncensored data the phenomenon can appear with Methods 1 and 2. Method 3 avoids this undesirable phenomenon. Our simulations showed further than the newly-proposed Method 3 exhibits performance comparable to that of the existing methods in terms of bias and variance. In terms of confidence interval coverage rates, we see with the existing methods that the coverage rate of the 95% confidence interval sometimes falls below 90%, while with the newly-proposed method this phenomenon

does not occur. The confidence interval coverage rates with Method 3 tended to be higher than the nominal 95%, with a coverage rate of 99% is some cases. In conclusion, with our newly-proposed estimator, the advantage of having the end-of-study total CIF equal to 1 is achieved without paying a price in terms of performance.

DATA AVAILABLITY STATEMENT

R code for performing the simulations and data analysis in this paper, along with the data used in the example, is posted on the following webpage:
`https://github.com/david-zucker/cumulative-incidence-function.git`

ACKNOWLEDGEMENTS

We thank the Associate Editor and two referees for their helpful comments that led to significant improvements in the paper.

FUNDING STATEMENT

There are no funders to report for this paper.

References

Aalen, O. O. and S. Johansen (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics 5*(3), 141–150.

Ajami, S., Y. Woodbridge, and D. G. Feitelson (2019). Syntax, predicates, idioms—what really affects code complexity? *Empirical Software Engineering 24*(1), 287–328.

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). *Statistical models based on counting processes*. Springer.

Beyersmann, J. and T. H. Scheike (2014). Classical regression models for competing risks. In J. P. Klein, H. C. van Houwelingen, J. G. Ibrahim, and T. H. Scheike (Eds.), *Handbook of survival analysis*, pp. 157–177. CRC Press, Boca Raton FL.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(2), 187–202.

Grambsch, P. M. and T. M. Therneau (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika 81*(3), 515–526.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The statistical analysis of failure time data, 2nd ed.* John Wiley & Sons.

Klein, J. P. and M. L. Moeschberger (2003). *Survival analysis: techniques for censored and truncated data, 2nd ed.* Springer.

Kosorok, M. R. and R. Song (2007). Inference under right censoring for transformation models with a change-point based on a covariate threshold. *The Annals of Statistics 35*(3), 957–989.

Putter, H., M. Fiocco, and R. B. Geskus (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine 26*(11), 2389–2430.

Schoop, R., J. Beyersmann, M. Schumacher, and H. Binder (2011). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal 53*(1), 88–112.

Table 1. Summary of simulation configurations under $Z \sim U(-0.5, 0.5)$.

Scenario	Hazard Type	Sample size	$\exp(\beta_j)$	z	Censoring rate
1	increasing	75	3	-0.4	0.0
2	increasing	150	3	-0.4	0.5
3	increasing	75	3	0.0	0.0
4	increasing	150	3	0.0	0.5
5	increasing	75	3	0.4	0.0
6	increasing	150	3	0.4	0.5
7	increasing	75	6	-0.4	0.0
8	increasing	150	6	-0.4	0.5
9	increasing	75	6	0.0	0.0
10	increasing	150	6	0.0	0.5
11	increasing	75	6	0.4	0.0
12	increasing	150	6	0.4	0.5
13	decreasing	75	3	-0.4	0.0
14	decreasing	150	3	-0.4	0.5
15	decreasing	75	3	0.0	0.0
16	decreasing	150	3	0.0	0.5
17	decreasing	75	3	0.4	0.0
18	decreasing	150	3	0.4	0.5
19	decreasing	75	6	-0.4	0.0
20	decreasing	150	6	-0.4	0.5
21	decreasing	75	6	0.0	0.0
22	decreasing	150	6	0.0	0.5
23	decreasing	75	6	0.4	0.0
24	decreasing	150	6	0.4	0.5
25	up-and-down	75	3	-0.4	0.0
26	up-and-down	150	3	-0.4	0.5
27	up-and-down	75	3	0.0	0.0
28	up-and-down	150	3	0.0	0.5
29	up-and-down	75	3	0.4	0.0
30	up-and-down	150	3	0.4	0.5
31	up-and-down	75	6	-0.4	0.0
32	up-and-down	150	6	-0.4	0.5
33	up-and-down	75	6	0.0	0.0
34	up-and-down	150	6	0.0	0.5
35	up-and-down	75	6	0.4	0.0
36	up-and-down	150	6	0.4	0.5

Table 2. Percentage of Simulations in Which Final Total CIF Exceeded 1, Uniformly Distributed Covariate With No Censoring

Scenario	Hazard Shape	Relative Risk	Covariate Value	Method 1	Method 2
1	increasing	3	−0.4	95.1	36.6
3	increasing	3	0.0	100	87.1
5	increasing	3	0.4	100	96.9
7	increasing	6	−0.4	93.1	44.3
9	increasing	6	0.0	100	97.0
11	increasing	6	0.4	100	99.8
13	decreasing	3	−0.4	95.0	34.1
15	decreasing	3	0.0	100	86.7
17	decreasing	3	0.4	100	96.3
19	decreasing	6	−0.4	90.0	37.5
21	decreasing	6	0.0	100	95.8
23	decreasing	6	0.4	100	99.5
25	up & down	3	−0.4	95.1	36.6
27	up & down	3	0.0	100	87.1
29	up & down	3	0.4	100	96.9
31	up & down	6	−0.4	92.9	43.2
33	up & down	6	0.0	100	96.9
35	up & down	6	0.4	100	99.8

Table 3. Percentage of Simulations in Which Final Total CIF Exceeded 1, Normally Distributed Covariate With No Censoring

Relative Risk	Covariate Value	Method 1	Method 2
3	−1.68	100	99.2
3	0.00	100	100
3	1.68	100	100
6	−1.68	98.8	90.5
6	0.00	100	100
6	1.68	100	100

Table 4. Summary of code snippets. All code segments check whether a number is in a set of non-overlapping ranges. If a code snippet is preceded by a number, it indicates the number of ranges.

Snippet	Description
as,bs,cs	Structure variants a,b,c - use if else with several conditions.
al,bl (b1l),cl	Logical variants a,b,c - use nested single condition if else.
an,an1,an2	Variants of snippet al that use negation.
f*	For loop that uses simple arithmetic manipulations of the for loop iteration variable.
f[]	For loop that uses pointer array for the ranges' limits.
lp0, . . . ,lp6	Special loops. The for loop statement is used differently from common practice. For example i counting down instead of up.

