# Chapter 7

# Simulation Study

## 7.1 Introduction

In this chapter we present a simulation study in order to illustrate, for the case of finite samples, the most significative results for the estimator proposed in Chapter 6. This study allows us to compare the proposed new methodology with the naive complete case analysis. Basic ideas about simulation techniques in statistics come from Bratley, Fox and Schrage (1987).

The structure of the chapter is as follows: In Section 7.2 we introduce all the elements needed for the simulation (*e.g.,* the reference distributions, the different proportion of censoring considered, the non-response patterns considered, ...) in order to obtain the configuration of all possible scenarios. In Section 7.3 we present the implementation and the algorithm for the simulation. The results are analyzed in detail in Section 7.4 and discussed in Section 7.5.

## 7.2 Design of the simulation

**Reference distributions:** We consider two populations, say, $X = 0$, with a decreasing hazard function ($\sigma > 1$) and Weibull(6.7, 1.4) distributed, and $X = 1$, with an increasing hazard function ($\sigma < 1$) and Weibull(7.7, 0.8) distributed. The choice of these laws and their parameters has been based on our HIV+PTB cohort.

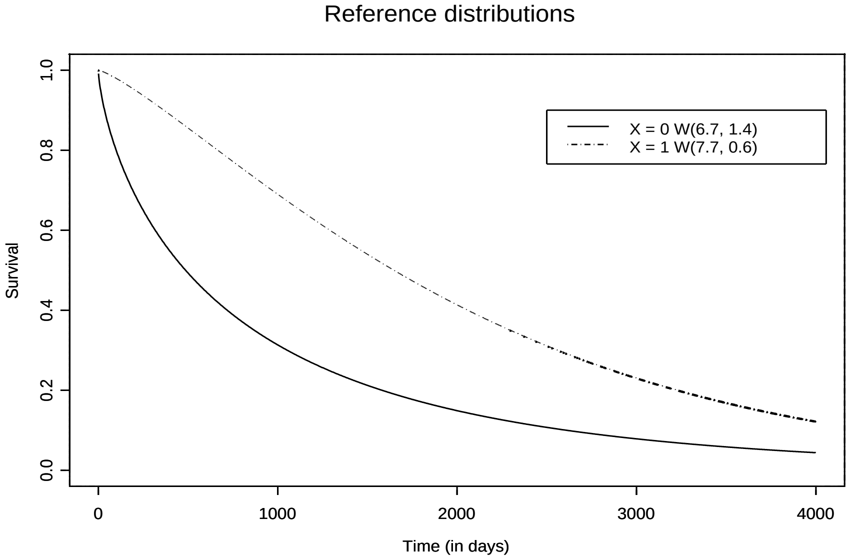

Figure 7.1: *Reference survival functions for the simulation*

Table 7.1 summarizes the survival function measured at different times and Figure 7.1 illustrates the difference between groups across time.

| | $P(T > t\|X)$ ($t$ in years) | | | | | | |
|---|---|---|---|---|---|---|---|
| Group | $t = 1$ | $t = 2$ | $t = 3$ | $t = 5$ | $t = 8$ | $t = 10$ | $t = 15$ |
| $X = 0$ | 0.569 | 0.396 | 0.290 | 0.168 | 0.083 | 0.054 | 0.020 |
| $X = 1$ | 0.900 | 0.778 | 0.660 | 0.455 | 0.242 | 0.154 | 0.045 |

Table 7.1: *True survival at different times (in years) for the reference distributions in each category of the covariate $X$*

**Proportion of censoring:** In order to simulate two different levels of censoring, we will consider that we observe the individuals in a window time $(0, T_{max}]$, where $T_{max} = 3{\cdot}365 = 1095$ days (*i.e.,* three years) or $T_{max} = 10{\cdot}365 = 3650$ days (*i.e.,* ten years). We will take the censoring mechanism as an uniform random variable in $(0, T_{max}]$ (*i.e.,* we will suppose that individuals can enter in -or equivalently leave- the study in any moment before $T_{max}$). This choice corresponds also to the idea that the accrual is uniform across time.

The proportion of censoring depends on the proportion of individuals in everyone of the categories in the covariate X (*i.e.,* depends on $P(X = 1)$). This probability is also unknown in a real analysis of a dataset with missing values in $X$. So, it is necessary to take into account this parameter in order to show the properties of the proposed estimator in the resulting scenarios.

Since the censoring time and the survival time are independent, the proportion of censoring can be computed by the expression

$$P(T > C) = P(T > T_{max}) + \int_0^{T_{max}} P(C < x) f_T(x) dx$$

for each of the categories in the covariate $X$.

Due to the symmetry it is enough, for example, to setup $P(X = 1)$ to 0.3 or 0.5. Table 7.2 summarizes the proportion of censoring in each scenario.

| | Censoring | | Pooled censoring | |
|---|---|---|---|---|
| $T_{max}$ (in years) | $X = 0$ | $X = 1$ | $P(X = 1) = 0.3$ | $P(X = 1) = 0.5$ |
| 3 | 0.5137 | 0.8375 | 0.6109 | 0.6756 |
| 10 | 0.2478 | 0.5027 | 0.3242 | 0.3752 |
| 30 | 0.0918 | 0.1878 | 0.1206 | 0.1398 |

Table 7.2: *Proportion of censoring for different values of* $P(X = 1)$ *and different observation windows* $(0, T_{max}]$

$\tau_k$ **partition:** The EGKM estimator depends, by definition, on the partition $0 < \tau_1 < \tau_2 < \ldots < \tau_K < T_{max}$ in $(0, T_{max})$. Since our times will be measured in days, we will perform the simulation considering the following three partitions: in years, in months and in weeks (the coarsest one, the medium and the thinest one, respectively).

