Received: 1 December 2019 | Revised: 24 September 2020 | Accepted: 26 September 2020

DOI: 10.1002/sim.8775

RESEARCH ARTICLE

Statistics in Medicine WILEY

Robust estimation in the nested case-control design under a misspecified covariate functional form

Michelle M. Nuño[1,2] | **Daniel L. Gillen**[3]

[1]Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA

[2]Children's Oncology Group in Monrovia, CA

[3]Department of Statistics, University of California Irvine, Irvine, California, USA

Correspondence
Michelle M. Nuño, Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA.
Email: mnuno@usc.edu

Funding information
National Institutes of Health, Grant/Award Numbers: P30AG066519, R01AG053555; National Science Foundation Graduate Research Fellowship, Grant/Award Number: DGE-1321846

The Cox proportional hazards model is typically used to analyze time-to-event data. If the event of interest is rare and covariates are difficult or expensive to collect, the nested case-control (NCC) design provides consistent estimates at reduced costs with minimal impact on precision if the model is specified correctly. If our scientific goal is to conduct inference regarding an association of interest, it is essential that we specify the model a priori to avoid multiple testing bias. We cannot, however, be certain that all assumptions will be satisfied so it is important to consider robustness of the NCC design under model misspecification. In this manuscript, we show that in finite sample settings where the functional form of a covariate of interest is misspecified, the estimates resulting from the partial likelihood estimator under the NCC design depend on the number of controls sampled at each event time. To account for this dependency, we propose an estimator that recovers the results obtained using using the full cohort, where full covariate information is available for all study participants. We present the utility of our estimator using simulation studies and show the theoretical properties. We end by applying our estimator to motivating data from the Alzheimer's Disease Neuroimaging Initiative.

KEYWORDS
efficient sampling, functional form, model misspecification, nested case-control

1 | INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease responsible for memory loss that also inhibits the ability to perform daily tasks. AD trials require that participants undergo various tests to help detect progression of the disease. One such examination is the Alzheimer's Disease Assessment Scale-11 (ADAS-11) which was created to evaluate cognitive and behavioral function,[1] both of which are compromised by AD. Along with these tests, participants often have to undergo genotype testing to check for the presence of the Apolipoprotein e4 (APOE e4) allele which is associated with higher risk of progression to AD.[2] Investigators are also often interested in collecting measures of Aβ, a biomarker of the disease. One way to measure levels of Aβ is in cerebrospinal fluid (CSF). However, participants are often unwilling to undergo this procedure,[3] so availability of CSF samples is limited. A sampling scheme such as the nested case-control (NCC) design would help reduce costs associated with the testing procedures by only requiring genotype testing and processing of CSF samples for a subsample from the original sample.[4]

The NCC design proposed by Thomas (1977) (the focus of this manuscript) is implemented using the Cox proportional hazards (PH) model. Under the PH model (we refer to the estimator as the partial likelihood [PL] estimator), subjects

Abbreviations: AD, Alzheimer's disease; NCC, nested case-control; SRS, simple random sample.

who experience an event provide more information than censored observations. The NCC design makes use of this fact by requiring full covariate information from all subjects who experience an event (ie, progress to AD) and a subsample from subjects who do not experience the event. At each event time, M (usually $1 \leq M \leq 4$) controls are randomly sampled without replacement from everyone who is still at risk at that time. In this way, the NCC design provides great reduction in costs when the event of interest is rare. As stated earlier, in the context of our problem only a subsample of subjects in the original sample would need to undergo genotype testing and a large portion of CSF samples would not have to be processed. When the proportional hazards assumption holds, the PL estimator under the NCC sampling scheme recovers the results from the PL estimator under the full cohort (FC). Moreover, because the NCC design uses all events, it is more efficient than taking a simple random sample of the same size (cf. Reference 5).

If the goal of a study is to conduct inference for the association between a pre-specified predictor of interest and a given outcome, it is crucial to specify the statistical model used for inference a priori to avoid multiple testing bias.[6–9] It is difficult, however, to ensure that all assumptions hold for a priori specified models. Because of this, it is important to consider properties of statistical models when the underlying assumptions used in the theoretical development of the model do not hold. In previous work, it has been shown that when the proportionality assumption does not hold under the NCC design and a binary predictor is of interest, the expectation of the sampling distribution of the usual NCC estimator will depend on M, the number of controls sampled at each event time.[10] Misspecification of the functional form of continuous covariates in the NCC design has not been investigated.

If the PH assumption holds under a correctly specified PH model, misspecification of the functional form of a covariate in the model will induce non-proportional hazards. Therefore, based upon the results provided in the work of Nuño and Gillen,[10] it is natural to hypothesize that if the functional form of a covariate is misspecified in the usual NCC design then the expectation of the sampling distribution of the usual NCC estimator will depend on the parameters of the design (ie, the number of controls sampled at each event time). It is important to note, however, that this dependence arises differently than that explored in the work of Nuño and Gillen[10] which considered the time-varying effect of a discrete covariate as opposed to an induced dependence on the sampling design parameters via a misspecified model.

In the context of our motivating AD research, if our hypothesis is true it would imply that one would have to a priori specify the functional form of baseline ADAS-11 in order to avoid dependence of the estimand on the sampling design. Failure to do so could reasonably lead to lack of scientific reproducibility and replicability. The functional form of a continuous covariate is not obvious, however, and since interests lies in conducting inference rather than data-driven modeling we might decide to fit a first-order linear trend to relate ADAS-11 to the log-hazard for time to dementia progression. As we will see later in the manuscript, the observed relationship is indeed not linear in nature. If interest lies in the association, changing the a priori specified model in a post-hoc fashion to fit the observed data will inflate the Type I error rate. It is therefore necessary to understand precisely how model misspecification impacts the resulting estimator in this case and to correct, if possible, any deleterious impacts of the misspecification.

In this manuscript, we begin with a brief introduction of the NCC design. We then show the dependence of the estimand on the sampling proportion (which in finite samples leads to a dependence on M). In Section 2.3, we present the asymptotic distribution of the proposed estimator and provide finite-sample estimators for the variance. We include simulation studies for the proposed estimator and end with an application of the proposed estimator to data stemming from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study[11] to investigate the association between the ADAS-11 and time to progression of AD.