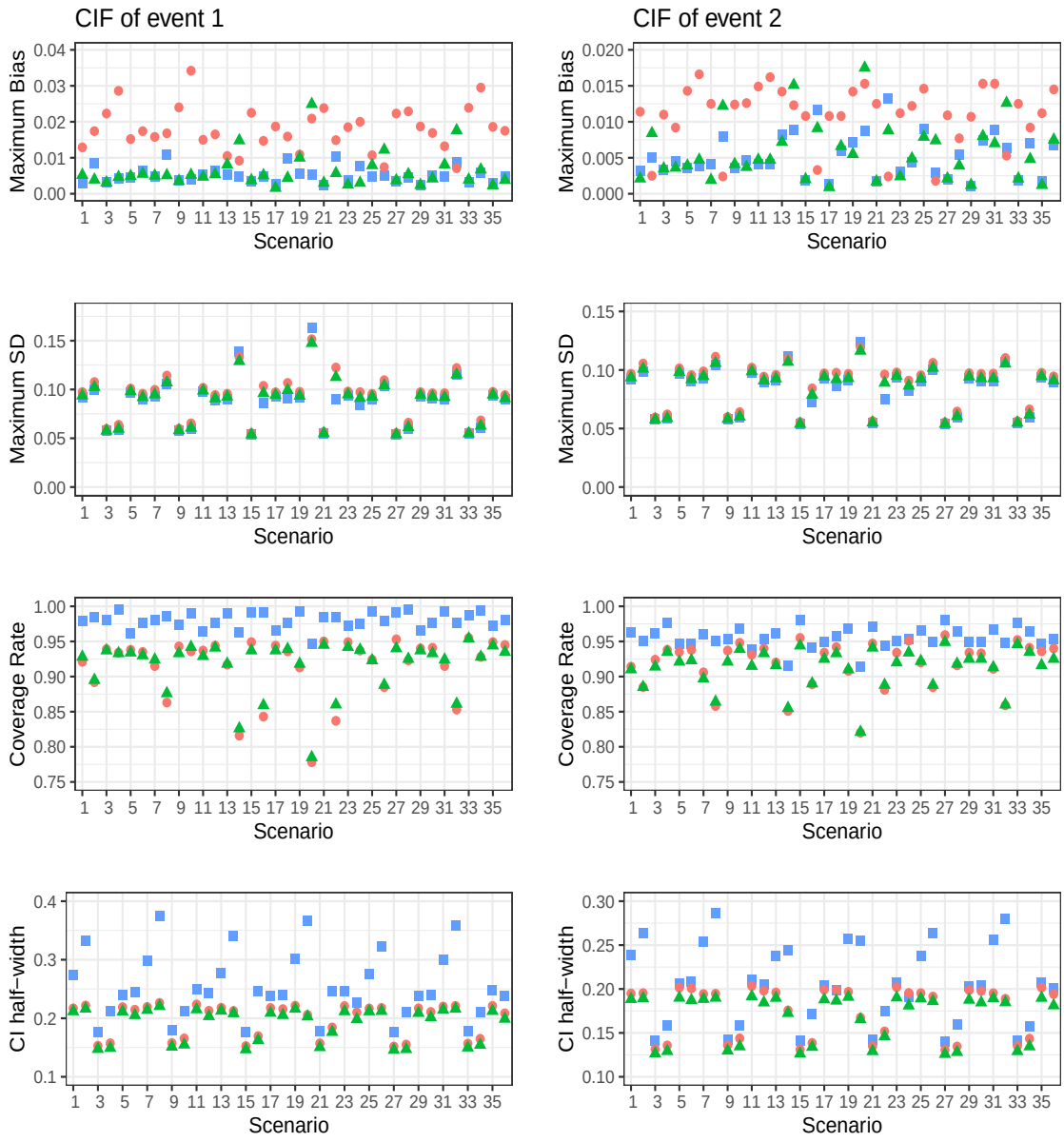

Fig. 1. Simulation results with uniformly distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square. Line 1 - the maximum mean bias of the CIF estimates up to the 90th percentile of the last observed event time; line 2 - the standard error of the estimates at the 90th percentile of the last observed event time (the standard error generally increased over time); line 3 - the empirical coverage rates of the 95% confidence bands; and line 4 - half width of the 95% confidence bands. See Table 1 for the scenarios' configurations.

