Received: 23 January 2019 | Revised: 8 July 2019 | Accepted: 11 July 2019

DOI: 10.1002/dmrr.3204

RESEARCH ARTICLE

WILEY

Nested case-control data analysis using weighted conditional logistic regression in The Environmental Determinants of Diabetes in the Young (TEDDY) study: A novel approach

Hye-Seung Lee | Kristian F. Lynch | Jeffrey P. Krischer | the TEDDY Study Group

Health Informatics Institute, Department of Pediatrics, University of South Florida Morsani College of Medicine, Tampa, FL, USA

Correspondence
Hye-Seung Lee, Health Informatics Institute, Department of Pediatrics, University of South Florida Morsani College of Medicine, Tampa, FL, USA.
Email: leeh@epi.usf.edu

Funding information
National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), Grant/Award Number: Contract No. HHSN267200700014C U01 DK63790, U01 DK63821, U01 DK63829, U01 DK63836, U01 DK63861, U01 DK63863, U01 DK63865, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, UC4 DK63821, UC4 DK63829, UC4 DK63836, UC4 DK63861, UC4 DK63863, UC4 DK63865, UC4 DK95300 and HHSN267200700014C; NIH/NCATS Clinical and Translational Science Awards to the University of Colorado, Grant/Award Number: UL1 TR001082; NIH/NCATS Clinical and Translational Science Awards to the University of Florida, Grant/Award Number: UL1 TR000064; JDRF; Centers for Disease Control and Prevention (CDC); National Institute of Environmental Health Sciences (NIEHS); National Institute of Child Health and Human Development (NICHD); National Institute of Allergy and Infectious Diseases (NIAID)

Abstract

Background: A nested case-control (NCC) design within a prospective cohort study can realize substantial benefits for biomarker studies. In this context, it is natural to consider the sample availability in the selection of controls to minimize data loss when implementing the design. However, this violates the randomness required for selection, and it leads to biased analyses. An inverse probability weighting may improve the analysis, but the current approach using weighted Cox regression fails to maintain the benefits of NCC design.

Methods: This paper introduces weighted conditional logistic regression. We illustrate our proposed analysis using data recently investigated in The Environmental Determinants of Diabetes in the Young (TEDDY). Considering the potential data loss, the TEDDY NCC design was moderately selective in its selection of controls. A data-driven simulation study was performed to present the bias correction when a nonrandom control selection was ignored in the analysis.

Results: The TEDDY data analysis showed that the standard analysis using conditional logistic regression estimated the parameter: −0.015 (−0.023, −0.007). The biased estimate using Cox regression was −0.011 (95% confidence interval: −0.019, −0.003). Weighted Cox regression estimated −0.013 (−0.026, 0.0004). The proposed weighted conditional logistic regression estimated −0.020 (−0.033, −0.007), showing a stronger negative effect size than the one using conditional logistic regression. The simulation study also showed that the standard estimate of β ignoring the nonrandom control selection tends to be greater than the true β (ie, positive relative biases).

Conclusion: Weighted conditional logistic regression can enhance the analysis by offering flexibility in the selection of controls, while maintaining the matching.

KEYWORDS

prospective cohort study, nested case-control design, selection bias, inverse probability weighting, weighted conditional logistic regression

1 | INTRODUCTION

7Prospective cohort studies are utilized to assess how incident events are influenced by the characteristics of interest in participants followed over time. However, the collection of prospective data can require substantial resources, especially when the incidence of events is low. When resources are limited, it may not be feasible to gather the data from the full cohort over the entire follow-up. A nested case-control (NCC) design is the primary choice in a prospective cohort study to avoid such situations without compromising many of the benefits from the full cohort analysis.[1,2] In modern epidemiological studies, as it becomes relatively easier to manage multicentre or international prospective cohort studies, the use of an NCC design has increased, especially when expensive biomarker analyses such as high-throughput genomics are pursued.[3-5]

An NCC design includes all event cases up to a specific follow-up time, but selects only a predetermined number of controls for each case from the event-free subjects at the time when a case developed the event.[6] Assuming that the selection of controls for each case was at random, conditional logistic regression is the standard statistical analysis. However, when the design is used for biomarker analyses, the selection of controls often depends on the availability of biospecimen samples since no data can be expected without the corresponding sample. This helps improve efficiency by reducing missing data, but it can introduce bias that may not be accounted for in the analysis using standard analytic tools.

In this paper, we propose an alternative selection bias corrected analysis in an NCC design. By adopting the approach by Lin and Paik[7] for a matched case-control data analysis, our approach maintains the matching and suggests how to obtain the control selection probability from the full cohort. This approach is illustrated in the application of the plasma 25-hydroxyvitamin D [25(OH)D] concentration analysis presented recently in an NCC study from The Environmental Determinants of Diabetes in the Young (TEDDY).[8] A TEDDY data-driven simulation study was conducted to assess the effect of bias correction. The performance of the simulation was described in relation to the effect size and the selection parameter of the factor of interest.

