Regression with a right-censored predictor using inverse probability weighting methods

Roland A. Matsouaka1,2 | Folefac D. Atem3

1Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA
2Program for Comparative Effectiveness Methodology, Duke Clinical Research Institute, Durham, North Carolina, USA
3Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston Houston, Houston, Texas, USA

Correspondence
Roland A. Matsouaka, Department of Biostatistics and Bioinformatics, Duke University, 200 Morris Street, Durham, NC 27701.
Email: roland.matsouaka@duke.edu

Summary
In a longitudinal study, measures of key variables might be incomplete or partially recorded due to drop-out, loss to follow-up, or early termination of the study occurring before the advent of the event of interest. In this paper, we focus primarily on the implementation of a regression model with a randomly censored predictor. We examine, particularly, the use of inverse probability weighting methods in a generalized linear model (GLM), when the predictor of interest is right-censored, to adjust for censoring. To improve the performance of the complete-case analysis and prevent selection bias, we consider three different weighting schemes: inverse censoring probability weights, Kaplan-Meier weights, and Cox proportional hazards weights. We use Monte Carlo simulation studies to evaluate and compare the empirical properties of different weighting estimation methods. Finally, we apply these methods to the Framingham Heart Study data as an illustrative example to estimate the relationship between age of onset of a clinically diagnosed cardiovascular event and low-density lipoprotein among cigarette smokers.

KEYWORDS
censored predictor, Cox proportional hazards model, inverse probability weighting, Kaplan-Meier estimator, regression model

1 | INTRODUCTION

Incomplete data, whether missing or censored, are ubiquitous in virtually all scientific disciplines, especially in economic, epidemiological, biological, biomedical, and environmental research studies. While several previously proven methods have been developed in the study of missing data and time to event outcomes, most of the literature on censored covariates deals the issue of the limit of detection.1-3 Nevertheless, over the past 5 years, Atem, Betensky, and colleagues have published a series of papers on randomly censored covariates.1,4-6

As opposed to a variable censored due to the limit of detection—where censoring is the result of inadequate instrument sensitivity to capture and quantify appropriately assay measures below (or beyond) some detectable limit7—randomly censored variable measurements arise when there is a time lag or limited follow-up between the time when the variable is measured and the occurrence of an event of particular interest that needs to happen for such a measurement to be accessible.1,6 For instance, when investigating familial aggregation in chronic disease incidence, Clayton modeled the possible influence parental age at the onset of a given disease might have on an individual’s risk of succumbing to a particular disease.8 Atem et al also considered the relationship between the maternal age of onset of dementia, particularly at a younger age, and amyloid deposition burden in offspring over the age of 60 years.1
Interestingly, issues about or pertaining to data analysis with a potentially randomly censored predictor are not limited to studies involving more than one generation (parent and offspring) of participants. In clinical research, for instance, a censored predictor may be present if one desires to estimate the effects of timely reperfusion via percutaneous coronary intervention (PCI) on hospital outcomes for patients presenting a ST-elevation myocardial infarction. Measures of time to PCI can be censored for several reasons, including the absence of PCI-capable hospitals in the area, door-to-balloon time greater than 90 minutes, and patient arriving at the off-hours. Censored predictors might also arise when investigating the association between the age of onset of cardiovascular disease (CVD) and the age of first cigarette smoking, the onset of chronic obstructive pulmonary disease and the number of years of smoking, or the relationship between the age of CVD onset and low-density lipoprotein (LDL) among cigarette smokers, as we will present in this paper. For the latter example, while risk factors and history of CVD can be ascertained, it is unlikely that all the participants would have developed CVD at the time of the investigation. Thus, the variable “age of onset of CVD” is not always guaranteed to be measured for all participants, that is, it usually not be fully observed.

Common analysis methods for handling censored covariates include the complete-case (CC) analysis, simple substitution methods (also known as ad hoc “fill-in” methods), and threshold methods. The complete case analysis—which consists of discarding censored observations and running the analyses on a subset of the data without censored observations—is generally easy to use and may sometimes provide unbiased results. However, it reduces the sample size, wastes valuable information by discarding observations, and often leads to inefficient results, especially under moderate or heavy censoring. Simple substitution methods—where censored observations are replaced by a constant, the overall mean, or the median of the covariate—have the advantage of using the entire dataset, but make the implicit assumption that these substituted values are actual (known) values as if they were never censored. However, treating substituted values as known may result in a biased and inefficient estimators with overly narrow confidence intervals since the methods tend to underestimate the true variability of the estimators. As argued by Helsel, there is no theoretical justification for such methods; their use is akin to data fabrication and thus should be eschewed. Threshold methods consist either of dichotomizing the randomly censored covariate or using a specific threshold (when available) to extract information relative to the covariate via dichotomization. While dichotomizing continuous covariates is already a bad idea, it does not help that much when the variable is censored. Finally, it is worth mentioning that, in the case of a time-to-event predictor and a linear dependent variable, reversing the independent and dependent variables will result in a correct hypothesis test but not in an unbiased parameter estimates.

More appropriate methods for randomly censored covariates have been inspired by methods for missing data. Indeed, censored covariates can be viewed as missing data with a specific structure since censored measures are known to lie beyond certain limiting values, even though the precise measures are unknown. Therefore, methods for randomly censored covariates include maximum likelihood estimation, estimating function approach, single and multiple imputations, as well as reverse survival regression methods. Under some specific conditions, imputation methods consist of filling censored values with some plausible values that validly represent (or replace) the censored observations. The uncertainty inherent to the imputation is judiciously incorporated in the analysis process to produce estimates and confidence intervals that reflect such an uncertainty.

Aside from the aforementioned methods, another framework that has been leveraged in the literature to handle missing data is weighting. Weighting allows us to correct for selection bias that may have been introduced by using CC analysis, that is, if we restrict the analysis solely to completely observed data. Similarly, even when the regression model is correctly specified, the presence of a censored predictor or of censored predictors can lead to biased and inefficient estimates, especially if the percentage of censored observations is high or the censoring is informative, that is, depends on patients’ demographic and medical history. Therefore, weighting improves our analysis by assigning a weight to each participant based on the occurrence of the event of interest which is used to improve efficiency via a weighted generalized linear model (GLM).

The intuitive idea of a weighting method is to define and estimate sampling weights based on the censoring mechanism to change the composition of subsample of patients with noncensored covariate measures. It generates a pseudo-population in which all noncensored participants represent themselves and provide additional copies of themselves, where the number of copies reflects the magnitude of their corresponding weight. Running such a weighted GLM may help reduce bias due to censoring and improve the precision of the CC analysis. Clearly, as in the context of missing data, using the weighting methods can also be a better alternative to censored covariate imputation.

However, to our knowledge, there is a paucity of publications that address the issue of randomly censored covariates using weighting methods. The only exceptions are the papers on treatment-duration policy for which a dynamic-treatment-regimes perspective was used by Johnson and Tsiatis, along with a handful of applications...
and extensions. Their questions of interest were “what is the optimal dose or how long treatment must be administered to see meaningful results?” On one hand, if stopped early, patients will experience minor or no side effects at all, the effect of the drug will not manifest and the drug will be wasted. On the other hand, administering a high dose of the drug may be more effective in treating the condition, but more and more patients might experience serious side effects and complications, eventually leading to treatment discontinuation. The goal was to find the optimal treatment infusion duration that can be recommended to inform future guidelines and clinical policy.