In each interval we will have to estimate the number of deaths and the number of censored individuals for each category of the covariate $X$. This means that the estimation will be less precise in those intervals with a lower number of individuals. If we consider a small sample size (*e.g.,* $n = 50$ o lower) the expected number of individuals in one category in each interval will be really low. For example, if we use the partition in weeks when $T_{max} = 3$ years then we define more than 150 intervals. So, in order not to waste too much information, we will consider three differents sample sizes for each of the censoring levels: if $T_{max} = 3$ years, then we take $n = 100$, 500 or 1000 and, if $T_{max} = 10$ years, then we take $n = 200$, 1000 or 2000.

**Non-response pattern:** We will generate the data according to the following non-response pattern

$$\text{logit } P(R = 1|Y, \delta, X) = \alpha_0 + \alpha_1 \cdot Y + \alpha_2 \cdot \delta + \tau \cdot X, \tag{7.1}$$

and we will setup parameters $\alpha_0, \alpha_1, \alpha_2$ and $\tau$ as in Table 7.3 is shown. It allows us to simulate the missing completely at random (MCAR), missing at random (MAR) and non-ignorable with non-ignorability parameter $\tau$ (NI($\tau$)) patterns.

| Model | $\alpha_0$ | $\alpha_1$ | $\alpha_2$ | $\tau$ |
|---|---|---|---|---|
| MCAR | 1 | 0 | 0 | 0 |
| MAR | -0.75 | 0.005 | 1 | 0 |
| NI(-2) | -0.75 | 0.005 | 1 | -2 |
| NI(2) | -0.75 | 0.005 | 1 | 2 |

Table 7.3: *Setup of parameters $\alpha_0, \alpha_1, \alpha_2$ and $\tau$ for each non-response pattern model*

The expected proportion of missing data can be computed by the expression

$$\begin{aligned} P(R = 0) &= \sum_{x=0,1} P(X = x) P(R = 0|X = x) \\ &= \sum_{x=0,1} P(X = x) \int_0^{T_{max}} P(R = 0|X = x, Y, \delta) dF_{Y,\delta|X=x} \end{aligned}$$

where $dF_{Y,\delta|X=x} = \left\{ I(\delta = 0) \frac{1}{T_{max}} S_{X=x}(y) + I(\delta = 1) f_{X=x}(y) \left(1 - \frac{y}{T_{max}}\right) \right\} dy$ and $f_{X=x}$ and $S_{X=x}$ are, respectively, the density and the survival function for the group $X = x$.

In Table 7.4 we present the expected proportion of missing data in each of the categories for every scenario.

| | | $P(R=0\|X=x)$ | | $P(R=0)$ | |
|---|---|---|---|---|---|
| $T_{max}$ | NR Model | $X=0$ | $X=1$ | $P(X=1)=0.3$ | $P(X=1)=0.5$ |
| 3 | MCAR | 0.2689 | 0.2689 | 0.2689 | 0.2689 |
| | MAR | 0.2636 | 0.2156 | 0.2492 | 0.2396 |
| | NI(-2) | 0.2636 | 0.5265 | 0.3425 | 0.3951 |
| | NI(2) | 0.2636 | 0.0474 | 0.1988 | 0.1555 |
| 10 | MCAR | 0.2689 | 0.2689 | 0.2689 | 0.2689 |
| | MAR | 0.1766 | 0.0868 | 0.1497 | 0.1317 |
| | NI(-2) | 0.1766 | 0.2372 | 0.1948 | 0.2069 |
| | NI(2) | 0.1766 | 0.0180 | 0.1290 | 0.0973 |

Table 7.4: *Proportion of missing data for different values of* $P(X=1)$ *and different observation windows* $(0, T_{max}]$, *for each non-response pattern*

Each set of data will be analyzed using the complete case methodology (CC) and also taking into account up to six different non-response patterns. In the CC analysis we will only consider those individuals with observed covariate value and we will apply the Kaplan-Meier estimator. For the other methodologies, we will semiparametrically estimate the survival function by using the EGKM estimator proposed in the previous section. Table 7.5 shows all the methods used in the analysis and the parameters in the non-response model (7.1) that they have to be estimated.

In fact, in order to prevent unlikelihood scenarios, we will analyze the data using models not too far from the real one that generated the data. In others words, we will assume that the analyst has a correct "common sense" about the non-response mechanism. Table 7.6 illustrates the 18 combined generating/analyzing non-response patterns.

**Possible scenarios:** Table 7.7 summarizes all the possible scenarios we will analyze. Since the CC analysis is independent on the grid option, we can configure $2 \cdot 3 \cdot 2 \cdot 4 + 2 \cdot 3 \cdot 3 \cdot 2 \cdot 14 = 48 + 504 = 552$ different scenarios to simulate. First, we will simulate the 276 cases corresponding to the less suitable scenarios (*i.e.,* when the

| NR Pattern | Presetup parameters | Parameters to estimate |
|---|---|---|
| CC | Complete Case analysis | none |
| MCAR | $\alpha_1 = \alpha_2 = 0$ and $\tau = 0$ | $\alpha_0$ |
| MAR | $\tau = 0$ | $\alpha_0, \alpha_1$ and $\alpha_2$ |
| NI(-2), NI(-1) | $\tau = -2, -1$ | $\alpha_0, \alpha_1$ and $\alpha_2$ |
| NI(1), NI(2) | $\tau = 1, 2$ | $\alpha_0, \alpha_1$ and $\alpha_2$ |

Table 7.5: *Non-response patterns used in the analysis of the simulated data*

| Generating | Analyzing pattern | | | | | | |
|---|---|---|---|---|---|---|---|
| pattern | CC | MCAR | MAR | NI(-2) | NI(-1) | NI(1) | NI(2) |
| MCAR | ● | ● | ● | ○ | ○ | ○ | ○ |
| MAR | ● | ● | ● | ○ | ● | ● | ○ |
| NI(-2) | ● | ● | ● | ● | ● | ○ | ○ |
| NI(2) | ● | ● | ● | ○ | ○ | ● | ● |

Table 7.6: *Generating vs analyzing non-response pattern used in the simulation study*

proportion of censoring is big, that is, when $T_{max} = 3$ years). Later, we will simulate those scenarios in the case $T_{max} = 10$ years that they can help us to illustrate the properties of the proposed EGKM estimator.