2 | BACKGROUND

2.1 | Nested case-control design

In a prospective cohort study, we observe time of event or censoring for each participant in follow-up, whichever comes first. When time of event is observed, a "risk-set" is constructed, which includes all participants in follow-up at the event time. Figure 1 illustrates a hypothetical prospective cohort study with five participants followed; of those, participants 1, 3, and 4 developed the event at times 1, 2, and 3. Hence, three risk-sets are constructed corresponding to each event: risk-set 1 by participant 1 including all five participants, risk-set 2 by participant 3 including participants 3, 4, and 5, and risk-set 3 by participant 4 including participants 4 and 5.

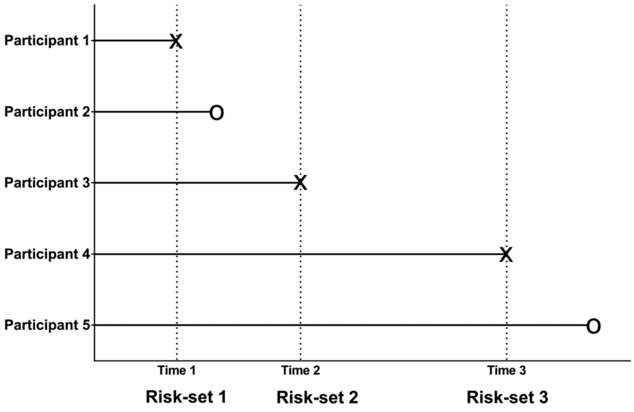

FIGURE 1 Hypothetical example to show a prospective cohort study with five participants followed

An NCC design includes all cases in follow-up but selects controls for a case from those event-free participants in the case's risk-set in a prospective cohort study. In Figure 1, participants 1, 3, and 4 are included as cases, and controls for each case are selected from the case's risk-set. For example, participants 2 to 5 are potential controls for participant 1 (case 1) in risk-set 1, but participant 5 is the only potential control for participant 4 (case 3) in risk-set 3. In this design, controls are expected to be randomly selected without replacement in a risk-set (ie, for better efficiency) but with replacement across risk-sets as long as the next risk-sets include them (ie, for the independence between risk-sets). Hence, a participant can appear more than once in different case-control sets since the participant can appear in different risk-sets by his/her observed time. However, case-control sets are independent of each other, assuming that risk-sets are independent of each other in the full cohort analysis. This implies that the information given for one case-control set is independent of the information given for another case-control set.

Since controls are selected from the same risk-set as the case, an NCC design is considered a matched case-control design with the risk-set as a matching factor. Therefore, this design can also match on potential confounders at a subject level. Also, with or without intention, this risk-set matching leads to matching on longitudinal data collected between a case and its matched controls (ie, a sample-level matching). As shown in Figure 2, through the risk-set matching, the longitudinal data in each matched case-control set are determined, depending on the case's event time. In the first set, only 3 observations per participant can be compared between the case and its matched controls, while 15 observations can be compared in the third set. If the matching is broken,

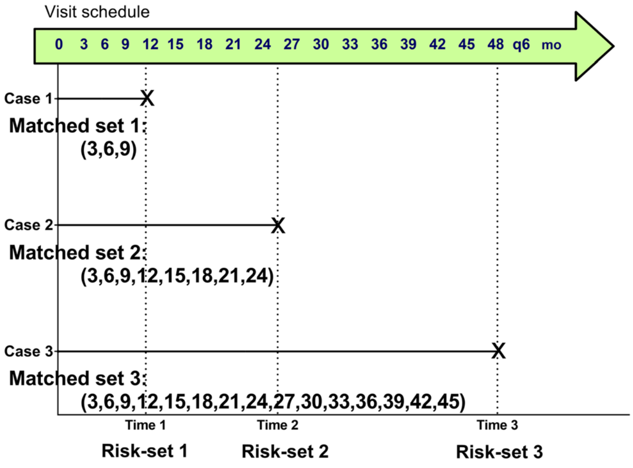

FIGURE 2 Hypothetical example to show possible variability in the number of longitudinal data between matched sets in a nested case-control design

this sample-level matching will introduce unforeseen bias in the analysis, in addition to that from the subject-level matching.