2 | METHODOLOGY

2.1 | Partial likelihood estimator under the usual NCC design

Let T_i, C_i, and Z_i denote the true event time, censoring time, and covariate for subject i, respectively. The observed time is $X_i = \min(T_i, C_i)$ and δ_i is an indicator for whether subject i experiences an event. Under the NCC design, M subjects are randomly sampled from everyone who is still at risk at each event time. In this setting, the estimating equation takes the form $U_{NCC}(\beta) = \sum_{i=1}^{n} \int_{t=0}^{\infty} \left\{ Z_i - \frac{S^{(1)}_{NCC}(\beta,t)}{S^{(0)}_{NCC}(\beta,t)} \right\} dN_i(t) = 0$ where $N_i(t) = I(X_i \leq t, \delta_i = 1)$, $S^{(r)}_{NCC}(\beta,t) = n^{-1} \sum_{j=1}^{n} Z_j^r \tilde{Y}_j(t) \exp(Z_j\beta)$

and $\tilde{Y}_j(t)$ is an indicator for whether subject j is in the NCC risk set at time t (ie, the subject either experienced an event at time t or was sampled as a control). For all times during which there are at least $M+1$ subjects at risk, we have that $\sum_{j=1}^{n} \tilde{Y}_j(t) = M+1$.

In previous work focusing on a binary predictor of interest it has been shown[10] that if the NCC sampling scheme is utilized and the proportionality assumption is not satisfied, the PL estimator is consistent for the solution to

$$\int_0^\infty E_Z \left\{ E_{Z^*|Z} \left(f_T(t|Z) S_c(t|Z) \gamma(a, Z^*, t) \times \left[Z - \frac{E_Z\{Z S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}}{E_Z\{S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}} \right] \right) \right\} dt = 0, \tag{1}$$

where $f_T(t|Z)$ and $S_T(t|Z)$ denote the density and survival function for the failure times, respectively, and $S_C(t|Z)$ denotes the survival function for the censoring times. Moreover, $\gamma(a, Z^*, t) = \frac{a \cdot S_T(t|Z^*) S_C(t|Z^*)}{S_T(t) S_C(t)}$ represents the probability of sampling a control with covariate value Z^* if an event is observed at time t with $\lim_{M,n\to\infty} M/n = a$. As described in our previous work, as $M, n \to \infty$, Equation (1) simplifies to

$$\int_0^\infty E_Z \left\{ a f_T(t|Z) S_C(t|Z) \times \left[Z - \frac{E_Z\{Z S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}}{E_Z\{S_T(t|Z) S_C(t|Z) \exp(Z\beta)\}} \right] \right\} dt = 0,$$

which results in the same estimand as that of the FC PL estimator. In finite samples, however, the estimates obtained using the NCC design will depend on M, the number of controls sampled at each event time. Under the NCC design, we alter the risks sets compared to those of the FC and, as a result, we also change the observed censoring distribution. As seen in Equation (1), the censoring distribution determines the weight given to each event time and therefore influences the estimates. Note that the censoring distribution (and in turn the weighting scheme) will differ for different values of M. When the PH assumption is satisfied, the weighting scheme will not impact the estimates because the relative hazards do not vary with time. When the functional form of a predictor is misspecified, however, we no longer satisfy the proportionality assumptions and the weights given to each event time will effect the estimates.

2.2 | Recovering the FC estimand for a single continuous variable with misspecified functional form

As described in the previous section, when the PH model is misspecified the expectation of the sampling distribution of the usual NCC estimator will depend on the number of controls sampled at each event time. This result is due to the changing censoring distribution, and hence, potentially changing covariate distributions of subjects included in the risk sets of the NCC design relative to the FC analysis. In order to mimic the risk-set representation of the FC, we propose imputing the values of subjects in the FC risk sets who were not included in the NCC sample. Because controls are randomly sampled at each event time, we can use information from previous risk sets to learn about subjects who are still at risk in the FC. Under the NCC design, we have full covariate information for subjects sampled into the NCC sample. We also know the at-risk status for all subjects in the FC at each event time. We can therefore use this information to impute the covariate values for subjects in the FC who were not sampled. Using the new risk sets with the imputed values, we can obtain estimates of the coefficients.

Proposition 1. *Let $\tilde{R}(t)$ be the risk set including the imputed values at time t and assume that the values of covariates in $\tilde{R}(t)$, $\tilde{Z}_j$, are sampled from the same distribution as those in $R(t)$, the FC risk set at time t. Denote β to be the estimand corresponding to the FC PL estimator and let $\hat{\beta}$ be the solution to*

$$U(\hat{\beta}) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i - \frac{n^{-1} \sum_{j\in\tilde{R}(t)} \tilde{Z}_j \exp(\hat{\beta}\tilde{Z}_j)}{n^{-1} \sum_{j\in\tilde{R}(t)} \exp(\hat{\beta}\tilde{Z}_j)} \right\} dN_i(t) = 0. \tag{2}$$

Then $\hat{\beta} \xrightarrow{P} \beta$.

The proof of Proposition 1 can be found in the Appendix. Note that subjects in $\tilde{R}(t_j)$ are the same as those in $R(t_j)$. We used different notation for both to emphasize that our proposed method relies on imputation of the unknown covariates for subjects not in the NCC sample. As stated earlier, for the result to hold we need the imputed values to be drawn from

the same distribution as those in the FC risk set at time t. While the covariate values can be imputed in several ways, one way to do so is via Algorithm 1. In this setting we estimate $\mu(t)$, the mean covariate value, for subjects in the risk set at each event time and calculate the mean squared error, $\sigma^2_{MSE}(t)$. To obtain $\hat{\mu}(t)$, we start by calculating the sample mean for the first event times (five in our example). The number selected here can differ and depends on the number of event times required to fit the natural spline. When fitting the natural spline, we include subjects sampled for previous event times only for the time at which they were sampled. Once we obtain $\hat{\mu}(t)$, we impute covariate values for subjects not in the NCC risk set (but who are still at risk in the FC) by randomly drawing values for the predictor of interest from a $N(\hat{\mu}(t), \sigma^2_{MSE}(t))$ distribution where $\hat{\mu}(t)$ represents the estimated mean covariate value at time t.