**Monte Carlo study:** We will conduct every simulation experiment based on 250 realizations. However, when the goal is to estimate the coverage probabilities, for a subset of the scenarios, we will extend the number of realizations to 1000. So, the estimated coverage probability of a true 95% confidence interval will have a simulation accuracy of approximately 1.35% ($1.96\sqrt{0.05 \cdot 0.95/1000} = 0.0135$).

At each realization, we will estimate the survival function, and its standard error, at 1 and 2 years when $T_{max} = 3$ years (or at 1, 2, 5 and 8 years when $T_{max} = 10$ years) for each category in $X$. Then, we will evaluate the absolute and the relative bias, and whether the true survivorship belongs to the respective nominal 95% confidence

| Configuration<br>parameter | Values | Number<br>of options |
|---|---|---|
| Observation window ($T_{max}$) | 3 or 10 years | 2 |
| Grid ($\tau_k - \tau_{k-1}$) | year / month / week | 3 |
| Sample size ($n$) | 100/500/1000 or 200/1000/2000 | 3 |
| $X = 1$ proportion ($P(X = 1)$) | 0.3 or 0.5 | 2 |
| Generating vs Analizing<br>non-response pattern | see Table 7.6 | 18 |

Table 7.7: *Configuration of the scenarios for the simulation study*

interval (coverage indicator).

Among all the Monte Carlo iterations, we report the Monte Carlo mean, the shortest half location parameter for the standard error estimates, the coverage probability of the nominal 95% large sample confidence intervals, the Monte Carlo bias (absolute and relative), the Monte Carlo standard error of the simulation and the mean squared error at each of the mentioned times for each category. For summarizing the standard error estimates we use the shortest half location parameter (Rousseeuw and Leroy, 1987), instead of the mean, because it is more robust in front of outliers samples and non-simetric distributions. All the results are given in Appendix V.

According to (6.11), for each of the reference distributions in the simulation, we can compute the upper bound for the relative bias of the Grouped Kaplan-Meier estimator in each of the mentioned times of estimation and for each scenario given in Table 7.7. This upper bound depends only on the observation window (*i.e.,* on $T_{max}$) and on the grid size (*i.e.,* on the partition $0 < \tau_1 < \ldots < \tau_K < T_{max}$). Since higher order terms in (6.11) are, in general, lower than $10^{-4}$, Table 7.8 only illustrates the linear term in the Taylor's expansion in the computation of these upper bounds. We can see that the relative bias depends on the true percentile we are estimating (in others words, the relative bias is lower if the true survival is higher). On the other hand, we can also observe that the relative bias is quite small if we consider a reasonable partition. For example, if we work in weeks, the relative bias for the distribution W(6.7, 1.4) is lower than 1% for the estimation of the 90% quantile (see Table 7.1 and Figure 7.1).

| | | | Upper bound at $t$ years (in %) | | | |
|---|---|---|---|---|---|---|
| Distr. | $T_{max}$ | Grid in | $t = 1$ | $t = 2$ | $t = 5$ | $t = 8$ |
| W(6.7,1.4) | 3 | years | 25.30 | 47.09 | – | – |
| | | months | 1.82 | 3.78 | – | – |
| | | weeks | 0.43 | 0.90 | – | – |
| | 10 | years | 7.59 | 12.43 | 26.66 | 47.44 |
| | | months | 0.49 | 0.84 | 1.92 | 3.61 |
| | | weeks | 0.11 | 0.20 | 0.45 | 0.86 |
| W(7.7,0.8) | 3 | years | 3.70 | 11.52 | – | – |
| | | months | 0.35 | 1.14 | – | – |
| | | weeks | 0.08 | 0.28 | – | – |
| | 10 | years | 1.11 | 2.85 | 11.44 | 29.88 |
| | | months | 0.09 | 0.23 | 0.91 | 2.42 |
| | | weeks | 0.02 | 0.05 | 0.22 | 0.58 |

Table 7.8: *Approximate upper bound (in percentage) for the relative bias for the Grouped Kaplan-Meier estimator for the two reference distributions in the Monte Carlo simulations as a function of the observation window* $(0, T_{max}]$ *and the grid size*

**Comparison between methodologies:** In order to compare the EGKM methodology and the complete case methodology we estimate the asymptotic relative efficiency, $ARE_1$, as

$$ARE_1 = \frac{1/MSE_{\text{alternative meth.}}}{1/MSE_{\text{CC analysis}}} = \frac{MSE_{CC}}{MSE_{a.m.}}.$$

To compare a different non-response pattern used in the analysis with the correct one, we will use the asymptotic relative efficiency between both, $ARE_2$, as:

$$ARE_2 = \frac{1/MSE_{\text{alternative non-resp. pat.}}}{1/MSE_{\text{correct non-resp. pat.}}} = \frac{MSE_{c.p.}}{MSE_{a.p.}}.$$

## 7.3 Implementation of the simulation

All the processes have been implemented in S-PLUS and the respective functions are listed and commented in Appendix IV.

