Cost-Effective Extreme Case-Control Design Using a Resampling Method

Young Min Kim1 and Jongho Im2

1Department of Statistics, Kyungpook National University, Daegu, Republic of Korea.
2Department of Applied Statistics, Yonsei University, Seoul, Republic of Korea.

ABSTRACT: Nested case-control sampling design is a popular method in a cohort study whose events are often rare. The controls are randomly selected with or without the matching variable fully observed across all cohort samples to control confounding factors. In this article, we propose a new nested case-control sampling design incorporating both extreme case-control design and a resampling technique. This new algorithm has two main advantages with respect to the conventional nested case-control design. First, it inherits the strength of extreme case-control design such that it does not require the risk sets in each event time to be specified. Second, the target number of controls can only be determined by the budget and time constraints and the resampling method allows an undersampling design, which means that the total number of sampled controls can be smaller than the number of cases. A simulation study demonstrated that the proposed algorithm performs well even when we have a smaller number of controls compared with the number of cases. The proposed sampling algorithm is applied to a public data collected for “Thorotrast Study.”

KEYWORDS: cost-effective design, extreme case-control design, nested case-control studies, partial likelihood, resampling method

Introduction

In practice, the large-scale studies are often limited because of time and budget constraints. One popular cohort-based sampling method that reduces the costs of data collection is the nested case-control (NCC) design that selects a set of controls from the risk sets defined in the cohort and then matches them to cases, respectively. For example, Grant et al1 considered the NCC design to reduce the cost and time efforts involved in the covariate ascertainment for the Life Span Study and the Adult Health Study cohorts of the Radiation Effects Research Foundation, in Japan. These NCC sample data having a time-to-event outcome are popularly analyzed using the Cox proportional hazards model based on partial likelihood. Many previous studies discuss the more general estimations and asymptotic properties of parameter estimators obtained from the partial likelihood in the NCC design.2–7

In general, the NCC design is only applicable to retrospective studies because censored subjects in control groups can be selected during the sampling process. However, researchers often want to add new covariates to statistical models (eg, epigenetic studies), such as genetic information and blood test results, as a prospective study. In this case, control groups should be observable subjects in the sense that we can obtain other information from the selected controls. One solution is to sample the controls at the end-point of the observation time period, which will generally be available for further data collection. For example, Yao et al8 considered a new retrospective cohort-based sampling design, called end-point sampling (EPS), to sample the observed controls from a cohort, and then applied the expectation and maximization (EM) method as a parameter estimation. Similarly, Sboner et al9 considered an extreme case-control (ECC) design with a naive statistical method in which the cases were patients who had died from prostate cancer within 10 years after diagnosis and the controls were patients who survived at least 10 years. Although these EPS or ECC designs were initially implemented to improve the efficiency of the design, we can use this sampling idea to reduce the costs of data collection and to observe additional information from the selected samples.

In this article, we propose a new cost-effective extreme case-control design (CECC) utilizing a resampling technique. First, the total number of controls, say \( s \), was determined by the budget constraints. Second, we drew the \( s \) controls at the end-point or the final risk set matched to the final event. Here, \( s \) is an arbitrary number, which means that it can be smaller than the number of cases and is not necessarily a multiple of the number of cases. However, the number of matched controls for each case was conserved as the same number for all of the cases using a resampling method. To adjust the length bias due to the ECC sampling design, we applied Salim et al10 weighted partial likelihood for parameter estimation. Thus, our proposed sampling algorithm inherits the main advantage of the ECC design. Also, the new sampling strategy is very effective in saving time and costs, because the total number of controls can be chosen freely from the constraints.