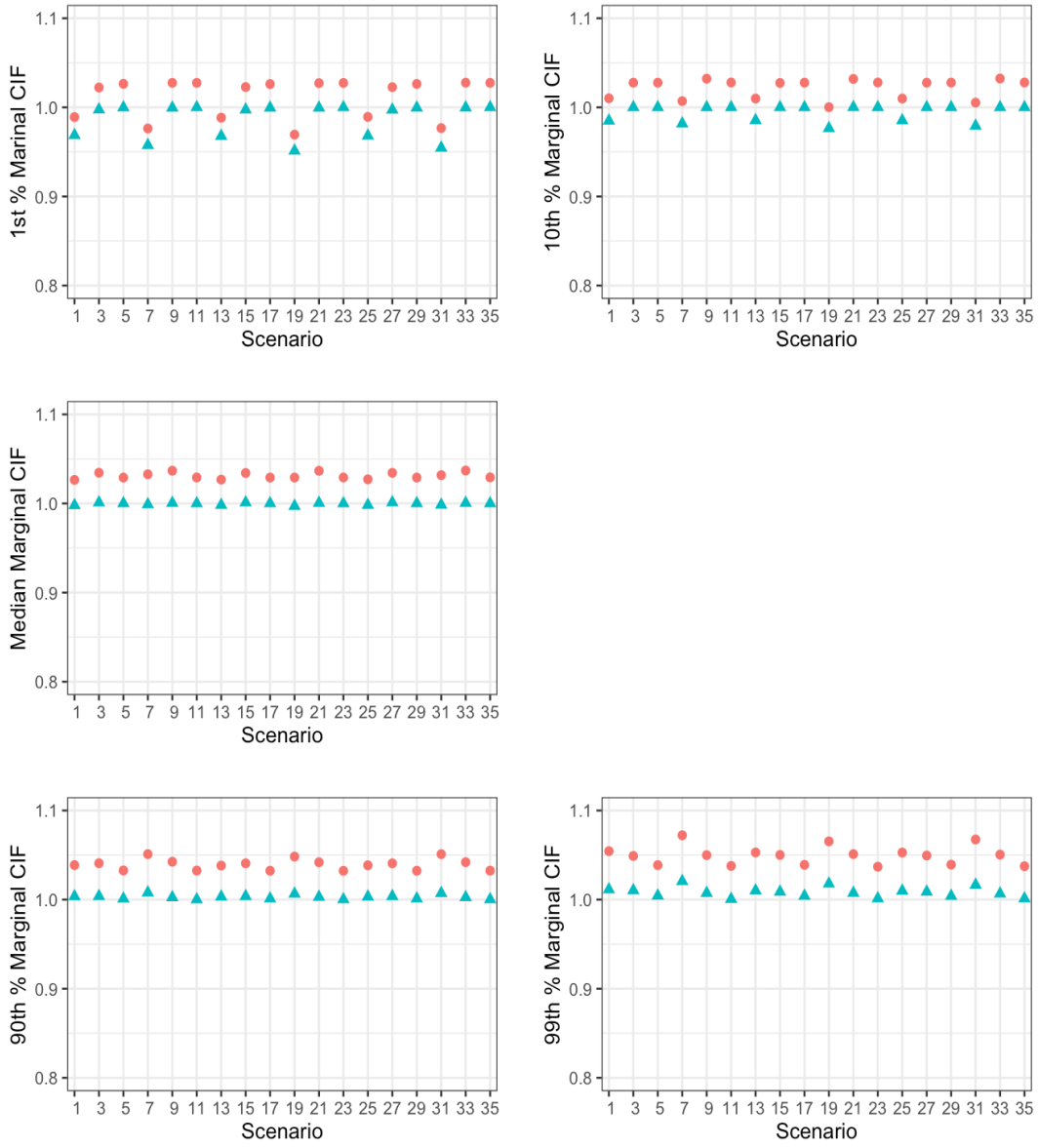

Fig. 2. Simulation results with uniformly distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle. Various quantiles (1%, 10%, 50%, 90%, 99%) of the marginal CIF (i.e. the CIF summed over the two risks) at the last event time, for the configurations without censoring. See Table 1 for the scenarios' configurations.

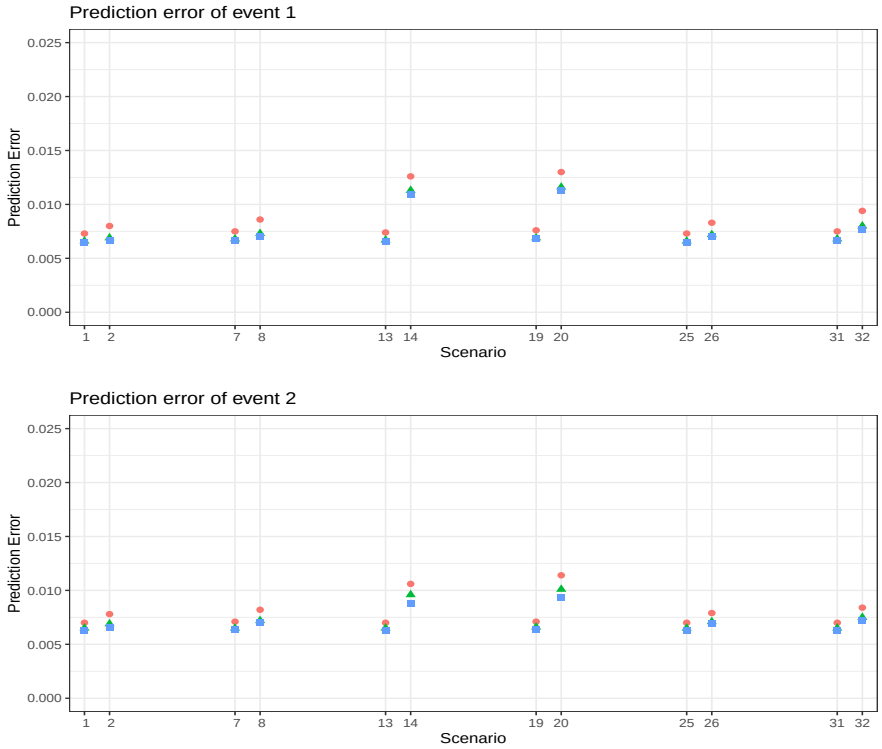

Fig. 3. Simulation results of prediction error (PE) with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square. For setting details of each scenario see Table 1.