Algorithm 1. Imputation approach for the univariate setting

D: number of events
$t_j, j = 1, \cdots, D$: ordered event times
R_j: risk set at time t_j under the FC
$\tilde{R}_j$: risk set at time t_j including the imputed values
M: number of controls sampled at each event time
s_0: is the intercept
$s(t)$: a natural spline with evenly spaced knots
$\mu(t)$: the mean covariate value at time t
z_{kc}: predictor of interest for subject k sampled at time t_c
procedure IMPUTATION OF THE PREDICTOR OF INTEREST
  for j in $1 : D$ **do**
    if $j \leq 5$ (Note: 5 was selected to allow enough time points to fit the natural spline). **then**
      Calculate $\hat{\mu}(t_j) = \bar{z} = \frac{1}{\sum_{c=1}^{j}\sum_{k=1}^{n}\tilde{Y}_k(t_c)} \sum_{c=1}^{j}\sum_{k=1}^{n} z_{kc}\tilde{Y}(t_c)$
    else
      Fit $\hat{\mu}(t) = s_0 + s(t)$ using subjects sampled for all $t_k \leq t_j$ to obtain $\hat{\mu}(t_j)$
      $\sigma^2_{MSE}(t_j) = \frac{1}{\sum_{k=1}^{j}|\tilde{R}_I(t_k)|} \sum_{k=1}^{j}\sum_{i=1}^{|\tilde{R}(t_k)|}(\hat{\mu}(t_k) - z_{ik})^2$
      Sample $|R(t_k)| - \sum_{i=1}^{n}\tilde{Y}_i(t_j)$ values from $N(\hat{\mu}(t_j), \sigma^2_{MSE}(t_j))$. These values, together with the original NCC controls, make up $\tilde{R}(t_j)$.
    end if
  end for
  Fit a Cox proportional hazards model using the imputed values.
end procedure

If the sample size is small, we can increase the number of controls sampled at the first event time to obtain better estimates of the mean covariate value at each time while not grossly impacting the overall efficiency of the NCC design. Moreover, all controls from previous risk sets can be used to estimate the means at each event time as long as those controls are still at risk.

Previous work also relies on imputation of subjects not sampled into the NCC risk sets.[10] However, in the binary case, the imputed value can only take on two values so estimating the number of subjects in each group is sufficient. When a continuous covariate is of interest, we must account for the variability in the covariate values (as is proposed in Algorithm 1) so that the imputed values are representative of the FC risk set.

2.3 | Asymptotic properties and estimation of the variance

In this section, we provide the asymptotic properties of the proposed estimator and introduce a finite-sample variance estimator.

Proposition 2. *Let $\hat{\beta}$ denote the solution to Equation (2). Suppose that $P(Y_i(\tau) > 0) > 0$ and let β denote the estimand corresponding to the FC PL estimator. If values in $\tilde{R}(t_j)$ are drawn from the same conditional distribution as those in $R(t_j)$, then $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, A^{-1}BA^{-1})$ where $A = \lim_{n\to\infty} A_n$, $B = \lim_{n\to\infty} B_n$, with $A_n(\beta) = n^{-1}\sum_{i=1}^{n}\delta_i\rho(X_i)(1-\rho(X_i))$, δ_i is an*

indicator for whether subject i experienced an event, X_i is the observed time for subject i, $\rho(X_i) = \frac{n^{-1}\sum_{j=1}^{n}\tilde{Y}_j(X_i)Z_j\exp(Z_j\beta)}{n^{-1}\sum_{j=1}^{n}\tilde{Y}_j(X_i)\exp(Z_j\beta)}$ and $\tilde{Y}_j(X_i)$ is an indicator for whether subject j was originally sampled to be in the NCC risk set at time X_i. Further, $B_n(\beta) = \sum_{j=1}^{D}\tilde{U}_j^*(\beta)\tilde{U}_j^*(\beta)^T$, where $t_1, t_2, \ldots, t_D$ are the unique event times and

$$\tilde{U}_j^*(\beta) = n^{-1}\sum_{i=1}^{n}\left[\delta_i\left\{Z_i - \frac{\sum_{k\in\tilde{R}(t_j)}\tilde{Z}_k\exp(\beta\tilde{Z}_k)}{\sum_{k\in\tilde{R}(t_j)}\exp(\beta\tilde{Z}_k)}\right\} - \frac{Z_i\exp(\beta Z_i)}{\sum_{k\in\tilde{R}(t_j)}\tilde{Z}_k\exp(\beta\tilde{Z}_k)} + \exp(\beta Z_i)\frac{\sum_{k\in\tilde{R}(t_j)}\tilde{Z}_k\exp(\beta\tilde{Z}_k)}{\{\sum_{k\in\tilde{R}(t_j)}\exp(\beta\tilde{Z}_k)\}^2}\right]. \quad (3)$$

The proof of Proposition 2 can be found in the Appendix. Based on the asymptotic properties of our estimator, we find that the finite-sample variance can be estimated using $\widehat{\text{Var}}(\hat{\beta}) = n^{-1}A_n^{-1}(\hat{\beta})B_n(\hat{\beta})A_n^{-1}(\hat{\beta})$. Notice that A_n is the variance under the usual NCC design when the model is correctly specified. B_n represents the true variance and accounts for imputation of the risk sets through a Taylor expansion.

2.4 | Incorporating adjustment covariates

So far we have introduced an estimator that recovers the FC results in the univariate setting. In observational studies, however, we often adjust for potential confounding variables to isolate the association of interest. As before, the imputation approach can be selected by the user, but in this manuscript we use a hot-deck multiple imputation approach.[12] Imputation of the risk sets can be accomplished as in Algorithm 2.

Algorithm 2. Imputation approach for the multivariate setting

D: number of events
$t_j, j = 1, \cdots, D$: ordered event times
R_j: risk set at time t_j under the FC
$\tilde{R}_j$: risk set at time t_j including the imputed values and the originally sampled NCC controls
$\underline{R}_j$: the risk set at time t_j under the NCC design
M: number of controls sampled at each event time
$s(t)$: a natural spline with evenly spaced knots
$\mu(t)$: the mean covariate value at time t
z_{ik1}: predictor of interest for subject i sampled in the NCC sample at time t_k
$\tilde{z}_{ik1}$: imputed value for the predictor of interest for subject i at time t_k
p: the number of covariates in the model
l: the number of previous risk sets to consider for hot-deck imputations
procedure IMPUTATION OF CONFOUNDING VARIABLES (DONE AFTER ALGORITHM 1)
    for j in 1:D **do**
        for m in 1:length($\tilde{R}(t_j)$) **do**
            if $\tilde{Y}_m(t_j) \neq 1$ **then** Find min $|\tilde{z}_{mj1} - z_{hj1}|$ for $h \in \cup_{k=(j-l)}^{j}\underline{R}(t_k)$.
                Let z_1^* be z_{hj1} for $h \in \cup_{k=(j-l)}^{j}\underline{R}(t_k)$ with the smallest absolute difference.
                Impute values of $\tilde{z}_{mj2}, \cdots, \tilde{z}_{mjp}$ using $z_2^*, \cdots, z_p^*$
            end if
        end for
    end for
    Fit a Cox PH model with the imputed subjects.
end procedure