To obtain each realization of the simulation we do as follows:

**Step 1.** Setup the scenario: $T_{max}$, grid, $n$, $p = P(X = 1)$, NR pattern for the generation, NR pattern for the analysis

**Step 2.** Generate the true data: $X \sim B(p)$, $T \sim W(6.7 + X, 1.4 - 0.6 \cdot X)$ and $C \sim U([0, T_{max}])$

**Step 3.** Generate the observed survival times: $Y = \min\{T, C\}$

**Step 4.** Generate the observed covariates: $R \sim B(\text{expit}(\alpha_0 + \alpha_1 \cdot Y + \alpha_2 \cdot \delta + \tau \cdot X))$ and redefine $X$ as $X \cdot R +$ "NA" $\cdot (1 - R)$

**Step 5.** Describe the sample: proportion of $X = 1$ in the initial sample, proportion of censoring, proportion of missing

If we are doing the Complete Case methodology:

**Step 6a.** Obtain the subsample of individuals with observed covariate

**Step 7a.** Compute the stratified Kaplan-Meier estimator, their standard errors and the coverage indicators: $\widetilde{\boldsymbol{S}}_{X=0}$ and $\widetilde{\boldsymbol{S}}_{X=1}$

otherwise

**Step 6b.** Estimate semiparametrically, using the methodology proposed in Section 6.3 with $\phi^{(2)}_{r=0}(Y,\delta) = (1, Y, Y^2)'$, the non-response parameter: $\widehat{\boldsymbol{\alpha}} = (\widehat{\alpha_0}, \widehat{\alpha_1}, \widehat{\alpha_2})'$

**Step 7b.** Estimate semiparametrically, using our methodology proposed in Section 6.3, the vector $\beta$ of number of deaths or censored individuals in each of the categories, and derive the Estimated Grouped Kaplan-Meier estimator, its standard error and the coverage indicator, in each category: $\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{S}}_{X=0}$ and $\widetilde{\boldsymbol{S}}_{X=1}$

and, in all the situations

**Step 8.** Compute the effective size of the sample and the effective proportion of individuals with $X = 1$

After doing all the 250 realizations in each scenario, we summarize the scenario, evaluating the following Monte Carlo parameters:

A) proportion of real data with $X = 1$, proportion of censoring, proportion of missing, effective size and effective proportion of $X = 1$

B) survival, median of the estimations of the standard error, coverage probability, absolute and relative bias, standard error of the simulation and mean square error, at each of the times of interest and for each category.

## 7.4 Results

The simulation has been developed in S-PLUS and performed in a Pentium II, 400 MHz, 64 MB RAM under Windows 98 operating system. In order to provide an idea about how much time is necessary to obtain the estimated survival funtions and confidence bands for a real univariate data set, we show in Table 7.9 the time needed for each iteration. Big part of the time is devoted to the computations involved in the estimation of the standard error from the sample; so, in the simulation, for computational time reasons, we only estimate the standard error (and the respective coverage indicator) for some of the most significative scenarios. Otherwise, this value is estimated by the Monte Carlo standard error of the simulation in each scenario.

| | $T_{max} = 3$ years | | | $T_{max} = 10$ years | | |
|---|---|---|---|---|---|---|
| Grid in | $n = 100$ | $n = 500$ | $n = 1000$ | $n = 200$ | $n = 1000$ | $n = 2000$ |
| years | 9 | 36 | 70 | 18 | 82 | 167 |
| months | 11 | 41 | 78 | 43 | 174 | 343 |
| weeks | 29 | 94 | 180 | 1071 | 2244 | 4810 |

Table 7.9: *Time in seconds for computing one iteration in each scenario for the analysis of a simulated data set with a non-ignorable generating and analyzing non-response pattern*

All the numerical results are listed in Appendix V. In this section we present the most significative results for the new methodology and those that allow us to assess some properties about the proposed estimator (for finite samples). In particular, we show all the results for the less informative situation, *i.e.,* when the censoring proportion is higher (in our case when $T_{max} = 3$ years), as well as some of the most conclusive ones for $T_{max} = 10$ years. For ease of exposition we will illustrate some of our results referring to specific pages in Appendix V (*e.g.,* [V, pp. 218–222]).

About the **parameters under control by design**, we obtain:

a) the Monte Carlo proportion of cases with real covariate $X$ equals 1 is the expected: the simulated proportion is 0.300 or 0.301, depending on the scenario, when $P(X = 1) = 0.3$ and 0.499 when $P(X = 1) = 0.5$, for all the scenarios,

b) something similar happens referring to the simulated proportion of censoring: The Monte Carlo proportion of censoring for $T_{max} = 3$ years stays in $[0.611, 0.612]$ if $P(X = 1) = 0.3$ and $[0.674, 0.676]$ if $P(X = 1) = 0.5$, which agree with the expected values by design shown in Table 7.2,

c) the simulated proportion of missing is also the designed in Table 7.4. Next table summarizes the Monte Carlo proportion of missing data for $T_{max} = 3$ years, different values of $P(X = 1)$, sample sizes and non-response patterns.

| $P(X = 1)$ | NR Model | Theoretical | $n = 100$ | $n = 500$ | $n = 1000$ |
|---|---|---|---|---|---|
| 0.3 | MCAR | 0.2689 | 0.275 | 0.268 | 0.268 |
| | MAR | 0.2492 | 0.250 | 0.248 | 0.249 |
| | NI(-2) | 0.3425 | 0.342 | 0.343 | 0.342 |
| | NI(2) | 0.1988 | 0.201 | 0.197 | 0.199 |
| 0.5 | MCAR | 0.2689 | 0.275 | 0.268 | 0.268 |
| | MAR | 0.2396 | 0.240 | 0.238 | 0.240 |
| | NI(-2) | 0.3951 | 0.393 | 0.395 | 0.395 |
| | NI(2) | 0.1555 | 0.157 | 0.155 | 0.156 |

Table 7.10: *Monte Carlo proportion of missing data for $T_{max} = 3$ years, different values of $P(X = 1)$ and different sample size $n$, for each non-response pattern*

Similar comments can be derived for $T_{max} = 10$ years scenarios.

About the **effective size of the sample**, the proposed methodology allows us to use the information of some of the individuals with missing covariate. More precisely, we can use the information of those individuals whose observed survival time, $Y$, shares the interval $(\tau_{k-1}, \tau_k]$ with other individuals with observed covariate. It means that the effective sample size will be larger if the grid size is also larger. Although, as we have shown in Table 7.8, for a coarse grid we can obtain a larger bias.