This manuscript is organized as follows. First, we describe the classical NCC sampling design and our proposed CECC sampling design. We conduct simulation studies to illustrate the relative bias and efficiency of our proposed method in comparison to the classical NCC study design. In the real data example section, we applied our new sampling algorithm to the “Thorotrast Study” data which investigated the relationship between liver cancer occurrence and volume of injected thorotrast.11,12 Finally, the concluding remarks and the limitations of our proposed method are provided.
Methods

Basic setup

Consider a cohort study with a size of \( N \). In the cohort, we observe vectors of variables, \( \{(Y_i, Z_i, \delta_i); i = 1, \ldots, N\} \), where \( Z_i \) is the time-invariant covariate, \( Y_i = \min(T_i, C_i) \) is an observed time, and \( T_i \) and \( C_i \) denote the event time and censoring time for \( i = 1, 2, \ldots, N \), respectively. Also, \( \delta_i \) is an event indicator function given by \( \delta_i = I(T_i \leq C_i) \).

Now assume that we have \( n = \sum_{i=1}^{N} \delta_i \) individuals observed to have an event and call them ‘cases,’ with their event times denoted by \( t_1, t_2, \ldots \). Let \( R(t_j) = \{ i: Y_i \geq t_j \} \) be the risk set, which consists of individuals not being observed as an event and censored up to time \( t_j \). For simplicity, we denote \( R(t_j) \), \( j = 1, \ldots, n \) as \( R_1, \ldots, R_n \) in order of the event times. As the cases specified at time \( t_k \) are deleted from the risk set at \( t_k \) for \( t_k > t_j \), or some subjects can be censored between two consecutive time points \( (t_{j-1}, t_j) \), we have a monotone decreasing structure on the risk set such that

\[
R_n \subset R_{n-1} \subset \cdots \subset R_1.
\]

We also denote \( r_j \) as the size of the risk set \( R_j \) and it also follows a non-increasing sequence, \( r_1 \geq r_2 \geq \cdots \geq r_n \). In this article, we considered cost-effective ECC designs as the sub-sample selection method. From the ECC design, we first obtained a sub-sample including \( n \) cases and \( r \) controls in the final risk set. In this ECC sample, we additionally observe \( X \), an exposure variable of research interest, which was not collected in the cohort sample due to its higher measurement cost.

NCC design

Thomas\(^2\) initially suggested that the NCC design can be implemented by taking control samples from each risk set. The theoretical properties using this NCC design were provided by Goldstein and Langholz\(^4\) and Borgán et al.\(^5\) under Cox's\(^3\) proportional hazard model. A conventional NCC design is as follows

(i) Specify the risk set \( R_i \) for a case \( i \).
(ii) Randomly select \( m \) controls from the risk set \( R_i \) without replacement.
(iii) Repeat (A) and (B) for all cases.

Note that some individuals can be repeatedly selected for different cases because non-censored individuals cannot be specified as the cases are included in all risk sets. Usually, the number of control subjects at event time points is considered in between \( m = 1 \) and \( m = 4 \). Figure 1 illustrates the classical NCC design. Figure 1A specifies the risk sets at each event time point and Figure 1B presents the selected NCC samples that have two controls for each case. Here, the two control subjects are randomly selected from each risk set.

For the NCC studies, the individual matching of controls to cases can be applied by adjusting the confounding or background covariates at the control-sampling stages. The matching technique in the NCC study can be used when researchers believe that all controls have the same value of specific characteristics as the corresponding case such as age and gender covariates. The procedure of matching the NCC study design is that at each event time, control subjects are randomly sampled from within strata defined by the matching covariates. Let \( Z \) be a matching covariate. The risk set is restricted to \( R(Z) \), where \( R(Z) \subset R \) includes the individuals that have the same category of the covariate \( Z \) as the value \( Z \) for the case \( i \).

According to the data type of \( Z \), the matching procedure can be implemented in two ways: category and caliper matching designs.\(^14\) When the covariate \( Z \) is categorical, the controls and cases are exactly matched with values of the covariate \( Z \). For the continuous or ordered categorical \( Z \), the controls are selected so that the distances of \( Z \) between the cases and controls, denoted by \( d(\cdot, \cdot) \), lie in a given tolerance. For a fixed tolerance level \( \varepsilon \), steps 1 and 2 in the conventional NCC design are changed by incorporating the matching procedure:
for a case, we reconstructed the controls, where \( m = \lfloor s/n \rfloor \)

draw \( s \) controls from the risk set \( R_i \).