In this paper, we consider inverse probability weighting methods to estimate parameters of a GLM with a randomly censored covariate. These methods restrict their attention to observations with complete covariate measures and weight each of their contributions to obtain consistent estimators. The goal is to up weight observed data to compensate and account for the loss of information due to censored observations. For this purpose, we discuss three different weighting methods and consider two types of censoring; one being the random censoring (or administrative censoring), which occurs at the end of the study and the other, the informative censoring, where the censoring mechanism depends on patients’ (measured) characteristics and outcomes. We start in Section 2 with some notation and provide an overview of the weighting methods in GLM. Then, we introduce the use of the inverse probability censoring weighting (IPCW) where the weights are defined via the predicted probabilities of being uncensored. Next, leveraging the structure of the censored covariate in Section 2.3, we derive nonparametric weights using the Kaplan-Meier (IPCW KM) estimates of censoring times. Furthermore, assuming proportional hazards of censoring times among patients, we develop a method to determine Cox proportional hazards related weights (IPCW Cox) in Section 2.4. Furthermore, we evaluate and compare the three methods, along with the CC analysis, through simulation studies in Section 3, looking at both independent and dependent censoring. Finally, in Section 4, we apply the four methods (CC, IPCW, IPCW KM, and IPCW Cox) to assess the relationship between age of parental onset of a clinically diagnosed cardiovascular event and LDL in offspring among smokers using data from the Framingham Heart Study. In Section 5, we conclude the paper with some discussion points.

2 | INVERSE PROBABILITY OF CENSORING WEIGHTS (IPCW)

2.1 | Notation, definition, and identification

We consider a GLM specified by

\[ \begin{align*}
E(Y|X_i = x_i, Z_i = z_i) &= \mu_i, \quad i = 1, \ldots, n \\
G(\mu_i) &= \eta_i = \beta_0 + \beta_1 x_i + \beta_2 z_i, \quad \text{with } z_i = (z_{i1}, \ldots, z_{ip}) \\
Var(Y|X_i = x_i, Z_i = z_i) &= \tau^2 v(\mu_i).
\end{align*} \tag{1} \]

The variable \( Y \) is the outcome of interest, \( g(\cdot) \) is the link function, \( X \) is the continuous variable that might be randomly censored for some participants, \( Z \) is the matrix of fully measured covariates of dimension \( n \times p \), \( \beta = (\beta_0, \beta_1, \beta_2')' \) is the \((p + 2)\)-length column vector of regression model coefficients to estimate, \( v(\mu) \) is the variance function, and \( \tau^2 \) is the dispersion parameter.

The choice of the link function \( g(\cdot) \) depends on the nature of the outcome of interest. For the linear regression with a continuous outcome, the identity function \( g(u) = u \), while \( g(u) = \log(u) \) and \( g(u) = \log(u/(1 - u)) \) correspond, respectively, to the log link for count data outcome and logit link for binary outcome. We say that the model (1) is correctly specified if \( \beta \) exits such that the equation is satisfied for all values of \( X \) and \( Z \).

For each subject \( i \) in the observed data consist of \( O_i = (V_i, \Delta_i, Z_i, Y_i) \), where \( V_i = \min(X_i, C_i) \), with \( C_i \) being the censoring variable and \( \Delta_i \) the censoring indicator, that is, \( \Delta_i = 1 \) if the \( i \)th subject’s covariate value is observed and \( \Delta_i = 0 \) if the value is censored. We assume that both the censoring time \( C \) and censored covariate \( X \) are continuous.

In the absence of censoring, the generalized estimation equation for \( \beta \) is

\[ \mathbf{U(\beta)} = \sum_{i=1}^{n} U_i(\beta) = \sum_{i=1}^{n} h(X_i, Z_i; \beta) (y_i - \mu_i) \{X_i \} = 0, \quad \tag{2} \]
where \( h(X_i, Z_i; \beta) = \tau^2 \nu(\mu_i)^{-1} \frac{d^2}{d\beta} \{ g^{-1}(\eta_i) \} \) is a given scalar function and \( X_i = (1, x_i, Z_i)' \). The function \( h \) is equal to \( \tau^2 \nu(\mu_i)^{-1} \) for the identity link and 1 for both the log and the logit link functions. Other functions \( h \) are presented and discussed by McCullagh and Nelder,\(^{24}\) as well as by Liang and Zeger.\(^{25}\) Depending on the nature of the outcome \( Y \), we can no longer solve appropriately the modified estimation Equation (2). The CC analysis is based on the data \( \mathcal{O}_i = (V_i, \Delta_i, Z_i, Y_i) : \Delta = 1, i = 1, \ldots, n \) where we ignore (remove the data) observations with censored values. Although the CC analysis is easy to implement, it results in a loss of efficiency due to data deletion. Also, it may yield inconsistent parameter estimates and lead to spurious results, especially when the censoring mechanism is covariate dependent.\(^{26}\)

To circumvent such limitations, we restrict our attention to subjects with a fully observed \( X \), but their contribution is leveraged using the selection probability \( \pi(Y_i, H_i; \theta) = P(\Delta = 1 | Y_i, H_i; \theta) \) as a weight to correct for a possible selection bias created by censoring, where \( \theta \) is the associated selection parameter and \( H \) contains \( Z \) as a subset. The vector \( H \) may include, in addition to \( Z \), auxiliary variables that are associated with \( X, Y \), and the censored indicator \( \Delta \) in such a way that \( C \) and \( X \) are independent given the covariate vector \( H \) and the response \( Y \). By auxiliary variables, we refer to variables that do not provide additional information to the regression model (1), but are considered potentially informative about and predictive of the censoring \( C \) which can reduce bias due to censoring.\(^{27,28}\) The choice of variables to include in the selection model \( \pi(Y_i, H_i; \theta) \) are beyond the scope of this paper. We suppose throughout this paper that there is a set of fully observed auxiliary variables. Interested readers can refer to papers that have extensively investigated the issue and proposed necessary steps to make such a choice,\(^{29-34}\) which closely resemble the steps required to select variables to include in propensity score models to adjust for confounding or selection bias.\(^{35,36}\)

The idea of using weights based on the selection probability is reminiscent of the Horwitz and Thompson’s principles of sampling design in the analysis of survey data.\(^{37}\) This idea has regained attention in the statistical literature of missing and censored data in recent years.\(^{19,27,28,38,39}\) If \( \pi(Y_i, H_i; \theta) \) is known, the estimating equation for \( \beta \) is given by

\[
0 = \sum_{i=1}^{n} U_i \Delta_i U_i(\beta) \quad \text{where} \quad W_i = \pi(Y_i, H_i; \theta)^{-1}.
\]