Table 7.11 shows the relative effective sample size of our methodology with respect to the complete case analysis, for $T_{max} = 3$ years and 10 years and $P(X = 1) = 0.3$. We can see that, if the grid is not too much coarse (in which situation we can use up to all the subjects in the sample, no matter which the sample size is), the improvement is increasing with the sample size. Analogously, for a fixed sample size and non-response model, it goes down if the grid gets thinner. It is interesting to note that for a sample size of $n = 1000$ the reduction is quite insignificant. We can also observe that the gain is higher when the proportion of missing is higher, and lower in the reverse sense. As an illustration, we can compare the values for the non-

| | | $T_{max} = 3$ years | $T_{max} = 10$ years |
|---|---|---|---|
| Grid in | NR Model | $n = 100$ / 500 / 1000 | $n = 200$ / 1000 / 2000 |
| years | MCAR | 1.38 / 1.37 / 1.37 | 1.36 / 1.37 / 1.37 |
| | MAR | 1.33 / 1.33 / 1.33 | 1.18 / 1.18 / 1.18 |
| | NI(-2) | 1.52 / 1.52 / 1.52 | 1.24 / 1.24 / 1.24 |
| | NI(2) | 1.25 / 1.25 / 1.25 | 1.15 / 1.15 / 1.15 |
| months | MCAR | 1.30 / 1.36 / 1.37 | 1.25 / 1.35 / 1.36 |
| | MAR | 1.25 / 1.33 / 1.33 | 1.15 / 1.17 / 1.18 |
| | NI(-2) | 1.35 / 1.51 / 1.52 | 1.18 / 1.23 / 1.24 |
| | NI(2) | 1.20 / 1.25 / 1.25 | 1.13 / 1.15 / 1.15 |
| weeks † | MCAR | 1.16 / 1.30 / 1.35 | — |
| | MAR | 1.10 / 1.26 / 1.31 | — |
| | NI(-2) | 1.14 / 1.38 / 1.47 | — |
| | NI(2) | 1.09 / 1.21 / 1.24 | — |

Table 7.11: *Monte Carlo relative effective sample size of the proposed methodology versus the complete case analysis, for* $T_{max} = 3$ *years and 10 years and* $P(X = 1) = 0.3$*, as a function of the grid, the non-response pattern and the sample size* $(n)$*.*
† *Results for* $T_{max} = 10$ *years and grid in weeks are not available*

ignorable scenarios with respect to the missing at random ones. For computational time reasons, results for $T_{max} = 10$ years and grid in weeks are not available. On the other hand, the corresponding table for $P(X = 1) = 0.5$, due to the proportion of missing, gives the same values for the MCAR scenario, quite the same for the MAR scenario, slightly higher for the NI(-2) scenario and slightly lower for the NI(2).

With respect to the **effective proportion of individuals with** $X = 1$, Table 7.12 shows the Monte Carlo estimates, for the scenarios with $T_{max} = 3$ years and 10 years, grid in months and sample size $n = 500$ or $n = 1000$, respectively; the table is organized as a function of the generating and analyzing non-response pattern and the true value of $P(X = 1)$ in the simulation. We observe that our settings for the MCAR and MAR analysis (see Table 7.3) provide quite the same estimators (the difference is lower than $10^{-3}$). This is due to the fact that when the $\tau_k$ partition is very thin (*e.g.,* grid in months or weeks) then the respective non-response models in (7.1) become very similar [V, pp. 218–222].

When the non-response pattern that generates de data is non-ignorable, we can see that the complete case methodology always presents a bias with the same sign of the non-ignorability parameter $\tau$. In fact, it underestimates the true proportion of $X = 1$ when $\tau < 0$, and it overestimates it when $\tau > 0$. This is because these individuals, in proportion, are underrepresented (if $\tau < 0$) in the complete case subsample versus those individuals with $X = 0$. In these scenarios, the proposed new methodology equilibrates these deficiencies and it provides a more adjusted estimated effective proportion (higher if $\tau < 0$, or lower, when $\tau > 0$). When $T_{max}$ =3 years and the proportion of missing is not too high (*e.g.,* around 15% or 20%, when $\tau = 2$), the proposed estimator works pretty well. If the proportion of missing is higher (*e.g.,* 35% or 40%, when $\tau = -2$) there is still a sensible underestimation of the true proportion. We can see that when the proportion of censoring is smaller, these biases get smaller.

We provide in Table 7.13 and for each scenario the Monte Carlo mean of the **estimated survivals** at 1 year and 2 years for each category as well as the standard error of the simulation (in parentheses). The table corresponds to the scenarios with $T_{max} = 3$ years, grid in months, sample size $n = 500$ and $P(X = 1) = 0.3$.

Looking at Table 7.13, we can see that the CC analysis always provides biased estimates for the survival in group $X = 0$, except for the MCAR generating pattern (the estimates are the same in all these scenarios because the observed subsample with $X = 0$ is independent on $\tau$ in the non-response mechanism). On the other hand, for the group $X = 1$, the CC analysis also provides good estimates if and only if the category $X = 1$ is well represented in the observed sample (in other words, when $\tau \gg 0$) [V, pp. 246, 263 and 279]. In a similar way that in Table 7.12, MCAR and MAR analysis provide quite similar estimators. The second one has a slightly higher standard error because we have to estimate two more parameters, $\alpha_1$ and $\alpha_2$, in the non-response mechanism (7.1) [V, pp. 282–284].