The estimating function in this setting takes the form $U_{HD}(\beta) = \sum_{i=1}^{n}\int_{t=0}^{\infty}\left\{\vec{Z}_i - \frac{\tilde{S}_{HD}^{(1)}(\beta,t)}{\tilde{S}_{HD}^{(0)}(\beta,t)}\right\}dN_i(t)$ where $\tilde{S}_{HD}^{(r)} = n^{-1}\sum_{j\in\tilde{R}(t_j)}\vec{Z}_j^r\exp(\vec{Z}_j\beta)$ and $\vec{Z}_j$ is the vector of covariates. $\tilde{R}(t_j)$ represents the covariate values for the imputed risk sets which include the originally sampled subjects and the imputed subjects. For subjects who were not originally sampled into the NCC sample, we draw values for the predictor of interest from a $N(\hat{\mu}(t), \sigma^2_{MSE}(t))$ as in the univariate setting. We match each of the imputed values to the subject from the l previous risk sets in the NCC sample with the

closest value to the imputed value (l is selected by the user). The values of the adjustment variables of the selected subject are used to impute the values for the imputed subject. If more than one NCC subject can be used to impute the covariate values, we randomly sample one subject from eligible subjects. As stated earlier, l can be selected by the user. If event times are far apart, we recommend selecting a small l because neighboring event times may have drastically different risk sets. If event times are close, a larger l may be selected. In fact, if event times are close, l can be selected to include all previous risk sets. When sampling subjects to impute covariate values, however, one must ensure that the subjects selected are still at risk during the current event time.

Some covariate information, such as demographic information, may be easily available for all study participants. If this is the case, we may use this information (instead of estimating the mean covariate values) along with the hot-deck imputation method to impute covariate values that may be difficult or expensive to collect.

3 | EMPIRICAL PERFORMANCE

3.1 | Univariate results

We begin by presenting simulation results for the univariate setting. Table 1 illustrates the performance of the usual NCC PL estimator and our proposed estimator when the functional form is misspecified. Values for the predictor of interest were sampled from a N($\mu = 1.5$, $\sigma = 0.5$) distribution. The true hazard function takes the form $\lambda(t) = \exp(\log(0.1) + \log(1.25)z + \log(0.5)z^2)$, failure times were drawn from Exp(rate = $\lambda(t)$), and censoring times were drawn from a Unif(0, 6) distribution. Observed event times were taken to be the minimum of the event and censoring times. Generating data in this way led to approximately 90% censoring and we included 2000 subjects in each of the 200 simulations. We considered NCC samples with one to four controls per event time. We did not consider more than four controls since in practice people often use up to four controls.[13] Moreover, it has been shown that using more than four controls does not provide a large benefit in terms of efficiency and power.[14] For each NCC sample, we sampled 60 controls at the first event time regardless of M. The additional sampling of controls was only performed at the first event time to supplement the risk set and was done for the usual NCC PL estimator and for the proposed estimator. While we selected 60 controls here, in practice the number of additional controls to be sampled at the first event should depend on variability of the predictor of interest, desired precision for the imputation estimate, and feasibility of data collection for the additional controls. Specifically, we recommend that one can select the number of additional controls to be used at the first event time by considering the desired precision of the estimated mean of the predictor of interest and weighing this against the relative increase of the sample size for the NCC design relative to that of the FC.

To illustrate the performance of the estimators under a misspecified functional form, we fit a model of the form $\lambda(t) = \lambda_0(t)\exp(\beta z)$. The analytic variance estimates for the usual NCC PL estimator were obtained using the robust variance estimator while those of the proposed estimator were obtained using the estimator presented in Section 2.3.

Table 1 shows that the NCC PL estimator performs poorly when the model is not specified correctly and that the results obtained depend on the value of M, the number of controls sampled at each event time. The proposed estimator, however, reduces the bias relative to the FC estimator from approximately 18% to less than 1% when $M = 1$ and from approximately 9% to 1% when $M = 4$. We compared the model fit (to the FC data) using Akaike's Information Criterion (AIC)[15] with the FC data and coefficient estimates from the standard NCC design and the proposed estimator. For brevity, we do not provide the numeric results, but the average AIC based on 200 simulations was lower for the proposed estimator compared to the standard NCC design for all M, indicating better fit. The robust variance estimator for the NCC PL estimator tends to under estimate the variance for smaller values of M. Our proposed sandwich estimator is conservative when $M = 1$ but performs well for $M = 2$ to 4. To assess the robustness of the proposed estimator, we also investigated the performance when the functional form is specified correctly. In this setting, data were generated as in the first scenario, but now the true hazard function takes the form $\lambda(t) = \exp(\log(0.008) + \log(2.5)z)$ and the failure times were drawn from Exp(rate = $\lambda(t)$). These data also had approximately 90% censoring. As seen on the right side of Table 1, when the model is specified correctly the NCC PL estimator estimates the same quantity as the FC estimator. Our proposed estimator also performs well regardless of M, yielding a bias (relative to the FC estimator) between 1% and 2% for all values of M. In this case, the AICs were nearly identical for the proposed and standard NCC estimators. The proposed variance estimator again gives conservative estimates of the variance when $M = 1$, but performs well for $M = 2$ to 4. When the

TABLE 1 200 simulations under a misspecified functional form (left) and a correctly specified functional form (right)

	Misspecified model					Correctly specified model				
	N	Coeff. Est.	% Est. Bias	Emp. Var.	An. Var.	N	Coeff. Est.	% Est. Bias	Emp. Var.	An. Var.
FC	2000.00	−1.4064	0.00	0.0167	0.0182	2000.00	0.9311	0.00	0.0239	0.0215
NCC										
M = 1	403.49	−1.6614	18.13	0.0808	0.0574	411.25	0.9272	−0.42	0.0595	0.0318
M = 2	544.65	−1.5686	11.53	0.0521	0.0371	553.62	0.9194	−1.26	0.0418	0.0254
M = 3	668.32	−1.5388	9.42	0.0324	0.0298	679.66	0.9237	−0.80	0.0332	0.0236
M = 4	778.99	−1.5308	8.85	0.0376	0.0271	791.18	0.9278	−0.35	0.0346	0.0228
Proposed estimator										
M = 1	403.49	−1.4136	0.52	0.0776	0.1728	411.25	0.9156	−1.67	0.0675	0.1300
M = 2	544.65	−1.4065	0.01	0.0505	0.0773	553.62	0.9147	−1.76	0.0524	0.0627
M = 3	668.32	−1.4075	0.08	0.0422	0.0529	679.66	0.9185	−1.36	0.0426	0.0472
M = 4	778.99	−1.4219	1.10	0.0417	0.0437	791.18	0.9167	−1.55	0.0391	0.0395

Note: The NCC samples included 60 controls at the first event time, regardless of M. Empirical variance and analytic variance estimates are also provided.

functional form is specified correctly, we observe a small loss in efficiency. However, this loss is nearly negligible and we have the added benefit that if the functional form is not specified correctly we still estimate the same quantity as that of the FC.