As long as \( \pi(Y_i, H_i; \theta) > 0 \),

\[
E \left[ \sum_{i=1}^{n} W_i \Delta_i U_i(\beta) \right] = E \left[ \sum_{i=1}^{n} \frac{\Delta_i}{\pi(Y_i, H_i; \theta)} U_i(\beta) \right] = E \left[ \sum_{i=1}^{n} E[\Delta_i | Y_i, H_i] U_i(\beta) \right] | Y_i, H_i = E \left[ \sum_{i=1}^{n} E[\Delta_i | Y_i, H_i] U_i(\beta) \right] = E \left[ \sum_{i=1}^{n} U_i(\beta) \right] = 0.
\]

Therefore, the solution to the estimating Equation (3) is a consistent and asymptotically normal estimator of \( \hat{\beta} \). In practice, however, \( \pi(Y_i, H_i; \theta) \) is unknown in most cases and must be estimated. In the next sections, we will determine the solution to the generalized estimate (3) given a specification of the selection probability model \( \pi(Y_i, H_i; \theta) \).
2.2 IPCW via a logistic regression

One of the methods to estimate the selection probability \( \pi(Y_i, H_i; \theta) \) is via logistic regression. In that case, an estimate of \( \pi(Y_i, H_i; \theta) \) is given by

\[
\hat{\pi}(Y_i, H_i; \hat{\theta}) = P(\Delta_i = 1|Y_i, H_i) = \frac{\exp(\hat{\theta}_0 + \hat{\theta}_1 Y_i + \hat{\theta}_2 H_i)}{1 + \exp(\hat{\theta}_0 + \hat{\theta}_1 Y_i + \hat{\theta}_2 H_i)},
\]

where \( \hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2) \) is the estimator of the parameter \( \theta \) from the logistic regression model.

The selection probability \( \hat{\pi}(Y_i, H_i; \hat{\theta}) \) thus defined as the probability of observing a measured (i.e., uncensored) \( X \) given the outcome \( Y \) and the covariate vector \( H \). Since the logistic regression model is based on the binary censoring indicator \( \Delta \), this weighting scheme resembles that of the propensity score inverse probability weighting method. The basic intuition is that weighting any subject in the CC sample by the weights \( \hat{W}_i = \hat{\pi}(Y_i, H_i; \hat{\theta})^{-1} \) is equivalent to allowing such a subject to represent him- or herself and \((\hat{W}_i - 1)\) other subjects who have the same propensity (or more generally have similar characteristics) to be censored as those for whom the variable \( X \) is censored and thus not observed. Hence, in the causal inference parlance, weighting via the selection probability \( \hat{\pi}(Y_i, H_i; \hat{\theta}) \) creates a pseudo-population (i.e., the weighted sample) which consists of copies of subjects whose measures of the covariate \( X \) are not censored.

2.3 IPCW via the Kaplan-Meier Estimator

Unlike the IPCW above, which is based on logistic regression and only uses limited information as to whether an observation is censored or not, our second approach is based on the (nonparametric) Kaplan-Meier estimator, which is particularly used in the analysis of time-to-event data with censored observations.

The corresponding selection probability \( \pi(Y_i, H_i; \theta) \) can be estimated as \( \hat{\pi}(Y_i, H_i; \theta) = \hat{K}(X_i) \) where \( \hat{K}(X_i) \) is the Kaplan-Meier estimator for the survival function of the censoring time and \( K(u) = P(C > u) \), the cumulative probability of remaining uncensored from the time \( u \) onward. Note that the Kaplan-Meier estimator \( \hat{K}(. ) \) is based on the data \((V_i, \Delta^*)\) where \( \Delta^* = 1 - \Delta \), which means \( \hat{K}(. ) \) is determined by reversing the roles of survival times \( X_i \) and censoring times \( C_i \), that is, the survival time \( T_i \) censors the censored time \( C_i \).\(^{40-43}\)

This estimation method of the selection probability via the Kaplan-Meier estimator is unbiased and yields reliable results when the censoring mechanism is purely random, that is, \( C \) is independent of \( X, H \), and the response \( Y \). The potential time to censoring is defined using the survivor function \( K(u) = P(C > u) \). Such a censoring occurs when values of \( X \) are potentially censored only through administrative censoring, which is strictly related to the study design.

2.4 IPCW using Cox proportional hazards model

Valid inference from a Kaplan-Meier estimator might be possible only under the assumption that no additional covariate is related to the censoring and the censoring mechanism is independent. In other words, the distribution of censored observations is similar to the distribution of uncensored observations. If the censoring mechanism is informative, that is, when censoring is strongly related to some other covariate(s), we need to control for or associated with the outcome since the Kaplan-Meier estimator may no longer be unbiased and optimal, which may lead to biased estimates of the selection probability. In that case, the parameter of the GLM model estimated using the above Kaplan-Meier estimator of the selection probability might be biased since such an estimator may not be sufficient enough to predict censoring adequately.

Similar to the KM estimator, the Cox proportional hazards model (Cox model) approach is another technique to predict the time to an event, under the proportional hazards assumption. It incorporates into its prediction model valuable information that is provided by all the measured potential confounders. This method works when the censoring mechanism is independent or when the censoring mechanism \( C \) may depend on the set of covariates \( H \) and even indirectly on \( Y \).\(^{27,44,45}\) The latter property is extremely valuable, especially when the measured covariates are strongly connected with censoring and thus might contain additional information about the probability of censoring.
Using the time-to-event framework for our censored variable, we can use a Cox proportional hazards model the same way it is commonly used in survival analysis, to adjust for informative censoring—along with substantial loss of information when there is a high percentage of censoring during the study follow-up. In that case, assuming that the censoring times are continuous, the selection probability \( \pi(Y_i, H_i; \theta) \) can be estimated by the way of the corresponding hazard function related to the censoring events

\[
\lambda_c(u | H_i; \theta) = \lambda_0(u) \exp(\theta_1 Y_i + \theta_2 \mathbf{h}_i),
\]

where \( \lambda_0(u) \) is the baseline hazard function. In that case, the selection probability is determined by the cumulative conditional probability that an individual will be uncensored through time \( u \) given \( Y_i \) and \( H_i \),

\[
\pi(u | Y_i, H_i; \hat{\theta}) = P(C > u | Y_i = y_i, H_i = \mathbf{h}_i; \hat{\theta})
\]

\[
= \prod_{s < u} \left[ 1 - d \Lambda_c(u | Y_i = y_i, H_i = \mathbf{h}_i; \hat{\theta}) \right] = \prod_{\{j; V_j < u, \Delta_j = 0\}} \left[ 1 - \lambda_0(V_j; \hat{\theta}) \exp[\theta_1 Y_i + \theta_2 \mathbf{h}_i] \right],
\]

where \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) \) is an estimator of \( \theta = (\theta_1, \theta_2) \), \( \Lambda_c(u | Y_i = y_i, H_i = \mathbf{h}_i; \hat{\theta}) = \int_0^u \lambda_c(s | H_i; \hat{\theta}) ds \) is the cumulative hazard, and \( \lambda_0(V_j; \hat{\theta}) = \Delta_s \left[ \sum_{k=1}^n I(V_k \geq V_j) \exp[\theta_1 Y_k + \theta_2 \mathbf{h}_k] \right]^{-1} \) is the Breslow estimator of the baseline hazard function \( \lambda_0 \).46

Similar to the weights from logistic regression, the weights defined herein allows subjects with an uncensored measure of \( X \) who resemble (with respect to \( Y \) and \( H \)) those with censored measures to receive more weights. Therefore, at each time \( u \), a subject with an observed measure of \( X \) can be considered as representing \( W(u) = \pi(Y_i, H_i; \hat{\theta})^{-1} \) individuals in the pseudo population, including him- or herself.44,45

### 2.4.1 Remarks

1. Even when the selection probabilities \( \pi(Y_i, H_i; \hat{\theta}) \) are known, Robins et al have shown that it is more efficient to use estimated probabilities than the true probabilities.28,47 A heuristic explanation of this seemingly counterintuitive result is that estimating the selection probabilities enables one to use all the available data more efficiently. Indeed, such an estimation incorporates more effectively measured covariates information on both censored and uncensored observations along with auxiliary variables, significant interactions, and high-order polynomial terms as well as the outcome \( Y \).