| NR Pattern | | Estimated effective prop. of $X = 1$ | |
|---|---|---|---|
| | | $P(X = 1) = 0.3$ | $P(X = 1) = 0.5$ |
| Generating | Analyzing | 3 years / 10 years | 3 years / 10 years |
| **MCAR** | CC | 0.300 / 0.300 | 0.499 / 0.499 |
| | **MCAR** | **0.301 / 0.297** | **0.500 / 0.496** |
| | MAR | 0.301 / 0.297 | 0.500 / 0.496 |
| **MAR** | CC | 0.314 / 0.323 | 0.516 / 0.525 |
| | MCAR | 0.301 / 0.300 | 0.500 / 0.499 |
| | **MAR** | **0.301 / 0.300** | **0.500 / 0.499** |
| | NI(-1) | 0.337 / 0.317 | 0.538 / 0.519 |
| | NI(1) | 0.278 / 0.291 | 0.467 / 0.484 |
| **NI(-2)** | CC | 0.216 / 0.285 | 0.391 / 0.481 |
| | MCAR | 0.189 / 0.250 | 0.347 / 0.429 |
| | MAR | 0.189 / 0.250 | 0.347 / 0.429 |
| | **NI(-2)** | **0.257 / 0.285** | **0.442 / 0.479** |
| | NI(-1) | 0.220 / 0.264 | 0.396 / 0.453 |
| **NI(2)** | CC | 0.358 / 0.339 | 0.564 / 0.543 |
| | MCAR | 0.350 / 0.321 | 0.552 / 0.523 |
| | MAR | 0.350 / 0.321 | 0.552 / 0.523 |
| | NI(1) | 0.328 / 0.311 | 0.529 / 0.511 |
| | **NI(2)** | **0.316 / 0.306** | **0.514 / 0.505** |

Table 7.12: *Monte Carlo estimated effective proportion of individuals with* $X = 1$*, for* $T_{max}$ *=3 / 10 years, grid in months and sample size* $n$ *= 500 / 1000, as a function of the non-response patterns we use and the true values of* $P(X = 1)$

| | | $\tilde{S}(t\|X=0)$ ($t$ in years) | | $\tilde{S}(t\|X=1)$ ($t$ in years) | |
|---|---|---|---|---|---|
| | | $t=1$ | $t=2$ | $t=1$ | $t=2$ |
| True values: | | **0.569** | **0.396** | **0.900** | **0.778** |
| NR Pattern | | | | | |
| Gen. | Anal. | | | | |
| MCAR | CC | *0.571* (0.035) | *0.395* (0.042) | *0.899* (0.031) | *0.775* (0.054) |
| | MCAR | 0.578 (0.031) | 0.409 (0.038) | 0.902 (0.031) | 0.789 (0.051) |
| | MAR | **0.577** (0.031) | **0.409** (0.038) | **0.903** (0.031) | **0.789** (0.050) |
| MAR | CC | 0.616 (0.033) | 0.428 (0.039) | **0.910** (0.029) | 0.785 (0.049) |
| | MCAR | 0.578 (0.030) | 0.407 (0.036) | 0.902 (0.032) | 0.784 (0.048) |
| | MAR | **0.578** (0.030) | **0.407** (0.036) | *0.902* (0.032) | **0.784** (0.048) |
| | NI(-1) | *0.575* (0.031) | *0.405* (0.037) | 0.881 (0.037) | 0.761 (0.051) |
| | NI(1) | 0.581 (0.030) | 0.410 (0.035) | 0.917 (0.028) | 0.799 (0.047) |
| NI(-2) | CC | 0.616 (0.033) | 0.428 (0.039) | 0.944 (0.029) | 0.820 (0.050) |
| | MCAR | 0.607 (0.027) | 0.434 (0.034) | 0.941 (0.032) | 0.827 (0.049) |
| | MAR | 0.607 (0.027) | 0.434 (0.034) | 0.941 (0.032) | 0.827 (0.049) |
| | NI(-2) | **0.029** (0.021) | **0.423** (0.035) | *0.899* (0.051) | **0.774** (0.058) |
| | NI(-1) | 0.600 (0.028) | 0.428 (0.035) | **0.922** (0.040) | 0.803 (0.053) |
| NI(2) | CC | 0.616 (0.033) | 0.428 (0.039) | **0.901** (0.028) | *0.776* (0.046) |
| | MCAR | 0.567 (0.030) | 0.398 (0.035) | 0.891 (0.031) | 0.772 (0.046) |
| | MAR | 0.567 (0.030) | *0.398* (0.035) | 0.891 (0.031) | 0.772 (0.046) |
| | NI(1) | *0.569* (0.030) | 0.400 (0.035) | 0.904 (0.028) | **0.786** (0.045) |
| | NI(2) | **0.571** (0.029) | **0.401** (0.035) | 0.911 (0.026) | 0.793 (0.044) |

Table 7.13: *Monte Carlo mean of the estimated survivals in the simulation at 1 year and 2 years (in parentheses the standard error of the estimates) for each category and for the $T_{max} = 3$ years, grid in months, sample size $n = 500$ and $P(X = 1) = 0.3$ scenarios.*
*Boldface: the least mean squared error estimate, Italic: the least biased estimate (if different from the least mean squared error estimate)*

No matter the level of censoring and $P(X = 1)$ value, about the proposed methodology we can see that it provides the best estimates (in the sense of minimizing the mean squared error) for category $X = 0$ [V, pp. 225, 241, 257 and 273]. For category $X = 1$ the new methodology provides the best estimates when the true $\tau$ in the non-response generating pattern is negative [V, pp. 230, 246, 263 and 279].

Comparing the upper bounds for the **relative bias** for the Grouped Kaplan-Meier estimator for the category $X = 0$ in Table 7.8 with those resulting from Table 7.13 both are of the same magnitude in MCAR and MAR generating patterns; when the missingness is non-ignorable and $\tau = -2$, the estimated relative biases are slightly larger, and they are shorter when $\tau = 2$ [V, pp. 224–225]. These conclusions are also true for the corresponding scenarios with $P(X = 1) = 0.5$ [V, pp. 240–241], and when $T_{max} = 10$ years [V, pp. 256–257 and 272–273]. Differences are bigger for category $X = 1$, in all scenarios [V, pp. 229–230, 245–246, 262–263 and 278–279].