if \( s = n \), randomly assign \( s(=n) \) controls to \( n \) cases.

else if \( s < n \)

set \( d = t \)

while \( s \leq n/d \) do

assign \( s \) controls to \( s \) cases randomly selected from \( C \).

end while

randomly select \( n - d \times s \) controls from the \( s \) controls and then randomly assign them to the remaining cases.

else if \( s > n \)

set \( S \) as the set of controls and \( d = t \)

while \( n \leq s/d \) do

randomly select \( n \) controls from \( S \) and then randomly assign to \( n \) cases.

end while

set \( S \) as the size of final \( S \) and \( v_s = n - v_i \).

assign the remaining \( v_i \) controls on \( v_i \) cases randomly selected from \( C \).

randomly select \( v_j \) controls from the initial \( s \) controls and then assign then to the remaining \( v_j \) cases.

end if

\[ m = \lfloor x \rfloor \text{ is the smallest integer that satisfies } m \geq x. \]

(i) Specify the risk set \( R_i(Z) \) for a case \( i \), where \( R_i(Z) = \{ j : Y_j \geq t_i, Z_j = Z_i \} \) or \( R_i(Z) = \{ j : Y_j \geq t_i, Z_j = Z_i \} \).

(ii) Randomly select \( m \) controls from the risk set \( R_i(Z) \) without replacement.

From the sampling design, there is no variation in exposure status \( Z \) between the cases and controls. The matching design provides efficiency gains relative to simple random sampling when \( Z \) is dichotomous, uncommon, and the interaction is positive.\(^{15}\)

**CECC design using a resampling method**

In this article, we propose a new cost-effective ECC design using a resampling technique that allows us to select an arbitrary number of controls \( s \), whereas the total number of controls \( s \) is equal to \( m(=1,2,\ldots) \), a multiple of the total number of cases in the classical NCC design. The size of controls \( s \) is determined as an arbitrary integer number by the budget and time constraints and can be even smaller than the size of the cases, that is, \( s < n \). From the selected \( s \) controls, we reconstructed \( m \times n \) controls matched to the cases using a resampling technique. As we selected all the cases from the cohort, the design efficiency is proportional to the size of \( s \). The proposed CECC is presented in Algorithm 1. For convenience, we implicitly assume \( m \) as the smallest integer number that is larger than \( s/n \).

The main difference compared with the classical NCC design is that the new algorithm only uses the final risk set, that is, all of the cases share the same risk set, while the classical NCC design specifies each risk set for each case (see Figure 2).

Figure 2A mimics the NCC design similar to Figure 1A, but the final risk set is only considered as the risk set in the new CECC design. Figure 2B describes the control samples selected in the final risk set corresponding to the cases randomly assigned to each stratum on each case-event time point. This new sampling design is similar to the conditional approach of the ECC proposed by Salim et al,\(^\text{11}\) apart from the resampling parts.

This new CECC design has some advantages compared with the conventional NCC design: (i) The new algorithm is more cost-effective because the number of selected controls can be smaller than the number of cases and does not have to be a multiple of the number of cases, (ii) it does not need to specify all the risk sets at the sampling stage, and (iii) it does not require censoring information. The advantage (ii) can be applied to practical analysis. Suppose that we are interested in the gene-exposure interaction effects based on cohort samples. We want to apply the NCC design because a gene-association study is expensive or some subjects were lost, and we can get some data from censored subjects. Thus, the proposed method only considers the candidate control subjects as the final risk, and we can then obtain extra bio-markers or genetic information.

For the matching design, we need to separate a given risk set by the matching variable \( Z \). We can mimic the NCC sampling procedure with a resampling technique within the identified strata. The example algorithm for the matching design with the binary \( Z \) and \( s < n(m = 1) \) is presented in Algorithm 2.

**Partial likelihood and parameter estimation**

In this article, we implicitly adopt the same assumptions required for the classical NCC designs: (A1) event is rare; (A2) the censoring time is independent of the event time; and (A3) the event time is independent of each other. However, this assumption does not guarantee the consistency in parameter estimation. Thus, we use Salim et al’s\(^\text{10}\) partial likelihood to adjust our length-biased samples.