2. In two of the examples we mentioned in the introduction, it is reasonable to assume that the censoring mechanism to which a covariate \( X \) measured on parents is unrelated to the outcome \( Y \) measured on their offspring. However, such an independence assumption cannot be generalized in all situations and should be considered only based on expert opinion, prior knowledge, informed understanding of the design, the predictor(s) of interest, and the outcome under study. Such an ascertainment of the reasons for censoring was conducted, for example, by the investigators of the ESPRIT (Enhanced Suppression of Platelet IIb/IIIa Receptor with Integrillin Therapy) study. In ESPRIT, the objective was to evaluate the effect of the duration of a continuous infusion of Integrillin for 18 to 24 hours in patients on a composite endpoint of mortality, myocardial infarction, and target urgent vessel revascularization.20 For patients who experienced serious complications during treatment, the infusion process was discontinued to provide appropriate medical attention. Therefore, when measures of the duration of the treatment were censored, it was at random and independent of the expected patient’s outcome(s).

In general, given that the censoring time \( C \) and the variable \( X \) are independent given \( Z \), we can show that

\[
f(Y | V, \Delta = 1, Z) = \frac{f(Y | X = v, C > v | Z) f(Y | Z)}{f(X = v, C > v | Z)} = \frac{f(X = v, C > v | Y, Z) f(Y | Z)}{f(X = v, C > v | Z)}
\]

\[
= \frac{f(X = v | Y, Z) P(C > v | Y, Z) f(Y | Z)}{f(X = v | Z) P(C > v | Z)} = \frac{f(X = v | Y, Z) P(C > v | Z)}{f(X = v | Z) P(C > v | Y, Z)}
\]

\[
= f(Y | X = v, Z) S_C(v | Y, Z) S_C(v | Z),
\]

provided that \( C \) and \( X \) are also independent given \( Y \) and \( Z \), where \( S_C(v,) = P(C > v.) \).
Equation (4) provides a theoretical justification of using the outcome $Y$ in the selection model when such dependence is warranted directly or indirectly. Such a modeling strategy, far from being a self-fulfilling prophecy, has become ubiquitous in the imputation procedures of missing data to ensure congeniality between the imputation and analysis models.31,48-51

3. **On stabilized weights:** Even when the condition that $\pi(Y_i, H_i; \hat{\theta}) > 0$ holds, the parameter estimator $\beta$ can be very unstable and perform poorly under small and moderate sample sizes if, for some subjects, the selection probability $\pi(Y_i, H_i; \hat{\theta})$ is very small.52-54 To mitigate the influential effects of large weights, Robins and Hernan recommend using stabilized weights, which consist of replacing the selection weights $W_i$ by $SW_i = f(C) \times W_i$ where $f(C)$ is the expected value of being uncensored.18 If we considered weights from the logistic regression, $f(C) = P(\Delta = 1)$. For the IPCW weights estimated via the Cox proportional hazards model, the expected value $f(C)$ is replaced by $f(C|u)$ which is derived using the usual Kaplan-Meir estimator $f(C|u) = \hat{K}^0(u) = P(C > u)$ (ie, the probability of being uncensored from the time $u$ onward),44,45,52,53 as indicated in Section 2.3.

4. **On Cox proportional hazards model:** Although Cox (semiparametric) proportional hazards model is undoubtedly the most popular method of analysis for right-censored data, there are other alternative parametric and semiparametric methods that can be used when appropriate (see Andersen and Keiding55 or Guo and Zheng,56 along with the references therein).

5. The above results can be extended easily to more multiple randomly censored covariates. Depending on the plausible assumptions we can make regarding the censoring distributions of these covariates, and estimate the selection probabilities $\pi(Y_i, H_i; \theta) = \prod_{k=1}^{K} \pi_k(Y_i, H_i; \theta)$, where $K$ is the total number of randomly censored covariates.

### 3 SIMULATION STUDY

We conducted simulation studies to investigate the performance of each of the proposed IPW methods. For comparison, we also included both the full model (Full), that is, the model without censoring and the CC method, to highlight the extent to which discarding observations impacts the results. Firstly, we generated two sets of data corresponding to two types of censoring (independent censoring or outcome-dependent censoring). For each set of data, we considered two different degrees of censoring, labeled “light” and “heavy” censoring, with 20% and 40% of censored observations, respectively.

#### 3.1 Simulation setup

In each dataset, we generated two fully observed covariates $Z_1$ and $Z_2$, with $Z_1 \sim N(18.5,3)$ and $Z_2 \sim Ber(0.5)$. We then generated the censored covariate $X \sim Weibull(0.2,0.25)$, and the outcome $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 X + \epsilon$, with $\epsilon \sim N(0,0.1)$. We set $\beta_0 = 0.005$, $\beta_1 = 0.01$, $\beta_2 = -0.01$, and $\beta_3 = -0.05$. Finally, we considered different scenarios for the distribution of the censoring mechanism $C$ of $X$ as shown in Table 1. For each scenario of the censoring mechanisms, we generated $M = 5000$ datasets of size $n = 400$ and $n = 600$, respectively.

| TABLE 1 | Distribution of the censoring variable C |
|----------|-----------------------------------------|
|          | **Censoring Type**                       |
|          | **Independent Censoring**               |
|          | **Outcome-Dependent Censoring**         |
| Full     | No censoring                            | No censoring                          |
| Light    | $C \sim Weibull(1,2)$                   | $C_1 \sim Weibull(1,2); C_3 \sim Weibull(1.5,2)$; $C = I(\epsilon > 0)C_1 + I(\epsilon \leq 0)C_3$ |
| (20% censored) |                                 |                                        |
| Heavy    | $C \sim Weibull(1,0.35)$                | $C_2 \sim Weibull(1.0,35); C_4 \sim Weibull(1.5,0.35)$; $C = I(\epsilon > 0)C_2 + I(\epsilon \leq 0)C_4$ |
| (40% censored) |                                      |                                        |

*Note: $I(.)$ is the standard indicator function.*
3.2 Evaluation criteria

From each generated data set, we run a weighted regression model—based on the aforementioned weighting methods—and determined the estimates \( \hat{\beta}_m \), \( m = 1, \ldots, M \), of the parameters of interest \( \beta_3 \) and considered the overall estimate \( \tilde{\beta}_3 = M^{-1} \sum_{m=1}^{M} \hat{\beta}_m \). Then, we calculated five different measures of performance: the bias \( (\hat{\beta}_3 - \beta_3) \); the bias percentage \( \% \text{Bias} = 100|\text{Bias}|/\beta_3 \); the model-based SE (ie, the average of all standard errors of \( \beta_3 \) from each fitted model); the Monte Carlo simulation SD (ie, the empirical standard error of the estimates of \( \beta_3 \) over all \( M \) simulated datasets); and the mean squared error (MSE = \( \text{Bias}^2 + \text{SE}^2 \)). The most consistent and efficient method will be the one with the smallest bias and MSE as well as similar SD and SE. Naturally, we expected the scenario with no censoring (full data) to yield results with the most consistent and efficient model parameter. Therefore, the performance of the different methods were then assessed based on how close the were from the full model.