3.2 | Multivariate results

When using data from observational studies, it is almost always necessary to adjust for confounding variables. In Section 2.4, we described a hot-deck imputation approach to impute the values for the missing covariates. In this section, we present simulation results when adjusting for a confounding variable. As before, we consider two scenarios- one in which the functional form of the predictor of interest is misspecified and one in which the functional form is correctly specified. In this setting, we assume that no covariate information is available for subjects not sampled into the NCC sample.

Table 2 presents the results for the multivariate setting with a misspecified functional form. The predictor of interest and the confounding variable are distributed as $Z_1 \sim N(\mu = 1.5, \sigma = 0.5)$ and $Z_2 \sim N(\mu = 0 + 2 \cdot I(x_1 \geq 1.6), \sigma = 1.5)$, respectively. The true hazard function for this scenario takes the form $\lambda(t) = \exp(\log(0.075) + \log(1.25)z_1 + \log(0.5)z_1^2 + \log(1.35)z_2)$ and failure times were drawn from Exp(rate $= \lambda(t)$). As before, censoring times were drawn from Unif(0, 6) and the observed times were the minimum of the observed and censoring times, yielding approximately 90% censoring. We sampled 60 controls at the first event time regardless of M and the model is assumed to take the form $\lambda(t) = \lambda_0(t)\exp(\beta_1 z_1 + \beta_2 z_2)$. We find that when the model is misspecified, the usual NCC PL estimator produces biased coefficient estimates (when compared to the FC estimates) and the estimates obtained depend on the number of controls sampled at each event time. The proposed estimator, however, yields results similar to those of the FC PL estimator. When $M = 1$, the bias relative to the FC estimator is approximately 14% under the usual NCC PL estimator. This is reduced to approximately 4% when the proposed estimator is used. In this setting, the proposed variance estimator gives conservative estimates of the variance, but performance of the variance estimator improves as M increases. While the estimates provided by our variance estimator can be conservative, it should be noted that those provided by the robust variance estimator for the NCC PL estimator tend to give anti-conservative estimates of the variance. The bottom portion of Table 2 presents the results for the usual NCC PL estimator and our proposed estimator when the model is correctly specified. Data were generated as in the previous scenario, but the true hazard function takes the form $\lambda(t) = \exp(\log(0.0125) + \log(2.5)z_1 + \log(0.5)z_2)$, with failure times being drawn from Exp(rate $= \lambda(t)$). In this

TABLE 2 200 simulations for hot-deck imputations under misspecification of the functional form (top) and a correctly specified functional form (bottom)

	N	$\hat{\beta}_1$	% Est. Bias	Emp. Var.	$\hat{\text{Var}}(\hat{\beta}_1)$	$\hat{\beta}_2$	% Est. Bias	Emp. Var.	$\hat{\text{Var}}(\hat{\beta}_2)$
Misspecified functional form									
Full Cohort	2000.00	−1.4446	0.00	0.0226	0.0215	0.2753	0.00	0.0024	0.0022
NCC									
M = 1	383.68	−1.6510	14.29	0.0926	0.0685	0.2941	6.83	0.0067	0.0041
M = 2	517.73	−1.6080	11.31	0.0671	0.0442	0.2883	4.72	0.0046	0.0031
M = 3	635.58	−1.5923	10.22	0.0540	0.0372	0.2896	5.19	0.0037	0.0028
M = 4	742.26	−1.5658	8.39	0.0478	0.0331	0.2900	5.34	0.0034	0.0026
Proposed estimator									
M = 1	383.68	−1.4964	3.59	0.1241	0.2869	0.2912	5.78	0.0086	0.0186
M = 2	517.73	−1.5055	4.22	0.0795	0.1188	0.2918	5.99	0.0063	0.0085
M = 3	635.58	−1.4596	1.04	0.0481	0.0768	0.2811	2.11	0.0045	0.0058
M = 4	742.26	−1.4707	1.81	0.0492	0.0617	0.2846	3.38	0.0042	0.0048
Correctly specified functional form									
Full Cohort	2000.00	0.9152	0.00	0.0076	0.0068	−0.6847	0.00	0.0074	0.0066
NCC									
M = 1	408.57	0.9366	2.34	0.0241	0.0175	−0.7046	2.91	0.0208	0.0145
M = 2	550.38	0.9234	0.90	0.0165	0.0117	−0.6876	0.42	0.0154	0.0101
M = 3	675.37	0.9219	0.73	0.0130	0.0099	−0.6908	0.89	0.0130	0.0089
M = 4	786.52	0.9189	0.40	0.0124	0.0089	−0.6907	0.88	0.0115	0.0081
Proposed Estimator									
M = 1	408.57	0.9349	2.15	0.0301	0.0974	−0.7354	7.40	0.0360	0.0882
M = 2	550.38	0.9122	−0.33	0.0247	0.0389	−0.7007	2.34	0.0224	0.0339
M = 3	675.37	0.9081	−0.78	0.0159	0.0251	−0.6949	1.49	0.0164	0.0225
M = 4	786.52	0.9093	−0.64	0.0188	0.0201	−0.6869	0.32	0.0158	0.0178

Notes: The NCC samples included 60 controls at the first event time, regardless of M. Empirical and analytic variance estimates are also provided.

setting, the fitted model takes the same form as the true data-generating mechanism. The usual NCC PL estimator and the proposed estimator perform similarly, both having a small bias relative to the FC estimator regardless of the selected M. In this setting, the proposed variance estimator is again conservative when $M = 1$, but its performance improves as M increases.