If we look at the **standard error estimates** and the **coverage probabilities** in Table 7.14 we will find that, in general, the analyzing patterns closer to the generating pattern give the best coverage probabilities; in particular when the generating non-ignorability parameter $\tau$ is $-2$. This fact is specially relevant at the beginning of the distribution if the observation window and the sample size are large enough (*e.g.,* when $T_{max} = 10$ years, grid in month, sample size is 1000 and the generating missing pattern is NI(-2), for the survival at 1 or 2 years for $X = 0$ we obtain a coverage probability of 6% and 27%, respectively, for the CC methodology, meanwhile the resulting coverage probabilities for the semiparametric approach are all around 95% [V, p. 283]). When $P(X = 1) = 0.5$ coverage probabilities for the group $X = 0$ slightly increase and those corresponding to the group $X = 1$ slightly decrease [V, pp. 282–284]. The CC methodology only works correctly when the non-response pattern is MCAR or we are trying to estimate the survival at the tail of the distribution [V, pp. 282–284]. On the other hand, standard errors increase with the non-ignorability level in the non-response analyzing model (*i.e.,* with $|\tau|$). When $P(X = 1) = 0.3$ we can see that the estimates for the standard errors in the category $X = 0$ are sensibly smaller than the corresponding to $X = 1$ [V, pp. 282, 283]. When $P(X = 1) = 0.5$ these estimates become more similar [V, pp. 282,284].

| NR Pattern | | $X=0$ | | | $X=1$ | | |
|---|---|---|---|---|---|---|---|
| Gen. | Anal. | lse | cp | sse | lse | cp | sse |
| MCAR | CC | 0.034 | 0.954 | 0.033 | 0.032 | 0.937 | 0.033 |
| | MCAR | 0.030 | 0.938 | 0.030 | 0.030 | 0.917 | 0.032 |
| | MAR | 0.030 | 0.940 | 0.030 | 0.030 | 0.918 | 0.032 |
| MAR | CC | 0.032 | 0.694 | 0.032 | 0.028 | 0.885 | 0.029 |
| | MCAR | 0.030 | 0.934 | 0.029 | 0.029 | 0.917 | 0.031 |
| | MAR | 0.029 | 0.935 | 0.029 | 0.029 | 0.916 | 0.031 |
| | NI(-1) | 0.032 | 0.963 | 0.030 | 0.037 | 0.968 | 0.036 |
| | NI(1) | 0.030 | 0.939 | 0.029 | 0.029 | 0.889 | 0.028 |
| NI(-2) | CC | 0.032 | 0.694 | 0.032 | 0.029 | 0.593 | 0.028 |
| | MCAR | 0.027 | 0.699 | 0.028 | 0.029 | 0.631 | 0.030 |
| | MAR | 0.027 | 0.706 | 0.028 | 0.029 | 0.636 | 0.030 |
| | NI(-2) | 0.032 | 0.925 | 0.031 | 0.045 | 0.897 | 0.049 |
| | NI(-1) | 0.029 | 0.861 | 0.029 | 0.037 | 0.834 | 0.039 |
| NI(2) | CC | 0.032 | 0.694 | 0.032 | 0.028 | 0.942 | 0.028 |
| | MCAR | 0.030 | 0.949 | 0.030 | 0.030 | 0.936 | 0.030 |
| | MAR | 0.030 | 0.949 | 0.030 | 0.030 | 0.937 | 0.030 |
| | NI(1) | 0.031 | 0.965 | 0.029 | 0.031 | 0.954 | 0.027 |
| | NI(2) | 0.033 | 0.973 | 0.029 | 0.033 | 0.956 | 0.026 |

Table 7.14: *Shortest half location parameter for the estimated standard errors (lse), coverage probability of the nominal* 95% *confidence intervals (cp) and simulated standard error (sse) for each category at 1 year for the* $T_{max} = 3$ *years, grid in months, sample size* $n = 500$ *and* $P(X = 1) = 0.3$ *scenarios*

| | | $ARE_1$ | | | | $ARE_2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| NR Pattern | | $X = 0$ | | $X = 1$ | | $X = 0$ | | $X = 1$ | |
| Gen. | Anal. | $t = 1$ | $t = 2$ | $t = 1$ | $t = 2$ | $t = 1$ | $t = 2$ | $t = 1$ | $t = 2$ |
| MCAR | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.84 | 0.93 | 0.97 | 0.91 |
| | MCAR | 1.19 | 1.07 | 1.03 | 1.09 | 1.00 | 1.00 | 1.00 | 1.00 |
| | MAR | **1.21** | **1.07** | **1.03** | **1.10** | **1.02** | **1.00** | **1.00** | **1.00** |
| MAR | CC | 1.00 | 1.00 | **1.00** | 1.00 | 0.29 | 0.55 | **1.08** | 0.98 |
| | MCAR | 3.45 | 1.82 | 0.93 | 1.02 | 1.00 | 1.00 | 1.00 | 1.00 |
| | MAR | **3.45** | **1.82** | 0.93 | **1.02** | **1.00** | **1.00** | 1.00 | **1.00** |
| | NI(-1) | 3.40 | 1.80 | 0.55 | 0.85 | 0.99 | 0.99 | 0.59 | 0.83 |
| | NI(1) | 3.32 | 1.77 | 0.86 | 0.93 | 0.96 | 0.97 | 0.93 | 0.91 |
| NI(-2) | CC | 1.00 | 1.00 | 1.00 | 1.00 | 0.45 | 0.77 | 0.93 | 0.80 |
| | MCAR | 1.52 | 0.99 | 1.05 | 0.91 | 0.68 | 0.76 | 0.98 | 0.72 |
| | MAR | 1.52 | 0.99 | 1.05 | 0.90 | 0.68 | 0.76 | 0.98 | 0.72 |
| | NI(-2) | **2.23** | **1.30** | 1.07 | **1.25** | **1.00** | **1.00** | 1.00 | **1.00** |
| | NI(-1) | 1.90 | 1.16 | **1.33** | 1.25 | 0.86 | 0.89 | **1.24** | 1.00 |
| NI(2) | CC | 1.00 | 1.00 | **1.00** | 1.00 | 0.26 | 0.48 | **1.07** | 1.01 |
| | MCAR | 3.73 | 2.06 | 0.75 | 1.01 | 0.96 | 0.99 | 0.80 | 1.02 |
| | MAR | 3.73 | 2.06 | 0.75 | 1.01 | 0.96 | 0.99 | 0.80 | 1.02 |
| | NI(1) | 3.84 | 2.07 | 0.97 | **1.05** | 0.99 | 1.00 | 1.03 | **1.05** |
| | NI(2) | **3.87** | **2.08** | 0.94 | 0.99 | **1.00** | **1.00** | 1.00 | 1.00 |