Applying Salim et al’s\(^\text{10}\) partial likelihood, the likelihood function according to our length-biased sample is

\[
L(\beta; x, z) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\exp(\beta_{ij} x_{ij} + \beta_{ij} z_{ij}) w_{ij}}{\exp(\beta_{ij} x_{ij} + \beta_{ij} z_{ij}) w_{ij}}
\]

where \((x_{ij}, z_{ij})\) are the covariates of case \( i \) and \((x_{j}, z_{j})\), \( j \neq i \), are the covariates of control \( j \) matched to case \( i \), and

\[
w_{ij} = \frac{S(t_i; x_{ij}, z_{ij})}{S(t_i; x_{ij}, z_{ij})}
\]

are the adjusting term for sampling bias defined by the fraction of survival times, \( S(t_i) \) and \( S(t_i) \), between event time \( t_i \) and the end-point of observation \( r_0 \).
Approximating the survival time as the Kaplan–Meier (KM) estimates, the weights $w_{ij}$ are estimated by

$$w_{Sij} = \frac{\exp(\beta_1 x_i - \pi) \cdot \exp(\beta_2 z_i - \pi)}{\hat{S}(t_x)}$$

where $\pi$ is the weighted mean of NCC samples and $\bar{\pi}$ is the average of cohort samples (see Salim et al. for more details).

Once we plug-in the estimated $w_{ij}$ to the likelihood function (1), we can obtain the estimator by maximizing the estimated partial likelihood or the logarithm of the estimated partial likelihood. For the matching design on $Z$, we have $z_i = \bar{z}_k$ for all $i \in C$ and so we need to remove a term associated with $\beta_2$ because they have the same value within in each case.

Results and Discussions

Simulation study

In this section, we first generated 2000 Monte Carlo (MC) cohorts, each of which consists of individuals of $N = 10,000$. To generate survival outcomes for cohort samples, the Cox proportional hazards model equation (3) is applied with constant baseline hazard $\exp(-4.5)$ and discrete covariates $Z$ which comes from the Bernoulli with a probability of 0.5. The censoring time is randomly generated from an exponential distribution with mean of 10. Here, $X$ is the main covariate of interest which is only available for the NCC samples, and $Z$ is used as matching variable for the matching design. The considered hazards ratios for $X$ and $Z$ are $\exp(-1)$ of 1.2, 1.5, 2.0 and $\beta_2 = 1$. The end-point $\tau_0$ is determined by the observation time for obtaining 250 or 500 events. The Cox proportional hazards model for the cohort is defined as

$$\lambda(t | X = x, Z = z) = \lambda(t) \exp(\beta_1 x + \beta_2 z)$$

where the baseline risk rate, $\lambda(t)$, is defined as $\exp(-4.5)$ in this simulation setting. This simulation setup is similar to that of Salim et al.

In this simulation study, we consider three estimators:

(i) NCC: estimators obtained from the classical NCC design samples;
(ii) CECC: estimators obtained from the CECC design samples with constant $w_{ij}$.
(iii) CECCW: estimators obtained from the CECC design samples with the estimated $w_{ij}$ in equation (2).

The CECC estimator does not use the sampling bias adjust term, but it does not require any censoring information for parameter estimation. Thus, we will also discuss if this naive CECC estimator without bias correction can be used in practice when the censoring time is not available or accurate. For estimation of the conditional logistic regression and the KM estimates, we used the R function `clogit` and `ncc` built in the R packages ‘survival’ and ‘Epi.’