4 | APPLICATION TO ADNI EXAMPLE

In this section, we apply the proposed estimator to data from the ADNI[11] to investigate the association between the ADAS-11 at baseline and time to progression to AD. The ADAS-11 is a cognitive test used to evaluate cognition and behavioral function, both of which are affected by AD.[1] We had 974 participants in our analysis. These participants had ADAS-11 and CSF Aβ at baseline and did not have a diagnosis of AD dementia at baseline. In our analysis, progression was defined as a stable clinical diagnosis of dementia or a diagnosis of dementia at the last visit. Based on this definition, approximately 15% of subjects experienced an event. Baseline characteristics of our sample can be found in Table 3. The mean age in the sample was 72.9 years, approximately 45% of participants were female, and approximately 42% of subjects had at least one APOE e4 allele.

TABLE 3 Baseline demographics for subjects in our study

Characteristic	mean (sd) or n(%)
N	974
Progressors (to AD dementia)	152 (15.6%)
Age	72.9 (7.0)
Female	442 (45.4%)
White	909 (93.3%)
≥1 APOE e4 allele	405 (41.6%)
ADAS-11	8.5 (4.6)
Mini-Mental State Examination	28.2 (1.7)
Education	16.2 (2.7)
Aβ	182.1 (53.7)

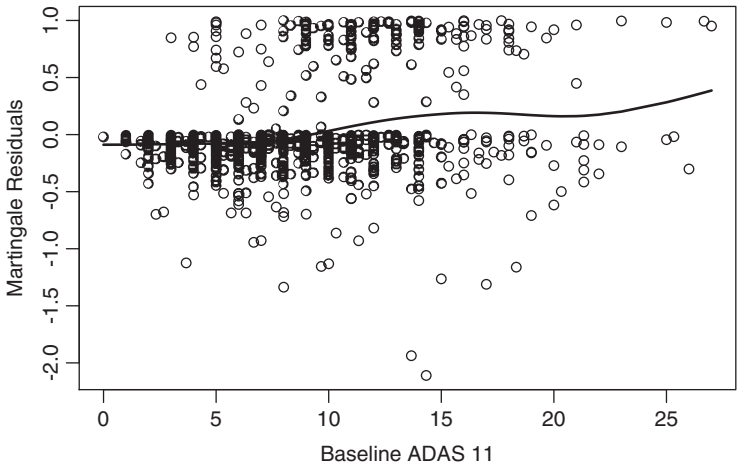

FIGURE 1 Martingale residuals plotted against baseline ADAS-11. The solid line represents a smoother

As stated before, the goal of this analysis is to investigate the association between ADAS-11 at baseline and time to progression to AD. In this case, one may a priori specify a model that assumes a linear relationship between ADAS-11 and time to progression. Using the Martingale residuals[16], we found that the functional form of ADAS-11 at baseline is not linear (Figure 1). Because the goal of the study was to investigate an association, changing the a priori selected model to fit the observed data could increase the Type I error rate and therefore is not recommended. Instead, we fit a first-order trend to investigate the behavior of the NCC design and the proposed estimator in this setting.

We fit the Cox proportional hazards model to the entire sample to obtain the FC estimates. We then obtained 200 NCC samples for each value of M and applied the PL estimator and the proposed estimator as if we only had full covariate information for subjects in the NCC sample. We a priori decided to sample 60 controls at the first event time for all NCC samples, regardless of M. This number of additional controls provided enough statistical information so that a 95% confidence interval for the mean ADAS-11 based upon data at the first event time had a total width of approximately 2 points, yet resulted in a minimal relative increase in total sample size. While this decision was made a priori, followup sensitivity analyses showed that the proposed estimator performed well for this application even when considering only 10 controls at the first event time. All models were adjusted for age, education, race, the presence of at least one APOE e4 allele, gender, and baseline CSF Aβ levels. Because APOE e4 status and CSF Aβ levels would be the most difficult covariates to collect, we applied the NCC sampling scheme as if these measurements were only available for subjects in the NCC sample. Demographic information, on the other hand, is easily collected for all study participants. Therefore, we assume that demographic information is available for all study participants, even if they were not sampled into the NCC sample. We used the hot-deck imputation method to impute values of APOE e4 and Aβ for participants not sampled into the NCC sample. Mahalanobis distance[17] was used to match subjects with missing values to sampled controls. When the covariance matrix was singular, we used Euclidean distance. Table 4 presents the coefficient estimates for a difference of five points in baseline ADAS-11, our predictor of interest, as well as for APOE e4 and CSF Aβ. Under the FC PL estimator,

TABLE 4 Mean coefficient estimates for ADAS-11, APOE e4, and Aβ based on 200 nested case-control samples from the FC data

		ADAS 11			APOE e4			Aβ		
	N	(HR) 5 points	% Est. Bias	Var. Est.	(HR)	% Est. Bias	Var.	(HR) 50 pg/ml	% Est. Bias	Var.
Full Cohort	974.00	0.651 (1.92)	0	0.008	0.099 (1.10)	0.00	0.046	−0.537 (0.58)	0.00	0.010
Usual NCC										
M = 1	310.03	0.748 (2.11)	14.75	0.016	0.248 (1.28)	150.71	0.046	−0.583 (0.56)	8.57	0.020
M = 2	393.55	0.770 (2.16)	18.22	0.013	0.249 (1.28)	151.72	0.044	−0.572 (0.56)	6.52	0.010
M = 3	462.10	0.764 (2.15)	17.32	0.012	0.240 (1.27)	142.73	0.043	−0.570 (0.57)	6.15	0.010
M = 4	518.22	0.747 (2.11)	14.69	0.011	0.201 (1.22)	102.93	0.042	−0.562 (0.57)	4.66	0.010
M = 5	564.60	0.724 (2.06)	11.16	0.010	0.182 (1.20)	83.54	0.042	−0.560 (0.57)	4.28	0.010
Proposed Est.										
M = 1	310.03	0.677 (1.97)	3.99	0.064	0.159 (1.17)	61.01	0.201	−0.583 (0.56)	8.57	0.070
M = 2	393.55	0.668 (1.95)	2.61	0.032	0.103 (1.11)	3.74	0.118	−0.592 (0.55)	10.24	0.030
M = 3	462.10	0.660 (1.94)	1.37	0.023	0.119 (1.13)	19.90	0.091	−0.580 (0.56)	8.01	0.030
M = 4	518.22	0.657 (1.93)	0.92	0.019	0.092 (1.10)	−7.58	0.079	−0.575 (0.56)	7.08	0.020
M = 5	564.60	0.656 (1.93)	0.75	0.016	0.072 (1.07)	−26.97	0.073	−0.583 (0.56)	8.57	0.020

Notes: 60 controls were sampled at the first event time, regardless of M.

we estimate that comparing two subpopulations that differ by five points in baseline ADAS-11, the risk of progression to AD is approximately 92% higher for the group with higher ADAS-11. When we estimate the coefficients using the PL estimator and the NCC sampling scheme, we find that as in the simulated examples, the estimates are different than those obtained using the FC PL estimator and that these differ by the value of M. The bias relative to the FC estimator in this case ranges from 11% to 18% compared to the FC PL estimates. Applying our proposed estimator reduces this to between 0.75% and 4% while using the same sample sizes as the usual NCC design. Notice also that, as expected, the variance estimates for the proposed estimator are larger than those for the usual NCC PL estimator and that both are larger than those of the FC PL estimator.