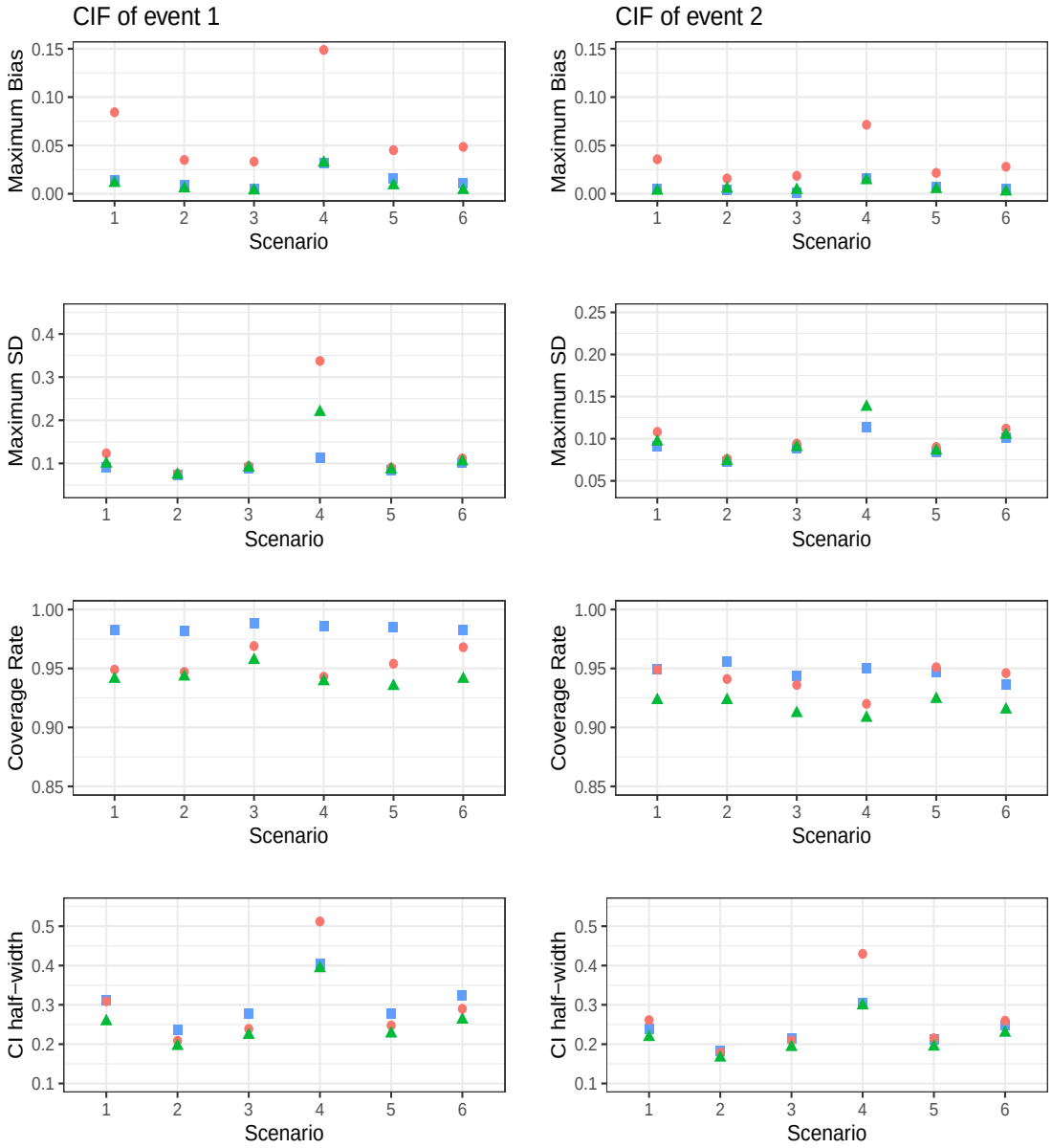

Fig. 4. Simulation results with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square. Line 1 - the maximum mean bias of the CIF estimates up to the 90th percentile of the last observed event time; line 2 - the standard error of the estimates at the 90th percentile of the last observed event time (the standard error generally increased over time); line 3 - the empirical coverage rates of the 95% confidence bands; and line 4 - half width of the 95% confidence bands. $n = 75$; no censoring; Scenarios 1 and 4 with $z = -1.68$.

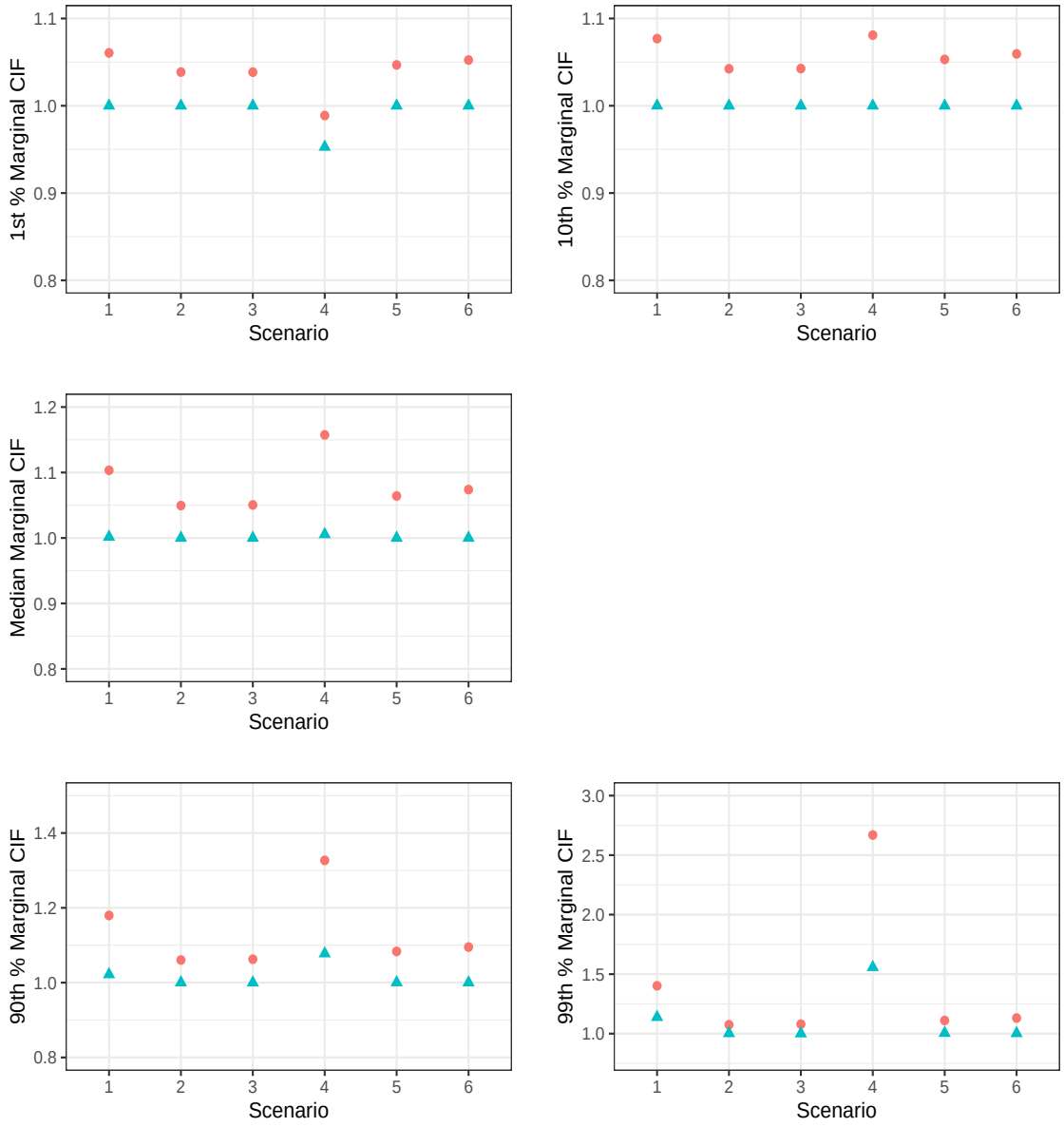

Fig. 5. Simulation results with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle. Various quantiles (1%, 10%, 50%, 90%, 99%) of the marginal CIF (i.e. the CIF summed over the two risks) at the last event time, for the configurations without censoring. $n = 75$; no censoring; Scenarios 1 and 4 with $z = -1.68$; Scenarios 2 and 5 with $z = 0$; Scenarios 3 and 6 with $z = 1.68$.