As the simulation outputs, we report the relative bias (R.Bias), standard errors (SE), and mean squared errors (MSE) of the estimators. Here, R.Bias is defined by

---

**Algorithm 2: CECC Design Matching on $Z$.**

**Require:** the set of cases $C_k$ with $Z = k$, $k = 1, 2$, the final risk set $R_k$.

**Input:** control size $s < n$ and $m = 1$.

**Output:** matched case-control samples.

1. draw $s_k$ controls from the risk set $R_k(Z = k)$, $k = 1, 2$.
2. for $k = 1, 2$, do
3. if $k = 1, 2$, randomly assign $s_j(= n_j)$ controls to $n_j$ cases.
4. else if set $d = 1$.
5. while $s_j \leq n_j / d$ do
6. assign $s_j$ controls to $s_j$ cases randomly selected from $C_k$.
7. set $d = d + 1$ and $C_k = C_k \setminus \{selected s_j cases\}$.
8. end while
9. randomly select $n_j - d \cdot s_j$ controls from the $s_j$ controls and then assign them to the remaining cases.
10. end if
11. end for

---

**Figure 2.** Sampling procedure of CECC design. (A) Cost-effective extreme case-control sampling design. (B) Cost-effective extreme case-control sampling design with two control subjects. Solid dot: case; empty dot: subject at risk set $R_i$, $i = 1, 2, \ldots, n$; $x$: censored subject; $S_i$, $i = 1, 2, \ldots, n$: stratum.
controls for according to the number of cases and respectively. The selected controls are re-distributed using the proposed CECC algorithm to construct NCC samples. For simplicity, we set $m = 1$ or $2$ according to the number of initial controls, which is the number of matched controls for each case.

Table 1 shows that the CECCW estimators have better performance than the CECC estimators, because the bias adjusting terms are represented during the estimation. The CECC estimators produce a larger bias when the length bias is increased as the number of cases is increased from 250 to 500. However, the size of biases is not critical in the sense that we may use the naive CECC estimators in practice when the censored time is not available or accurate. Among three estimators, the CECCW estimators are the most efficient under the same number of controls. This because the CECCW estimators incorporate the cohort information through the estimated survival function. We can also confirm that the proposed method performs well even when the number of sampled controls is smaller than the number of cases. Although we need to pay some efficiency loss due to the size of controls, the biases are well controlled for this under sampling design. This is desirable results as the cost-effective NCC design. The efficiency of the proposed estimators depends on the size of total number of controls selected as the CECC sample.

Similar to Table 1, Table 2 also presents the MC relative biases, standard errors, and mean squared errors of the NCC, CECC, and CECCW estimators for matching scenario. The simulation results presented at Table 2 are almost similar to those of no matching case, which implies that when the matching variable is considered, the proposed method also works well. Thus, in summary, all estimators produce nearly unbiased estimates in this simulation setup. However, the CECCW is slightly more efficient compared with other estimators.

Real data example

Andersson et al.\textsuperscript{11,12} studied the association between chronic $\alpha$-particle irradiation from Thorotrast and the liver cancer incidence. All of the study subjects took cerebral angiography with or without Thorotrast from 1935 to 1947 or from 1946 to 1963, respectively. The ‘thoro’ data are available from R package ‘Epi.’ The modified data for the NCC and the CECC designs are employed. In particular, we consider both the CECC and the CECCW for this data analysis. For simplicity, we consider the small number of variables and a simple model. The variable is as follows:

- Sex: 0 for male and 1 for female.
- Event: indicator of liver cancer diagnosis.
- Exposure: injected volume of thorotrast in milliliter. Control patients have a 0 in this variable.
- Censored: 0—-not censored and 1—-censored.
- Incidence age: age of liver cancer diagnosis.
- Exit age: age of exit year from the study including the incidence age.
- Birth: birth cohort as 0 for birth date earlier than 1920 year and 1 for birth date later than 1920 year.
- Time = Exit age—Incidence age.

Tables 3 and 4 show the total number of observation is 2468. We considered the final exit date, February 20, 1992, for data application. The number of liver cancer cases is 130 and censored observations are 40 and the numbers of male and female are 1291 and 1177. The range of birth date is from January 7, 1868 to February 1, 1958. The number of subjects which was born before January 1, 1920 is 1816, otherwise 652. There are 1479 non-exposed and 898 exposed subjects in this study. The median of the follow-up years is approximate 26 years and the median of the incidence age (years) is 62.32 years old. For the age at injection (year), the median value is 40.35 years old.