To calculate the usual NCC PL estimator and the proposed estimator, we would only have to collect full covariate information for subjects in the NCC samples. That is, we would only have to perform genotype testing and process CSF samples for subjects who progressed to AD or those who were sampled as controls. This reduces costs associated with these tests and allows us to use CSF samples to answer other questions that we may have about AD.

5 | DISCUSSION

It has been shown that the expectation of the sampling distribution of the usual NCC estimator will depend on the number of controls sampled at each event time when the PH assumption is violated. Previous work has proposed an estimator that yields the same results as those obtained using the FC data when the predictor of interest is binary.[10] In this scenario, the functional form of the covariate of interest is specified correctly, but the effect of the covariate is assumed to be constant when in reality it varies with time. In our current work, we consider the performance of the PL estimator under the NCC design when the effect of the covariate is constant over time, but the functional form is misspecified. We again observe that the estimates obtained using the PL estimator under the NCC design also depend on the number of controls sampled at each event time. We therefore propose a method that estimates the same quantity as the FC PL estimator under misspecification of the functional form, while only using the information from the usual NCC design. By only requiring full covariate information from the NCC sample, our proposed estimator maintains the reduction in costs afforded by the NCC design. The proposed estimator recovers the FC estimates when the model is misspecified, both in the univariate and multivariate scenarios. When the model is specified correctly, the proposed estimator still recovers the FC estimates regardless of M. While the proposed estimator increases the bias relative to the FC estimator for $M = 1$ in the multivariate setting, it should be noted that M is usually larger than one in practice. Our proposed finite-sample variance estimator performs well for M greater than one but yields conservative estimates when $M = 1$.

The purpose of this manuscript was to show the dependence on M in the standard NCC design (as proposed by Thomas (1977)[4]), when the model is mis-specified. While we present one fix, it should be noted that the Samuelsen estimator[18] might also be considered as an alternative approach via weighting. Moreover, it is known that the estimand corresponding to the FC PL estimator depends on the censoring distribution when the model is misspecified.[19-21] In our previous work, we introduced an estimator for the FC censoring distribution that only requires the NCC sample. The estimator for the censoring distribution can also be used to reweight the estimating function to yield a censoring-robust estimator in this setting.[10]

The NCC design provides great reduction in costs when the event of interest is rare. When the model is specified correctly, the NCC design estimates the same quantity as the FC PL estimator. If the functional form is misspecified, however, the results obtained from the usual NCC estimator depend on the number of controls sampled at each event time. The proposed estimator uses the same information as the usual NCC design but recovers the FC results even when the functional form is misspecified. We therefore recommend application of the proposed estimator since the estimator performs well even when the functional form is specified correctly and still affords the cost reductions offered by the NCC sampling scheme. When using our estimator, however, we do recommend using M larger than one (which is commonly done in practice).

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1321846 and the National Institutes of Health under Grant Nos. R01AG053555 and P30AG066519. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID


Michelle M. Nuño https://orcid.org/0000-0003-2031-929X

REFERENCES

1. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. *Am J Psychiatry*. 1984;141(11):1356-1364.
2. Corder E, Saunders A, Strittmatter W, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 1993;261(5123):921-923.
3. Nuño MM, Gillen DL, Dosanjh KK, Brook J, Elashoff D, Ringman JM, Grill JD. Attitudes toward clinical trials across the Alzheimer's disease spectrum. *Alzheimer's Research & Therapy volume*. 2017;9(1):81.
4. Thomas D. Addendum to a paper by FDK Liddel JC McDolad and DC Thomas. *J Royal Stat Soc Ser A*. 1977;140:483-485.
5. Nuño MM, Gillen DL. Alternative sampling designs for time-to-event data with applications to biomarker discovery in Alzheimer's disease. *Handbook of Statistics*. Vol 36. Amsterdam, Netherlands: Elsevier; 2017:105-166.
6. Ioannidis JP. Why most published research findings are false. *PLoS Med*. 2005;2(8):e124.
7. Gelman A, Loken E. *The Garden of Forking Paths: Why Multiple Comparisons can be a Problem, Even When There is no "Fishing Expedition" or "p-hacking" and the Research Hypothesis was Posited Ahead of Time*. New York, NY: Department of Statistics Columbia University; 2013.
8. Motulsky HJ. Common misconceptions about data analysis and statistics. *Br J Pharmacol*. 2015;172(8):2126-2132.
9. de Groot AD. The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychol*. 2014;148:188-194.
10. Nuño MM, Gillen DL. On estimation in the nested case-control design under non-proportional hazards. *Under Review*; 2019.
11. Alzheimer's disease neuroimaging initiative. http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowled%gement_List.pdf.
12. Fellegi IP, Holt D. A systematic approach to automatic edit and imputation. *J Am Stat Assoc*. 1976;71(353):17-35.
13. Ernster VL. Nested case-control studies. *Prev Med*. 1994;23(5):587-590.
14. Taylor JM. Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Stat Med*. 1986;5(1):29-36.
15. Akaike H, Parzen E, Tanabe K, Kitagawa G. Information theory and an extension of the maximum likelihood principle. *Selected Papers of Hirotugu Akaike*. New York: Spring Science and Business Media; 1998:199-213.

16. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. Berlin, Germany: Springer Science & Business Media; 2005.
17. Mahalanobis PC. *On the Generalized Distance in Statistics*. Jatani, India: National Institute of Science of India; 1936.
18. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. *Biometrika*. 1997;84(2):379-394.
19. Struthers CA, Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika*. 1986;73(2):363-369. http://dx.doi.org/10.1093/biomet/73.2.363.
20. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics*. 2000;1(4):423-439.
21. Boyd AP, Kittelson JM, Gillen DL. Estimation of treatment effect under non-proportional hazards and conditionally independent censoring. *Stat Med*. 2012;31(28):3504-3515.
22. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. Vol 360. Hoboken, NJ: John Wiley & Sons; 2011.
23. Eriksson F, Martinussen T, Nielsen SF. Large sample results for frequentist multiple imputation for Cox regression with missing covariate data. *Ann Inst Stat Math*. 2019;72:1-28.
24. Lu K, Tsiatis AA. Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*. 2001;57(4):1191-1197.
25. Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann Stat*. 1992;20(4):1903-1928.
26. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc*. 1989;84(408):1074-1078.
27. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. *Ann Stat*. 1982;10(4):1100-1120.