3.3 Simulation results

Tables 2 and 3 summarize the results from our simulation study. They indicate, as expected, that the performance of the different methods (along with the CC analysis) depends on the sample size, the proportion of censored observations on the potential censored predictor, and the type of censoring mechanism. For each method, at a given percentage of the censored predictor, the bias as well as the percentage decreases in magnitude as the sample size increases. Also, for a given sample size, the bias increases as the percentage of censored covariate increases from 20% to 40%.

For Table 2, the censoring mechanism is independent, that is, it is neither a function of other covariate(s) nor the dependent variable. These simulated results for sample sizes \( n = 400 \) and \( n = 600 \), showed that the three IPCW methods together with the CC approach were marginally biased. The CC approach and the IPCW Cox are least biased. In the former, censored observations are deleted that might result to drop in power. Thus, the SE of the CC analyses were high and increased tremendously as the censoring rate increased. This explains the high MSEs for the CC analysis when censoring is heavy as compared to the other the three IPCW methods. The IPCW Cox performed best as compared to the other IPCW approaches because it used all available information to determine its weights. The IPCW based on logistic regression to predict the propensity of \( X \) not being censored and the IPW based on Kaplan-Meier to predict the distance of an event without taking into consideration other covariates appeared to be indistinguishable. These approaches resulted in the highest bias and MSE. The IPCW based on Cox model that tries to construct the weight slightly differently by predicting a (time) distance to an event while taking into consideration of all the variables, instead of the propensity of event performed the best. Its bias and MSE are the smallest.

| Censoring | Method  | \( n = 400 \) | \( n = 600 \) |
|-----------|---------|--------------|--------------|
|           | Bias    | % Bias       | SE  | SD  | MSE | Bias    | % Bias       | SE  | SD  | MSE |
| Light     | Full    | 0.01 2       | 0.48 | 0.23 | 0.0  | 0.38  | 0.39       | 0.15 |
|           | CC      | 0.01 2       | 1.20 | 0.15 | 0.01 | 1.21  | 0.98       | 0.97 |
|           | IPCW    | -0.10 20     | 0.41 | 1.17 | -0.10| 0.21  | 0.38       | 1.16 |
|           | IPCW KM | 0.09 18      | 0.52 | 1.10 | 0.08 | 0.38  | 0.23       | 0.81 |
|           | IPCW Cox| 0.02 4       | 0.49 | 0.28 | 0.01 | 0.47  | 0.39       | 0.16 |
| Heavy     | Full    | 0.01 2       | 0.48 | 0.23 | 0.0  | 0.38  | 0.39       | 0.15 |
|           | CC      | 0.07 14      | 3.23 | 3.21 | 0.04 | 3.21  | 2.63       | 2.64 |
|           | IPCW    | 0.18 36      | 1.48 | 1.48 | 0.18 | 3.44  | 2.63       | 4.43 |
|           | IPCW KM | -0.19 38     | 1.73 | 6.60 | 0.18 | 3.60  | 1.13       | 2.21 |
|           | IPCW Cox| 0.03 6       | 1.42 | 2.11 | 0.03 | 2.11  | 1.32       | 1.83 |

**Note:** Bias, SD, and SE provided in \( 10^{-1} \); MSE provided in \( 10^{-4} \).

**Abbreviations:** CC, complete-case analysis; IPCW, inverse probability of censoring weights; IPCW KM, IPCW via Kaplan-Meier estimator; IPCW Cox, IPCW via a Cox proportional hazards model for censoring; MSE, mean squared error.
Thus, in the case of independent censoring, even though censoring is completely random, the (final) outcome $Y$ still depends indirectly on this random process and only the IPCW Cox takes into consideration all these components. In other words, the IPCW Cox uses all the information at hand while the other approaches use partial information. The IPCW Cox takes explicitly into consideration the time at censoring, all possible available predictors and the outcome $Y$, which is included as a predictor in the Cox model, to come up with an optimal weights. Although the IPCW adjust also for possible available predictors and the outcome, the logistic regression model used to estimate the weights is based on a binary indicator of whether censoring occurs or not. This logistic model does not adjust for the time at which each censored event occurs. Whether censoring happens at the very beginning of the follow-up period or not, it is treated exactly the same as the one occurring toward the end of the follow-up period. Thus, the loss of information and precision inherent to the IPCW method. Finally, while the IPCW KM accounts for when the censoring event occurs, it does not factor in the information provided by the available predictors or the outcome, leading also to less precise estimates of the censoring weights compared to the IPCW Cox. In the case of missing data, the IPCW approach uses all available information and should produce optimal weights. Since we do not have a time-to-event scenario in missing data, the IPCW Cox should not be considered in such a context.

For Table 3, when censoring is dependent on the outcome, that is, the censoring mechanism is a function of the dependent variable. These simulated results for sample sizes $n = 400$ and $n = 600$ showed significant differences between the three IPCW methods together with the CC approach in terms of the bias and MSE. The bias of the CC, IPCW, and IPCW KM increases tremendously as censoring increased from light to heavy. The MSE of these three approaches also experienced a similar increase. The performance of the IPCW Cox is slightly biased as compared to that of the Full model. The bias and MSE reduce with an increased sample size from $n = 400$ to $n = 600$, but increased slightly as censoring increased from light to heavy. It is also worth mentioning that, as for heavy censoring, the model-based SEs and the simulation SD for the proposed IPCWs are not very close for the heavy censoring cases. This is due to a reduction in precision as censoring increases. Thus, the Kaplan-Meier estimator and Cox proportional hazard model suffered a reduction in power in estimating its parameters, thus reducing the precision of the weights.