Table 5 provides estimates of exposure $\exp(\beta_1)$, exponential function of estimates of exposure $\exp(\beta_1)$, and standard errors of estimates of exposure $\exp(\beta_1)$, 1 NCC data, 3 CECCs and CECCWs with ratios of cases and controls as 130:130 (1:1), 130:150 (1:2), and 130:200 (1:2) with or without matching. Here, the matching variable, “Birth,” is considered. Since the sex variable is not significant in the cohort analysis, we ignore the variable. Cohort with matching means the matching variable is used as a covariate in the cohort analysis. Table 5 shows that the CECCW results provide better standard errors than less bias than the NCC and the CECC. The reason is that we conjecture this follow-up time of this cohort is relative long and the cohort size is small.

Conclusions

In this article, we propose a new CECC utilizing a resampling idea to make flexibility the choice of the number of control samples. This new procedure provides two advantages compared with the classical NCC sampling design. It allows an under sampling design, which means the number of controls is less than the number of cases. Also, it does not need to specify
Table 1. Monte Carlo relative biases (R.Bias), standard errors (SE), and mean squared errors (MSE) of estimators from the NCC, CECC, and CECCW without matching.

| METHODS | CONTROLS | \( m \) | \( \beta_i = \log(1.2) \) | \( \beta_i = \log(1.5) \) | \( \beta_i = \log(2.0) \) |
|---------|----------|--------|----------------------------|----------------------------|----------------------------|
|         |          |        | R.BIAS (%) | SE | MSE | R.BIAS (%) | SE | MSE | R.BIAS (%) | SE | MSE |
| NCC     | 250:250  | 1:1    | -0.0453 | 0.1810 | 0.0328 | 1.2464 | 0.1867 | 0.0349 | 1.2540 | 0.1967 | 0.0388 |
| CECC    | 250:150  | 1:1    | -0.1486 | 0.2172 | 0.0472 | 0.1626 | 0.2202 | 0.0485 | 1.4069 | 0.2239 | 0.0502 |
|         | 250:250  | 1:1    | 1.7952 | 0.1818 | 0.0331 | 2.0897 | 0.1818 | 0.0331 | 2.3031 | 0.1909 | 0.0367 |
|         | 250:400  | 1:2    | 2.9312 | 0.1649 | 0.0272 | 2.9360 | 0.1653 | 0.0275 | 2.2260 | 0.1707 | 0.0294 |
| CECCW   | 250:150  | 1:1    | -1.4886 | 0.2141 | 0.0458 | -1.2853 | 0.2173 | 0.0473 | 0.1989 | 0.2206 | 0.0487 |
|         | 250:250  | 1:1    | 0.3004 | 0.1791 | 0.0321 | 0.5998 | 0.1790 | 0.0321 | 0.8325 | 0.1878 | 0.0353 |
|         | 250:400  | 1:2    | 1.4172 | 0.1624 | 0.0264 | 1.4408 | 0.1628 | 0.0265 | 0.7590 | 0.1680 | 0.0282 |
| NCC     | 500:500  | 1:1    | 1.9980 | 0.1236 | 0.0153 | 0.7447 | 0.1287 | 0.0166 | 1.0367 | 0.1334 | 0.0179 |
| CECC    | 500:300  | 1:1    | 4.3142 | 0.1504 | 0.0227 | 6.0110 | 0.1496 | 0.0230 | 4.9052 | 0.1601 | 0.0268 |
|         | 500:500  | 1:1    | 5.0875 | 0.1239 | 0.0154 | 4.4117 | 0.1286 | 0.0169 | 3.8251 | 0.1353 | 0.0190 |
|         | 500:800  | 1:2    | 3.4215 | 0.1163 | 0.0136 | 4.1937 | 0.1173 | 0.0141 | 3.9960 | 0.1202 | 0.0152 |
| CECCW   | 500:300  | 1:1    | 0.5894 | 0.1443 | 0.0208 | 2.5592 | 0.1446 | 0.0210 | 1.2784 | 0.1517 | 0.0231 |
|         | 500:500  | 1:1    | 1.3928 | 0.1194 | 0.0143 | 0.8520 | 0.1239 | 0.0154 | 0.4505 | 0.1304 | 0.0170 |
|         | 500:800  | 1:2    | -0.2131 | 0.1119 | 0.0125 | 0.6494 | 0.1130 | 0.0128 | 0.6140 | 0.1158 | 0.0134 |