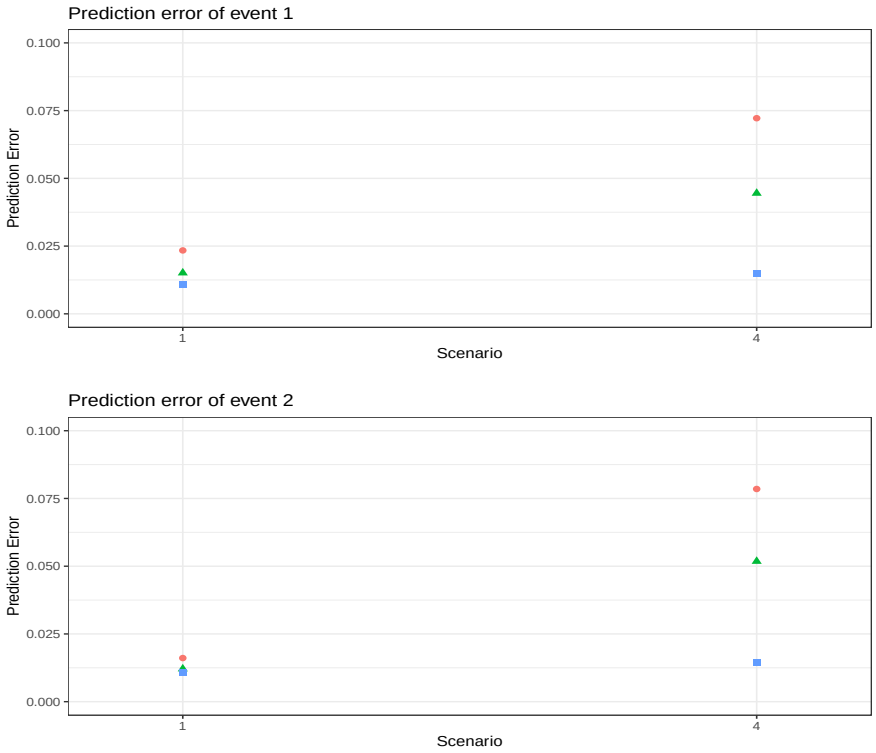

Fig. 6. Simulation results of prediction error (PE) with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square.

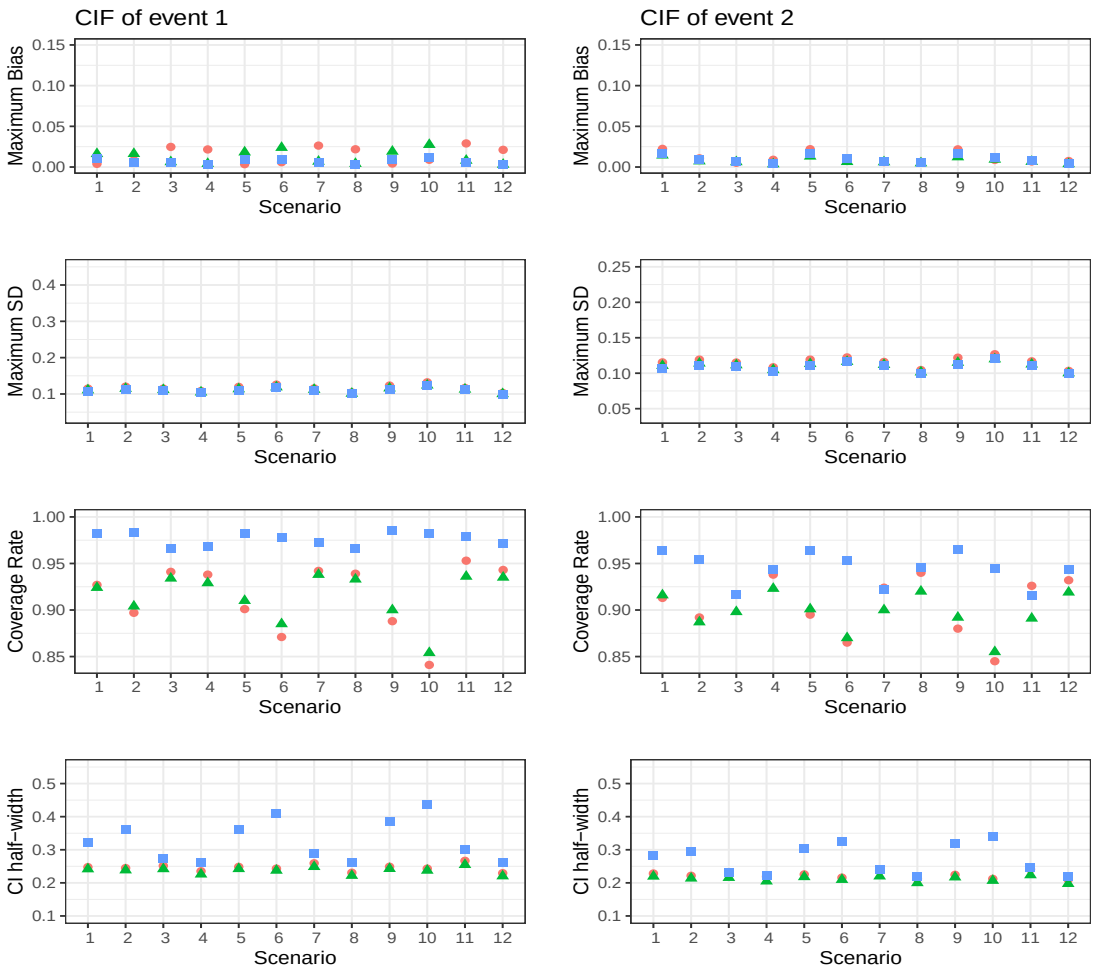

Fig. 7. Simulation results with two covariates covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square. Line 1 - the maximum mean bias of the CIF estimates up to the 90th percentile of the last observed event time; line 2 - the standard error of the estimates at the 90th percentile of the last observed event time (the standard error generally increased over time); line 3 - the empirical coverage rates of the 95% confidence bands; and line 4 - half width of the 95% confidence bands.