### Additional simulations

#### 3.4.1 Simulations to mimic the real data example

Similar to the real data example, we generated two fully observed covariates $Z_1$ and $Z_2$, with $Z_1 \sim Ber(0.53)$ and $Z_2 \sim Ber(0.52)$. We then generated the potential censored covariate $X \sim Uniform(0.3, 1.30)$, and the outcome $Y = \beta_0 + $
TABLE 4  Simulation results: A covariate-dependent censoring to mimic the real data example

| Censoring | Method  | Bias | % Bias | SE  | SD  | MSE |
|-----------|---------|------|--------|-----|-----|-----|
| 20%       | Full    | 0    | 0      | 0.12| 0.12| 1.4 |
|           | CC      | −0.001| 0.28   | 0.14| 0.14| 4.1 |
|           | IPCW    | 0.02 | 5.4    | 0.14| 0.31| 2.0 |
|           | IPCW KM | 0.004| 0.99   | 0.14| 0.14| 2.0 |
|           | IPCW Cox| 0.004| 1.09   | 0.14| 0.14| 0.2 |
| 40%       | Full    | 0    | 0      | 0.12| 0.12| 1.4 |
|           | CC      | 0.02 | 5.4    | 0.16| 0.16| 2.6 |
|           | IPCW    | 0.03 | 8.1    | 0.16| 0.33| 2.7 |
|           | IPCW KM | −0.02| 5.4    | 0.15| 0.17| 2.3 |
|           | IPCW Cox| −0.02| 5.2    | 0.15| 0.16| 2.4 |
| 65%       | Full    | 0    | 0      | 0.12| 0.12| 1.4 |
|           | CC      | 0.04 | 10.8   | 0.21| 0.21| 4.6 |
|           | IPCW    | 0.07 | 18.9   | 0.21| 0.23| 4.9 |
|           | IPCW KM | 0.06 | 16.2   | 0.16| 0.21| 2.9 |
|           | IPCW Cox| 0.05 | 13.5   | 0.16| 0.21| 2.8 |

Note: Bias, SD, and SE provided in 10⁻²; MSE provided in 10⁻⁶.
Abbreviations: CC, complete-case analysis; IPCW, inverse probability of censoring weights; IPCW KM, IPCW via Kaplan Meier estimator; IPCW Cox: IPCW via a Cox proportional hazards model for censoring; MSE, mean squared error.

β₁Z₁ + β₂Z₂ + β₃X + ε, with ε ~ N(0,0.01). To mimic the real data analysis, we increased the sample size to n = 850 and chose the variance of ε to be 0.01, that is, the variance of the outcome is about 10 times smaller than in the previous two simulated. We set β₀ = 4.90, β₁ = 0.0037, β₂ = 0.10, and β₃ = 0.045. Finally, we considered different scenarios for the distribution of the censoring mechanism C of X, we set C = C₁ ~ Weibull(0.75, q) if Z₁ = 0 and C = C₂ ~ Weibull(1.25, q) if Z₁ = 1, where q = 2.50, 1.50, and 0.70 for 20%, 40%, and 65% censoring, respectively. For each covariate-dependent censoring mechanisms, we generated M = 5000 datasets of sample size n = 850. We also simulated data with an interaction term between the potential censored covariate X and Z₁. The regression coefficient (β₄) for this interaction term was set at 0.05. All other parameters were kept as described in the linear regression without an interaction.

3.4.2  Simulation results to mimic the real data example

Table 4 summarizes the results from this covariate-dependent simulated data. The choice of the variance of the outcome together with the increase in the sample size of the simulated data resulted to increase precision. The four approaches, that is, CC, IPCW, IPCW KM, and IPCW Cox result in relatively unbiased estimates for the three proportions of censoring. Nonetheless, as censoring increases to 65% the three IPCWs approaches percentage bias increase significantly from less than 10% to 13.5 ≤ bias ≤ 18.9, with the least percentage bias from the IPCW Cox and the highest from the IPCW. However, based on MSE, the effect of these biased estimates from the IPCW KM and IPCW Cox approaches are minimal, since estimates from these approaches perform better than minimal percentage bias CC approach. The results from the model with the interaction term are presented in Table A1.

4  APPLICATION TO FRAMINGHAM HEALTH STUDY

Cigarette smoking is a major risk factor for CVD. The adverse effects of smoking on CVD risk involve complex mediation through multiple interrelated mechanisms, including increased oxidative stress, endothelial injury and dysfunction,
Table 5 Relationship between age of onset of cardiovascular disease and low-density lipoprotein among cigarette smokers

| Method     | Variable | Estimate | SE  | t-value | P-value |
|------------|----------|----------|-----|---------|---------|
| CC*        | Intercept| 5.0091   | 0.0535| 93.58   | <.0001  |
|            | Age      | 0.0014   | 0.0010| 1.36    | .1762   |
|            | Beer     | 0.0195   | 0.0226| 0.87    | .3876   |
|            | Male     | 0.0311   | 0.0234| 1.33    | .1855   |
| IPCW       | Intercept| 5.0297   | 0.0328| 153.19  | <.0001  |
|            | Age      | 0.0002   | 0.0006| 0.30    | .3767   |
|            | Beer     | 0.0158   | 0.0145| 1.08    | .2784   |
|            | Male     | 0.0419   | 0.0145| 2.89    | .0040   |
| IPCW KM    | Intercept| 4.8039   | 0.0198| 242.45  | <.0001  |
|            | Age      | 0.0036   | 0.0005| 7.03    | <.0001  |
|            | Beer     | 0.0531   | 0.0142| 3.74    | .0002   |
|            | Male     | 0.1079   | 0.0142| 7.62    | <.0001  |
| IPCW Cox   | Intercept| 4.8074   | 0.0203| 237.12  | <.0001  |
|            | Age      | 0.0037   | 0.0005| 6.81    | <.0001  |
|            | Beer     | 0.0431   | 0.0140| 3.08    | .0021   |
|            | Male     | 0.1176   | 0.0138| 8.50    | <.0001  |

Abbreviations: CC, complete-case analysis; IPCW, inverse probability of censoring weights; IPCW KM, IPCW via Kaplan-Meier estimator; IPCW Cox, IPCW via a Cox proportional hazards model for censoring.

*Age at onset of CVD; Beer represents the beer consumption.

altered blood coagulation, altered metabolism, and derangements of lipid composition. The association between LDL and CVD has been well studied and documented among individuals receiving certain therapeutic interventions, but not much is known about this association among cigarette smokers.

To assess this question, we compared the CC analysis to the three different IPCW methods using data from the Framingham Heart Study (FHS) database, a well-known premier, longitudinal prospective cohort study for studying CVDs and other terminal diseases. For this paper, we consider data from the Framingham Offspring Cohort (FHSO) launched in 1971 and whose participants have been examined, on average, every 3 to 4 years since enrollment. The FHSO dataset consists of a sample of 3514 biological descendants of the Original FHS Cohort, 1576 of their spouses and 34 adopted offspring for a total sample of 5124 subjects (48% males). Our analyses used observations collected during the FHSO Exam 7 (1998-2001, n = 3,539) and FHSO Exam 8 (2007, n = 2898). After cleaning and merging the required data by including only smokers, the sample size was reduced to n = 886, of which 292 (32.96%) were clinically diagnosed with CVD, 468 (52.82%) were regular beer consumers, and 466 (52.60%) were male.

In Table 5, we provided a summary of all four approaches in the association between age of onset of CVD as a predictor and log(LDL) as a dependent variable, while controlling for sex (male vs female) and beer intake (yes or no) as confounders. As shown, the estimates of the age of onset of CVD differed across the different methods. The CC analysis estimate, based on about 33% noncensored observations, was greater than the IPCW but less IPCW KM and IPCW Cox. For the IPCW approach, a logistic regression for observing a noncensored event was fitted with beer intake, male, and LDL as covariates. For the IPCW KM the weight was derived through a Kaplan-Meier estimator approach for time to onset of CVD, while for the IPCW Cox, the weight was derived from the Cox regression model for time to onset of CVD with beer intake, male, and LDL as explanatory variables.