# of cases = 250

# of cases = 500

Abbreviation: NCC, nested case-control.
| METHODS  | CONTROLS  | m     | $\beta_1 = \log(1.2)$ |          | $\beta_1 = \log(1.5)$ |          | $\beta_1 = \log(2.0)$ |          |
|----------|-----------|-------|------------------------|----------|------------------------|----------|------------------------|----------|
|          |           |       | R.BIAS (%) | SE | MSE | R.BIAS (%) | SE | MSE | R.BIAS (%) | SE | MSE |
| # of cases = 250 | | | | | | | | |
| NCC      | 250:250   | 1:1   | 1.6775 | 0.1856 | 0.0345 | 1.0622 | 0.1898 | 0.0360 | 1.0010 | 0.2056 | 0.0423 |
| CECC     | 250:150   | 1:1   | 2.8795 | 0.2167 | 0.0470 | 1.9887 | 0.2239 | 0.0502 | 1.4207 | 0.2340 | 0.0549 |
|          | 250:250   | 1:1   | -1.4748 | 0.1881 | 0.0354 | 2.3676 | 0.1906 | 0.0364 | 2.8904 | 0.1999 | 0.0404 |
|          | 250:400   | 1:2   | 1.6386 | 0.1796 | 0.0323 | 2.4780 | 0.1810 | 0.0328 | 2.4021 | 0.1937 | 0.0378 |
| CECCW    | 250:150   | 1:1   | 1.4270 | 0.2135 | 0.0456 | 0.5853 | 0.2208 | 0.0488 | 0.0391 | 0.2306 | 0.0532 |
|          | 250:250   | 1:1   | -2.8539 | 0.1855 | 0.0344 | 0.9647 | 0.1879 | 0.0353 | 1.5007 | 0.1971 | 0.0390 |
|          | 250:400   | 1:2   | 0.2178 | 0.1770 | 0.0313 | 1.0722 | 0.1785 | 0.0319 | 1.0205 | 0.1911 | 0.0366 |
| # of cases = 500 | | | | | | | | |
| NCC      | 500:500   | 1:1   | 0.2045 | 0.1309 | 0.0171 | 1.3677 | 0.1331 | 0.0178 | 0.3275 | 0.1386 | 0.0192 |
| CECC     | 500:300   | 1:1   | 5.6533 | 0.1546 | 0.0240 | 4.7117 | 0.1597 | 0.0259 | 4.0727 | 0.1616 | 0.0269 |
|          | 500:500   | 1:1   | 4.2535 | 0.1293 | 0.0168 | 4.4172 | 0.1320 | 0.0177 | 3.9260 | 0.1386 | 0.0199 |
|          | 500:800   | 1:2   | 5.0455 | 0.1263 | 0.0160 | 4.6792 | 0.1239 | 0.0157 | 4.1397 | 0.1317 | 0.0182 |
| CECCW    | 500:300   | 1:1   | 2.3357 | 0.1498 | 0.0224 | 1.5596 | 0.1549 | 0.0240 | 1.0381 | 0.1567 | 0.0246 |
|          | 500:500   | 1:1   | 1.0361 | 0.1253 | 0.0157 | 1.2788 | 0.1281 | 0.0164 | 0.9012 | 0.1344 | 0.0181 |
|          | 500:800   | 1:2   | 1.8110 | 0.1225 | 0.0150 | 1.5437 | 0.1203 | 0.0145 | 1.1168 | 0.1278 | 0.0164 |

Abbreviation: NCC, nested case-control.
all risk sets required for the classical NCC design. Considering these advantages, the new algorithm can be applied to the real data analysis when the budget and time constraints are limited to include new measurements. We can use CECCW estimators when the censored times are all available for cohort and use CECC estimators otherwise.