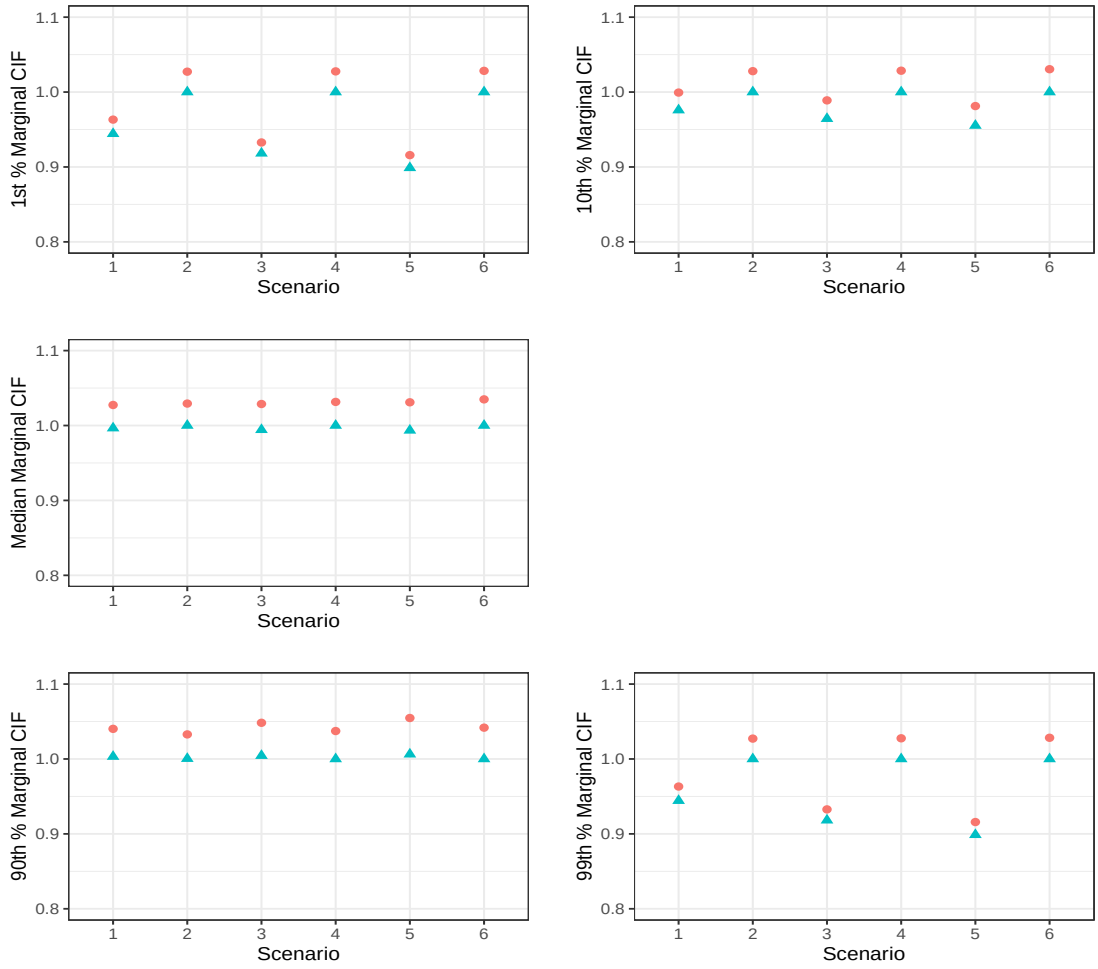

Fig. 8. Simulation results with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle. Various quantiles (1%, 10%, 50%, 90%, 99%) of the marginal CIF (i.e. the CIF summed over the two risks) at the last event time, for the configurations without censoring.

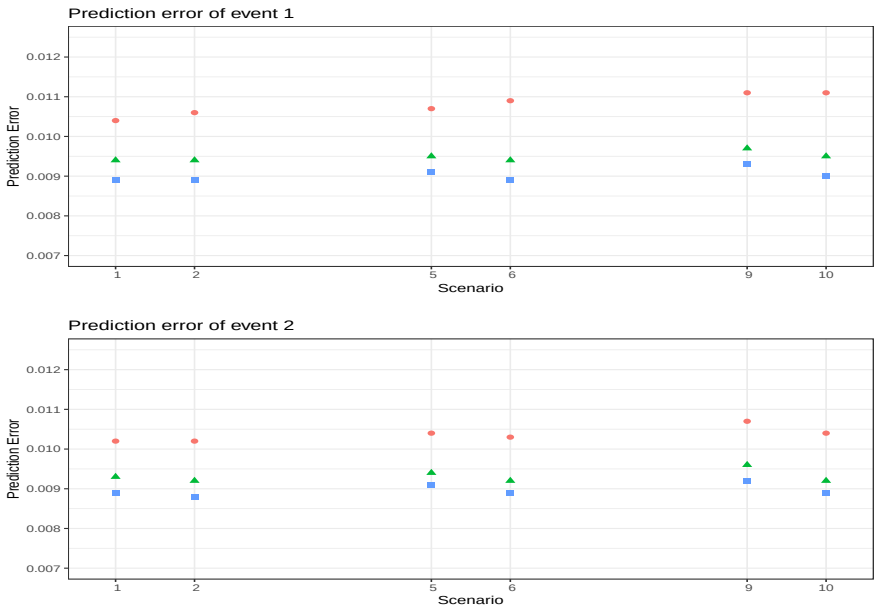

Fig. 9. Simulation results of prediction error (PE) with normally distributed covariate. $\hat{F}_j^{(1)}$ - red circle; $\hat{F}_j^{(2)}$ - green triangle; $\hat{F}_j^{(3)}$ - blue square.