How to cite this article: Nuño MM, Gillen DL. Robust estimation in the nested case-control design under a misspecified covariate functional form. *Statistics in Medicine*. 2021;40:299–311. https://doi.org/10.1002/sim.8775

APPENDIX

Proof. Let T_i, C_i, and $X_i = \min(T_i, C_i)$ be the event, censoring, and observed times for subject i, respectively. $N_i(t) = I(X_i \leq t, \delta_i = 1)$ is a right-continuous counting process. Define $S^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^{n} Z_i^r \exp(\beta Z_i) Y_i(t)$ $(r = 0, 1, 2)$ where $Y_i(t) = I(X_i \geq t)$. Let $s^{(r)}(\beta, t) = \lim_{n\to\infty} S^{(r)}(\beta, t)$ and $\tilde{S}^{(r)}(\beta, t) = n^{-1}\sum_{i=1}^{n} \tilde{Z}_i^r \exp(\beta \tilde{Z}_i) Y_i(t)$ where $\tilde{Z}_i = Z_i$ if subject i was originally sampled into the NCC sample and the imputed value otherwise. For subjects not in the original NCC sample, $\tilde{Z}_i \sim N(\hat{\mu}(t), \sigma^2_{MSE}(t))$ where $\hat{\mu}(t)$ is an estimate of $E[ZY(t)]$. We prove the asymptotic properties of the proposed estimator using theorem 5.3 of Kalbfleish and Prentice (2011)[22] which implies Rebodello's theorem. This requires that there exists an open neighborhood % of β and $s^{(r)}(\beta, t)$, $r = 0, 1, 2$ defined on $B \times [0, \tau]$ that satisfy the following: (1) $\sup_{\beta\in\%, t\in[0,\tau]} ||\tilde{S}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$; (2) $s^{(0)}(\beta, t)$ is bounded away from 0 for $t \in [0, \tau]$; (3) For $r = 0, 1, 2$, $s^{(r)}(\beta, t)$ is a continuous function of β uniformly in $t \in [0, \tau]$, $s^{(1)}(\beta, t) = \frac{\partial s^{(0)}(\beta,t)}{\partial \beta}$ and $s^{(2)}(\beta, t) = \frac{\partial^2 s^{(0)}(\beta,t)}{\partial \beta^2}$; (4) $\Sigma(\beta, t) = \int_0^\tau \nu(\beta, u) s^{(0)}(\beta, u)\lambda_0(u)du$ is positive definite $\forall \beta \in \%$; (5) Z_i is bounded $\forall t \in [0, \tau]$; (6) $\lambda_0(u)du < \infty$. As in Eriksson et. al (2019)[23] our results require that the imputed values are drawn from the same conditional distribution as the covariates for subjects in the full cohort and that missing values are missing at random. The latter is satisfied by design.

We assume that $P(Y_i(\tau) > 0) > 0$ (ie, there is positive probability that subject i is at risk over the inferential support interval) which implies that conditions (2) and (6) hold. We also assume that conditions (4) and (5) hold. Condition (5) along with the dominated convergence theorem ensures that (3) is also satisfied. For condition (1) to hold, we need $\sup_{\beta\in\%, t\in[0,\tau]} ||\tilde{S}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$. We have that $s^{(r)}(\beta, t) = \mathrm{E}[S^{(r)}(\beta, t)]$ and that $||S^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ by the strong law of large numbers. Now, suppose that $Z \sim f_Z$ and $\tilde{Z} \sim f_Z$. This gives us that $s^{(r)}(\beta, t) = \mathrm{E}[\tilde{S}^{(r)}(\beta, t)]$ and that $\sup_{\beta\in\%, t\in[0,\tau]} ||\tilde{S}^{(r)}(\beta, t) - s^{(r)}(\beta, t)|| \xrightarrow{P} 0$ by the strong law of large numbers. This argument is similar to that used in the work of Lu and Tsiatis (2001).[24]

Equation (2) is a sum over stochastic integrals of a predictable process with respect to a martingale and the predictability of the NCC sampling scheme holds by the same argument used in Goldstein and Langholz (1992).[25] Notice that at each event time, the proposed estimator only considers controls that were sampled into risk sets up to the current time so the proposed estimator maintains predictability. Therefore, theorem 5.3 of Kalbfleisch and Prentice (2011)[22] with the sandwich variance estimator of Lin and Wei (1989)[26] and a Taylor expansion of the estimating function about $s^{(0)}(\beta, t), s^{(1)}(\beta, t)$ and $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} N_i(t)$ implies that $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, A^{-1}BA^{-1})$. $A = \lim_{n\to\infty} A_n$ and $B = \lim_{n\to\infty} B_n$ where $A_n(\beta) = n^{-1}\sum_{i=1}^{n} \delta_i \rho(X_i)(1 - \rho(X_i))$ and $B_n(\beta) = \sum_{j=1}^{D} \tilde{U}_j^*(\beta)\tilde{U}_j^*(\beta)^T$. δ_i is an indicator for whether subject i experienced an event and $\rho(X_i) = \frac{n^{-1}\sum_{j=1}^{n} \tilde{Y}_j(X_i) Z_j \exp(Z_j\beta)}{n^{-1}\sum_{j=1}^{n} \tilde{Y}_j(X_i)\exp(Z_j\beta)}$ where $\tilde{Y}_j(X_i)$ is an indicator for whether subject j was originally sampled

to be in the NCC risk set at time X_i. $\tilde{U}_j^*(\hat{\beta})$ is defined as in Equation (3). The proof of consistency follows from the work of Andersen and Gill (1982).[27] Using the fact that $\sup_{\beta\in\%,t\in[0,\tau]} ||\tilde{S}^{(r)}(\beta,t) - s^{(r)}(\beta,t)|| \xrightarrow{P} 0$ and under the assumption that $\tilde{Z}_i$ and Z_i are drawn from the same distribution, it is easy to show that the log PL of the proposed estimator converges in probability to a concave function maximized at β. ■