This study is not free from limitations. When the size of the final risk set is very small with respect to the cohort, that is, most units in the cohort have an event or are censored during the observation time; the proposed algorithm may lead to unintended biases during the estimation. This phenomenon is similar to a coverage problem in the sense that the small final risk set can easily fail to cover the whole range of risk sets used for the

| METHODS   | CONTROLS | $m$ | NO MATCHING | MATCHING |
|-----------|----------|-----|--------------|----------|
|           |          |     | $\beta$     | $\exp(\beta)$ | $SE(\beta)$ | $\beta$ | $\exp(\beta)$ | $SE(\beta)$ |
| Cohort    |          |     | 0.0758       | 1.0787    | 0.0046 | 0.0739 | 1.0767 | 0.0047 |
| NCC       | 130:130  | 1:1 | 0.0918       | 1.0962    | 0.0172 | 0.0989 | 1.1039 | 0.0171 |
| CECC      | 130:130  | 1:1 | 0.0958       | 1.1005    | 0.0167 | 0.1028 | 1.1083 | 0.0174 |
|           | 130:150  | 1:2 | 0.1002       | 1.1054    | 0.0137 | 0.1017 | 1.1070 | 0.0135 |
| CECCW     | 130:130  | 1:1 | 0.0659       | 1.0681    | 0.0086 | 0.0706 | 1.0731 | 0.0079 |
|           | 130:150  | 1:2 | 0.0732       | 1.0760    | 0.0067 | 0.0602 | 1.0620 | 0.0053 |
|           | 130:200  | 1:2 | 0.0642       | 1.0663    | 0.0062 | 0.0631 | 1.0651 | 0.0058 |

Abbreviation: NCC, nested case-control.
candidates of controls. Thus, the proposed sampling technique will efficiently work when we have rare events and observation time is relatively short avoiding too many censored units.

Acknowledgements
The authors would like to thank the editor and reviewers for their careful readings and thoughtful comments.

Author Contributions
All authors contributed equally to this work.

ORCID iD
Young Min Kim https://orcid.org/0000-0001-6295-1758

REFERENCES
1. Grant EJ, Cologne J, Sharp GB, et al. Bioavailable serum estradiol may alter radiation risk of postmenopausal breast cancer: a nested case-control study. *Int J Radiat Biol*. 2018;94:97–105.
2. Thomas DC. Addendum to: “methods of cohort analysis: appraisal by application to asbestos mining” by Liddell FDK, McDonald JC and Thomas DC. *J Roy Stat Soc Ser A*. 1977;140:469–491.
3. Oakes D. Survival times: aspects of partial likelihood (with discussion). *Int Stat Rev*. 1981;49:235–252.
4. Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in Cox regression models. *Ann Stat*. 1992;20:1903–1928.
5. Borgan O, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann Stat*. 1995;23:1749–1778.
6. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci*. 1996;11:35–53.
7. Langholz B, Goldstein L. Conditional logistic analysis of case-control studies with complex sampling. *Biostatistics*. 2001;2:63–84.
8. Yao Y, Yu W, Chen K. End-point sampling. *Stat Sin*. 2017;27:415–435.
9. Shoner A, Demichelis F, Calza S, et al. Molecular sampling of prostate cancer: a dilemma for predicting disease progression. *BMC Med Genom*. 2010;3:8.
10. Salim A, Ma X, Fall K, Andrén O, Reilly M. Analysis of incidence and prognosis from “extreme” case-control designs. *Stat Med*. 2014;33:5388–5398.
11. Andersson M, Carstensen B, Storm HH. Mortality and cancer incidence after cerebral angiography. *Radiat Res*. 1995;142:305–320.
12. Andersson M, Vyberg M, Visfeldt J, Carstensen B, Storm HH. Primary liver tumours among Danish patients exposed to Thorotrast. *Radiat Res*. 1994;137:262–273.
13. Cox DR. Regression models and life tables. *J Roy Stat Soc Ser B*. 1972;34:262–220.
14. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Epidemiol*. 1973;33:547–446.
15. Cologne J, Langholz B. Selecting controls for assessing interaction in nested case-control studies. *J Epidemiol*. 2003;13:193–202.
16. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457–481.