Table 7.15: *Asymptotic Relative Efficiency of the different methodologies used in the simulation at 1 year and 2 years and for each category.* $ARE_1$ *takes the CC methodology as the reference and* $ARE_2$ *uses the generating non-response pattern as analyzing pattern and reference. The scenarios correspond to* $T_{max} = 3$ *years, grid in months, sample size* $n = 500$ *and* $P(X = 1) = 0.3$.

*Boldface: the most efficient estimate*

To **compare the different methodologies**, Table 7.15 shows the Asymptotic Relative Efficiency (ARE) for the scenarios with $T_{max} = 3$ years, grid in months, sample size $n = 500$ and $P(X = 1) = 0.3$. The ARE's are computed at 1 year and 2 years and for each category. As we introduced in subsection 6.4.1, $ARE_1$ takes the CC methodology as the reference and $ARE_2$ uses, as the reference, the analyzing non-response pattern that it is equal to the generating one. For group $X = 0$ we can see that the proposed methodology is more efficient than the CC analysis. In particular, more than two/three times at the beginning of the distribution if the non-response pattern is non-ignorable. For group $X = 1$, while the proposed methodology seems to be appropriated when $\tau$ is negative, it is not so clear for the other scenarios. When $T_{max} = 10$ years, we obtain smaller mean squared errors than those resulting from the more censored case, and a similar table to Table 7.15, but with more extreme values, can be derived [V, pp. 257 and 263]. When $P(X = 1) = 0.5$ corresponding ARE's values for $X = 0$ decrease, and those for $X = 1$ increase [V, pp. 273 and 279].

## 7.5 Discussion

As we have seen in the previous section, a first crucial point is the choice of the $\tau_k$ partition. On one hand, in order to obtain less biased estimates, better coverage probabilities and less mean squared errors it would be interesting to choose a thin partition; but, on the other hand, a thinner partition implies more computation time (mainly for estimating standard errors) and a reduction in the effective sample size. So, if we have observed sample size enough, a recommendation would be to select the grid size according to the expected bias, instead of the effective sample size gain. A first approximation to the expected bias can be derived from the distributions of the observed sample and the expression (6.11). If the proportion of missing data is large (for example more than 65%) the price to pay is that we will have to use medium grids with potentially more biased estimates and less coverage probabilities.

We conclude that, in general, the new methodology works better than the CC methodology; in particular, when the analyst has some idea about the type of the non-response mechanism (because it helps him/her to specify a non-response model and to control the non-ignorability parameters). We have seen that the best analysis is the one in which we are using an analyzing non-response pattern closest to the true non-response pattern that generated the data. It is interesting to see that the proposed estimator also provides good estimates in the less informative case (*i.e.,* $T_{max} = 3$ years, $P(X = 1) = 0.3$ and $\tau = -2$) in which the group $X = 1$ has an odds ratio of being observed, versus the group $X = 0$, lower than 1 ($exp(-2) = 0.135$), with a 83.75% of censoring (see Table 7.2) and a 52.65% of missing (see Table 7.5). However, while the semiparametric approach always provides more efficient estimates for group $X = 0$ than the CC analysis, when $\tau \gg 0$ (*i.e.,* the group $X = 1$ is well represented in the observed subsample) the CC methodology is more efficient than the proposed one (mainly if we use coarsened grids).

In practice, we never know which is the closest non-response model to the true missing data mechanism. So, in order to make correct inferences it is always necessary to perform a sensitivity analysis over the non-ignorability parameters and the non-response model. One strategy to apply could be:

1. Specify a plausible non-response model that incorporates all the non-response patterns (MCAR, MAR and NI).
2. Understand the role of the non-ignorability parameters.
3. Based on the information a priori, decide a region of plausible non-ignorability parameters (for the assumed non-response model).
4. Estimate the model and the survival values for each set of parameters in the region.
5. Analyze the sensitivity of the inferences as a function of the different parameters.
6. Repeat previous steps 1 to 5 for others reasonable non-response models, if it is necessary.

About the efficiency of the proposed methodology by itself, it is important to comment that in the simulation we are setting $\phi_{\boldsymbol{r}}^{(1)} = \mathbf{0}$ and a specific $\phi_{\boldsymbol{r}}^{(2)}$ that they are not necessarily the efficient choice. It explains that in some wrong specified non-response model we can obtain a more efficient estimator. For instance, in Table 7.15 it happens when the generating mechanism is NI(-2) and the analyzing mechanism is NI(-1) and we are estimating the survival for the $X = 1$ group at the beginning of the distribution. If we take a thinner grid (*e.g.,* in weeks) this effect is reduced and it does not exist if the sample size increases (see Appendix V, for more details).