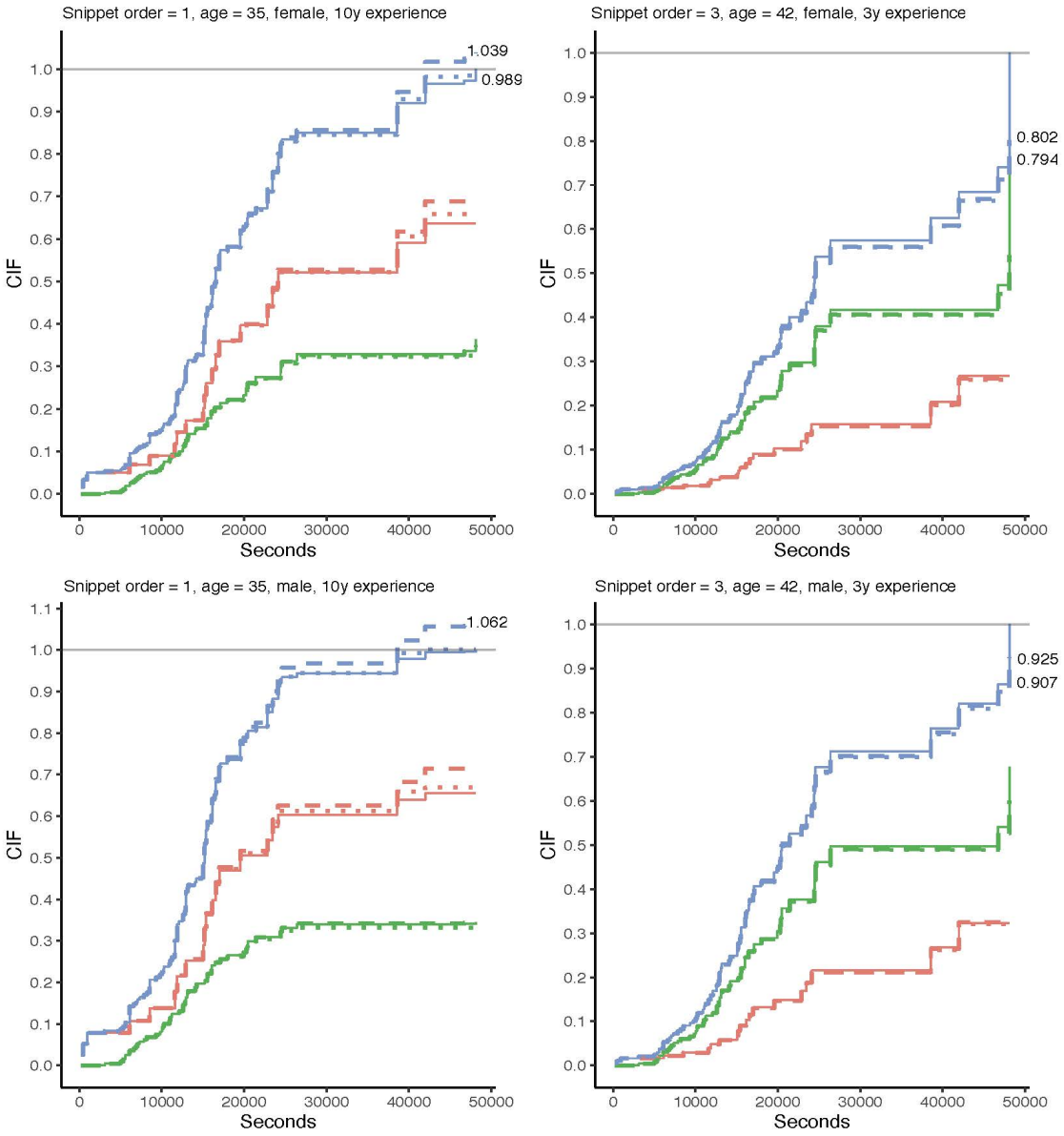

Fig. 10. Data Analysis, Snippet `lp3`, less than a year of programming experience. $\hat{F}_j^{(1)}$ - dashed line; $\hat{F}_j^{(2)}$ - dotted line; $\hat{F}_j^{(3)}$ - continuous line. Red - incorrect answer; green - correct answer, blue - marginal probability. The value of $\hat{F}_{\cdot}^{(m)}(T_{(K)}|\mathbf{z})$, $m = 1, 2$, are presented whenever deviates from 1.