The estimates for beer intake and sex (male vs female) were similar for IPCW KM and IPCW Cox but not for CC or IPCW. Interestingly, the parameter of interest, age of onset of CVD, was significant for IPCW KM and IPCW Cox (P < .0001) but not for CC or IPCW. The point estimate for the IPCW KM and IPCW Cox method was significantly larger (≈0.036) than those of the other two methods (resp. .0014 and .0002 for CC and IPCW).

In the CC analysis method, deleting observations (≈67%) resulted in a significant drop in precision as shown by the increased SEs. As demonstrated by the simulation studies, the CC analysis method together with the IPCW and IPCW KM
methods resulted in biased estimates when censoring depends on the outcome. However, when censoring is dependent on other covariates along with a sufficiently large number of events, as it is the case in this example, the IPCW KM results to estimates that are comparable to IPCW Cox. Since the censoring mechanism in the FHSO data likely depends on baseline variables, the IPCW KM and the IPCW Cox (unlike the CC and the IPCW) resulted in statistically significant estimates for the age of onset of CVD and thus provided a more realistic assessment of the relationship between the onset of CVD and LDL. Finally, the standard errors of the parameter of interest were roughly the same across all the three probability weighting methods.

5 | DISCUSSION

In this paper, we have presented three different inverse probability weighting methods, together with analyses from the CC method to model data when the primary predictor of interest is right-censored. Throughout our simulation study and the analysis from the Framingham Heart Study data, the need to properly account for censoring and employing a statistical model that provides unbiased estimates of the parameter of interest irrespective of the type of censoring cannot be overemphasized.

The simulation study showed varying percentage bias for the four methods, that is, CC analysis, inverse probability weight via a logistic regression model, the inverse probability weighting via Kaplan-Meier estimator and the inverse probability weighting via Cox model. For well-powered studies, all three IPCW approaches together with the CC resulted in less biased estimates when the censoring was either dependent on baseline covariates and completely independent. As precision dropped, the IPCW and IPCW KM resulted in biased estimates. Furthermore, when censoring was dependent on the outcome of interest, the CC, IPCW, and IPCW KM approach resulted in biased estimates. Since it is very difficult to have a sample where censoring is completely independent or dependent (of the other covariates and outcome), these approaches are likely to result in biased estimates and poor inferences.

This becomes more evident in the analysis of real-world data we considered. The estimates from CC and IPCW methods resulted in an insignificant association between the age of onset of CVD and LDL among cigarette smokers. These two approaches that resulted in biased estimates failed to take into consideration the time of censoring in the model. However, by accounting for both the variable dependence of the censoring mechanism and incorporating the actual time of censoring in its analytic approach, the inverse probability weight via Kaplan-Meier and Cox model provided a significant estimate. Furthermore, it is worth mentioning that the inverse probability weight via the Kaplan-Meier estimator resulted in estimates that are like those based on the inverse probability weight via the Cox model. Nonetheless, since the inverse probability weight via Cox is the only reliable method that resulted in an unbiased estimate irrespective of the type of censoring mechanism, the estimates for the association between age of onset of CVD and LDL among smokers should be based on IPCW Cox.

This work highlights the robustness provided by the distance (or time) to event inverse probability weighting method via the Cox model in estimating the effect of a potentially right-censored covariate on an outcome. Even though the weights estimated via a Kaplan-Meier estimator was also based on the distance to the event, this method does not consider possible measured confounders and, thus, lack of robustness when censoring is dependent on the outcome. Finally, the traditional inverse probability weight based on the propensity score of not being censored carries less information and appeared weaker when censoring is directly related to the outcome of interest.

ORCID

Folefac D. Atem https://orcid.org/0000-0002-1142-7806

REFERENCES

1. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Linear regression with a randomly censored covariate: application to an Alzheimer’s study. *J Royal Stat Soc Ser C (Appl Stat)*. 2017;66(2):313-328.
2. Kong S, Nan B. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*. 2016;103(1):161-174.
3. Atem FD, Matsouaka RA, Zimmern VE. Cox regression model with randomly censored covariates. *Biometr J*. 2019;61(4), pp.1020-1032.
4. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Multiple imputation of a randomly censored covariate improves logistic regression analysis. *J Appl Stat*. 2016;43(15), pp.2886-2896.
5. Qian J, Chiou SH, Maye JE, Atem F, Johnson KA, Betensky RA. Threshold regression to accommodate a censored covariate. *Biometrics*. 2018;74(4):1261-1270.
6. Atem FD, Sampene E, Greene TJ. Improved conditional imputation for linear regression with a randomly censored predictor. Stat Methods Med Res. 2019;28(2):432-444.
7. Schisterman EF, Little RJ. Opening the black box of biomarker measurement error. Epidemiology. 2010;21(Suppl 4):S1.
8. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978;65(1):141-151.
9. Hira RS, Bhatt DL, Fonarow GC, et al. Temporal trends in care and outcomes of patients receiving fibrinolytic therapy compared to primary percutaneous coronary intervention: insights from the get with the guidelines coronary artery disease (GWTG-CAD) registry. J Am Heart Assoc. 2016;5(10):e004113.
10. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere. 2006;65(11):2434-2439.
11. Austin PC, Hoch JS. Estimating linear regression models in the presence of a censored independent variable. Stat Med. 2004;23(3):411-429.
12. Rigobon R, Stoker TM. Bias from censored regressors. J Bus Econ Stat. 2009;27(3):340-353.
13. Yuan KH, Savalei V. Consistency, bias and efficiency of the normal-distribution-based MLE: the role of auxiliary variables. J Multivar Anal. 2016;150:277-290.
14. Irwin JR, McClelland GH. Negative consequences of dichotomizing continuous predictor variables. J Mark Res. 2003;40(3):366-371.
15. Fitzsimons GJ. Death to dichotomizing. J Consum Res. 2008;35(1):5-8.
16. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7(1):19.
17. Tsimikas JV, Bantis LE, Georgiou SD. Inference in generalized linear regression models with a censored covariate. Comput Stat Data Anal. 2012;56(6):1854-1868.
18. Hernan MA, Robins JM. Causal Inference. Boca Raton, FL: CRC Press; 2018.
19. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. Stat Methods Med Res. 2013;22(4):1070-1087.
20. Johnson BA, Tsiatis AA. Estimating mean response as a function of treatment duration in an observational study, where duration may be informatively censored. Biometrics. 2004;60(2):315-323.
21. George SL, Wang X, Pang H. Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis. Boca Raton, FL: CRC Press; 2016. 474pp.
22. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham heart study and the epidemiology of cardiovascular disease: a historical perspective. Lancet. 2014;383(9921):999-1008.
23. Tsoa CW, Vasan RS. Cohort profile: the Framingham heart study (FHS): overview of milestones in cardiovascular epidemiology. Int J Epidemiol. 2015;44(6):1800-1813.
24. McCullagh P, Nelder JA. Generalized Linear Models. 2nd ed. London, UK: Chapman & Hall; 1989.
25. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13-22.
26. Hernán MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-625.
27. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. AIDS Epidemiology. New York, NY: Springer; 1992:297-331.
28. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89(427):846-866.
29. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.
30. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Methods. 2001;6(4):330.
31. Schafer JL. Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica. 2003;57(1):19-35.
32. Yuan KH, Savalei V. Consistency, bias and efficiency of the normal-distribution-based MLE: the role of auxiliary variables. J Multivar Anal. 2014;124:353-370.
33. Raykov T, West RT. On enhancing plausibility of the missing at random assumption in incomplete data analyses via evaluation of response-auxiliary variable correlations. Struct Equ Model Multidiscip J. 2016;23(1):45-53.
34. Enders CK. Multiple imputation as a flexible tool for missing data handling in clinical research. Behav Res Ther. 2017;98:4-18.
35. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-1156.
36. Greenland S, Pearl N. Statistical foundations for model-based adjustments. Ann Rev Public Health. 2015;36:89-108.
37. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47(260):663-685.
38. Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. Ann Stat. 1981;9(6):1276-1288.
39. Bartlett JW, Carpenter JR, Tilling K, Vansteelandt S. Improving upon the efficiency of complete case analysis when covariates are MNAR. Biostatistics. 2014;15:719-730.
40. Zhao H, Tsiatis AA. A consistent estimator for the distribution of quality adjusted survival time. Biometrika. 1997;84(2):339-348.
41. Lin D. Linear regression analysis of censored medical costs. Biostatistics. 2000;1(1):35-47.
42. Anstrom KJ, Tsiatis AA. Utilizing propensity scores to estimate causal treatment effects with censored time-lagged data. Biometrics. 2001;57(4):1207-1218.
43. Bang H, Tsiatis AA. Median regression with censored cost data. Biometrics. 2002;58(3):643-649.
Robins JM. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. Proceedings of the Biopharmaceutical Section. Vol 24. San Francisco, CA: American Statistical Association; 1993:3.

Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics. 2000;56(3):779-788.

Breslow NE. Contribution to the discussion of the paper by DR Cox. J R Stat Soc Ser B. 1972;34(2):216-217.

Robins JM, Rotnitzky A, Zhao LP. Analysis of semi-parametric regression models for repeated outcomes in the presence of missing data. J Am Stat Assoc. 1995;90(429):106-121.

Moons KG, Donders RA, Stijnen T, Harrell FE. Using the outcome for imputation of missing predictor values was preferred. J Clin Epidemiol. 2006;59(10):1092-1101.

Little RJ. Regression with missing X’s: a review. J Am Stat Assoc. 1992;87(420):1227-1237.

Meng XL. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26(16):3057-3077.

Little RJ. Regression with missing X’s: a review. J Am Stat Assoc. 1992;87(420):1227-1237.

Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Stat Med. 2007;26(16):3057-3077.

Meng XL. Multiple imputation inferences with uncongenial sources of input. Stat Sci. 1994;9(4):538-558.

Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-560.

Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Commun Health. 2006;60(7):578-586.

Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008;13(4):279.

Andersen PK, Keiding N. Survival Analysis, Overview. In Armitage P and Colton T, eds. Encyclopedia of Biostatistics. Wiley, Chichester; 1995.8pp.4452-4461.

Guo S, Zeng D. An overview of semiparametric models in survival analysis. J Stat Plann Infer. 2014;151:1-16.

Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38(11):2074-2102.

Gossett LK, Johnson HM, Piper ME, Fiore MC, Baker TB, Stein JH. Smoking intensity and lipoprotein abnormalities in active smokers. J Clin Lipidol. 2009;3(6):372-378.

Silverman MG, Ference BA, Im K, et al. Association between lowering LDL-C and cardiovascular risk reduction among different therapeutic interventions: a systematic review and meta-analysis. Jama. 2016;316(12):1289-1297.

**How to cite this article:** Matsouaka RA, Atem FD. Regression with a right-censored predictor using inverse probability weighting methods. Statistics in Medicine. 2020;39:4001–4015. https://doi.org/10.1002/sim.8704
### APPENDIX A. ADDITIONAL SIMULATION RESULTS

**TABLE A1** Simulation results for a covariate-dependent censoring with interaction between fully observed covariate (FOC) and a censored covariate \((n = 850)\)

| Censoring | Method     | Potential Censored Covariate (PCC) | Interaction between FOC and PCC |
|-----------|------------|-----------------------------------|-------------------------------|
|           |            | Bias | % Bias | SE  | SD  | MSE | Bias | % Bias | SE  | SD  | MSE |
| 20%       | Full       | 0    | 0      | 0.17| 0.17| 2.9 | 0    | 0      | 0.24| 0.24| 5.8 |
|           | CC         | 0.01 | 2.7    | 0.21| 0.27| 5.8 | 0.01 | 0.2    | 0.28| 0.28| 7.9 |
|           | IPCW       | 0.01 | 2.7    | 0.24| 0.27| 5.8 | 0.08 | 1.6    | 0.30| 0.44| 9.6 |
|           | IPCW KM    | 0.03 | 8.1    | 0.20| 0.20| 4.1 | 0.12 | 2.4    | 0.25| 0.25| 7.7 |
|           | IPCW Cox   | 0.01 | 2.7    | 0.21| 0.22| 4.4 | 0.07 | 1.4    | 0.26| 0.26| 7.3 |
| Full      |            | 0    | 0      | 0.17| 0.17| 2.9 | 0    | 0      | 0.24| 0.24| 5.8 |
|           | CC         | 0.02 | 5.4    | 0.23| 0.24| 5.3 | 0.03 | 0.6    | 0.32| 0.32| 10  |
|           | IPCW       | 0.05 | 13.5   | 0.22| 0.24| 5.1 | 0.40 | 2.8    | 0.28| 0.31| 24  |
|           | IPCW KM    | 0.09 | 24     | 0.20| 0.26| 4.8 | 0.20 | 4      | 0.27| 0.30| 11  |
|           | IPCW Cox   | 0.05 | 13.5   | 0.21| 0.27| 4.6 | 0.20 | 5      | 0.26| 0.31| 1.0 |
| 65%       | Full       | 0    | 0      | 0.17| 0.17| 2.9 | 0    | 0      | 0.24| 0.24| 5.8 |
|           | CC         | 0.02 | 5.4    | 0.30| 0.30| 9.0 | 0.03 | 0.6    | 0.43| 0.44| 19  |
|           | IPCW       | 0.05 | 13.5   | 0.24| 0.33| 6.0 | 0.20 | 4      | 0.30| 0.51| 13  |
|           | IPCW KM    | 0.11 | 30     | 0.20| 0.32| 5.2 | 0.30 | 6      | 0.27| 0.41| 16  |
|           | IPCW Cox   | 0.08 | 21     | 0.21| 0.33| 5.1 | 0.20 | 4      | 0.26| 0.37| 11  |

*Note:* Bias, SD, and SE provided in \(10^{-2}\); MSE provided in \(10^{-6}\).
Abbreviations: CC, complete-case analysis; IPCW, inverse probability of censoring weights; IPCW KM, IPCW via Kaplan Meier estimator; IPCW Cox, IPCW via a Cox proportional hazards model for censoring; MSE, mean squared error.