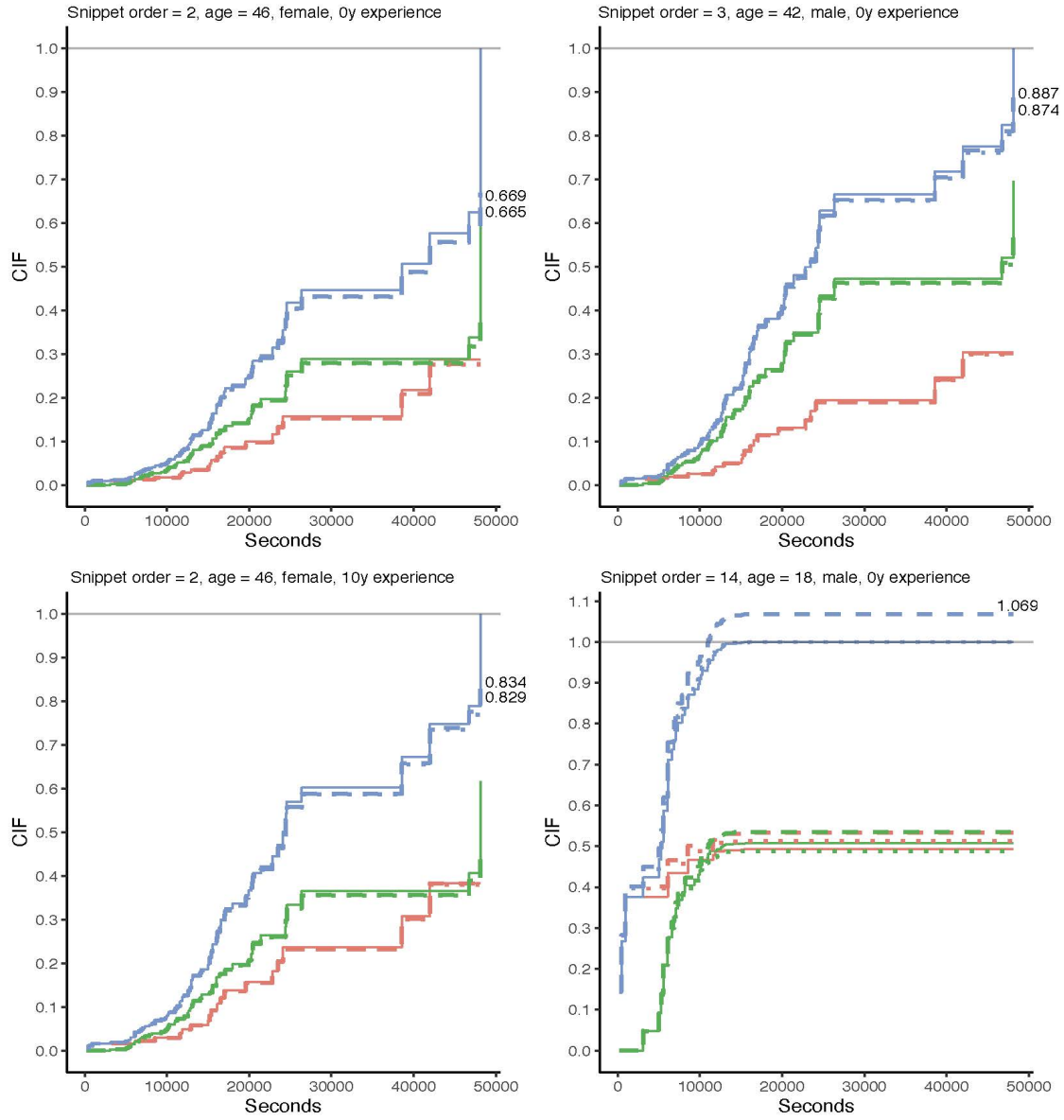

Fig. 11. Data Analysis, Snippet `lp3`, 5 years programming experience. $\hat{F}_j^{(1)}$ - dashed line; $\hat{F}_j^{(2)}$ - dotted line; $\hat{F}_j^{(3)}$ - continuous line. Red - incorrect answer; green - correct answer, blue - marginal probability. The value of $\hat{F}_{\cdot}^{(m)}(T_{(K)}|\mathbf{z})$, $m = 1, 2$, are presented whenever deviates from 1.

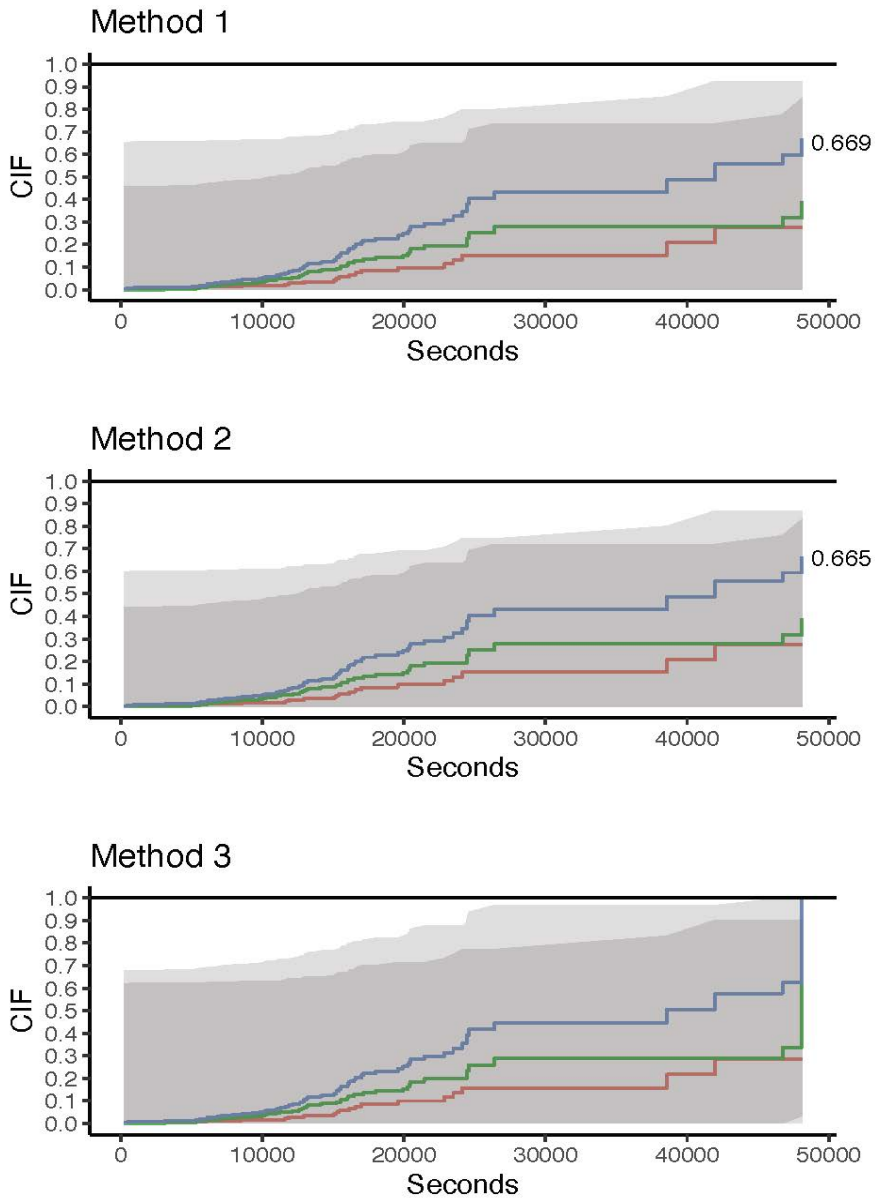

Fig. 12. Data Analysis, Snippet `lp3`, $\mathbf{z} = (\text{snippet order} = 1, \text{age} = 35, \text{female}, \text{YoE} = 5)$. Method 1 - $\hat{F}_j^{(1)}$; Method 2 - $\hat{F}_j^{(2)}$; Method 3 - $\hat{F}_j^{(3)}$. Red - incorrect answer; green - correct answer, blue - marginal probability. The value of $\hat{F}_{\cdot}^{(m)}(T_{(K)}|\mathbf{z})$, $m = 1, 2$, are presented whenever deviates from 1.