2.2 | Nested case-control data analysis

2.2.1 | Conditional logistic regression

In a matched case-control data analysis, conditional logistic regression is primarily used to examine the association between the event and characteristics measured by the time of event. The conditional odds can be written as below:

$$\frac{P(Y=1\,|\,X,Z,S=1)}{P(Y=0\,|\,X,Z,S=1)}=\frac{P(S=1\,|\,Y=1,X,Z)}{P(S=1\,|\,Y=0,X,Z)}\frac{P(Y=1\,|\,X,Z)}{P(Y=0\,|\,X,Z)} \quad (1)$$

where *Y* is the indicator of being the case, *X* is the vector of characteristics of interest, *Z* is the vector of matching factors, and *S* is the indicator of being included in the matched case-control design. The assumption that the selection is at random implies that the selection does not depend on *X*, which is the characteristics of interest {ie, $P(S = 1|\,Y,X, Z) = P(S = 1|\,Y,Z)$}. Therefore, Equation (1) can be written as

$$logitP(Y=1|S=1,X,Z)=logitP(Y=1|X,Z)+f(Z) \quad (2)$$

where $f(Z)=\log\left\{\frac{P(S=1\,|\,Y=1,X,Z)}{P(S=1\,|\,Y=0,X,Z)}\right\}=\log\left\{\frac{P(Y=1\,|\,S=1,Z)P(Y=0|Z)}{P(Y=0\,|\,S=1,Z)P(Y=1|Z)}\right\}=\log\left\{\frac{P(Y=0|Z)}{P(Y=1|Z)}\right\}-\log(m)$, for 1 to *m* (the number of controls) matched case-control design. Then, the function *f* (*Z*) is canceled out, and the conditional likelihood for standard conditional logistic regression becomes

$$L(\beta)=\prod_{i=1}^{n}\frac{\exp\{X_{i0}(t)\beta\}}{\sum_{j\in Rs_i}\exp\{X_{ij}(t)\beta\}} \quad (3)$$

where *n* is the number of cases, and the set Rs_i includes the case 0 and *m* controls matched to the case in the *i*th matched case-control set, *j* = 0,1,2,. . *m*. For an NCC design, $X_{ij}(t)$ can be defined as the *j*th subject's characteristics of interest by the event time *t* of the case in the *i*th set, since the design is matched by the case's risk-set. Although the likelihood (3) is the same as the partial likelihood for full cohort analysis, the risk in Rs_i is fixed by the design, as opposed to the one constructed by chance in the full cohort analysis. The regression parameter *β* corresponds to the log of the odds ratio for a unit change of $X_{ij}(t)$, as the likelihood is formed by modeling the odds.

2.2.2 | Weighted Cox regression

If the matching is broken in an NCC design, the participants included in the design may be considered as a subcohort selected from the full cohort. Then, weighted Cox regression can be a choice for selective cohort analysis with the weight being the inverse selection probability for each participant.[9-11] The weighted partial likelihood can be written as

$$L(\beta)=\prod_{i=1}^{n}\frac{\exp\{X_{i0}(t)\beta+Z_{i0}\Upsilon\}}{\sum_{j\in M_i}W_j\exp\{X_{ij}(t)\beta+Z_{ij}\Upsilon\}} \quad (4)$$

where W_i is the inverse of the selection probability (p_i) for the *i*th subject in the subcohort (ie, $W_i=\frac{1}{p_i}$) and M_i is the risk-set including the subjects in an NCC design who were being followed at the case *i*"s event time. Note that this approach assumes that the risk in M_i is constructed at random among the subjects included in an NCC design. Here, the regression parameter *β* may correspond to the log of the hazard ratio for a unit change of $X_{ij}(t)$, after adjusting for the matching factors.

This approach has been used to analyze secondary events observed other than the primary for cases in the design.[12,13] This approach can be viewed as a selection bias-corrected analysis but breaks the matched design. When the study design implements the matching at a sample level, breaking the matching introduces the variability that cannot be properly controlled in the analysis. Also, it may reduce the efficiency. For example, in Figures 1 and 2, if participant 5 was a control for participant 1, the pair could have processed three samples at 3, 6, and 9 months in the same batch by the sample-level matching. In this weighted Cox regression analysis, participant 5 can be also in risk-sets 2 and 3, but this participant's information is incomplete for those risk-set analyses since those three samples would have been only analyzed by the NCC design.

3 | WEIGHTED CONDITIONAL LOGISTIC REGRESSION FOR NESTED CASE-CONTROL ANALYSIS

In Equation (1), the assumption that the selection in the design is random leads to the standard conditional likelihood for inference in Equation (3). In an NCC design, all cases are included, so the assumption remains true for cases {ie, $P(S = 1|\,Y = 1,X,Z) = 1$}. However, the assumption for the selection of controls {ie, $P(S = 1|\,Y = 0,X,Z) = P(S = 1|\,Y = 0,Z)$} may not be true. When $P(S = 1|\,Y = 0,X,Z) \neq P(S = 1|\,Y = 0,Z)$, instead of Equation (2), the log odds for an NCC design becomes as follows:

$$logitP(Y=1|S=1,X,Z)=logitP(Y=1|X,Z)-\log\{P(S=1|Y=0,X,Z)\} \quad (5)$$

Then, by denoting W_{ij} as the inverse of the selection probability {ie, $\frac{1}{P(S=1\,|\,Y=0,X,Z)}$} for the *j*th subject in the *i*th set, the standard conditional likelihood (3) becomes

$$L(\beta)=\prod_{i=1}^{n}\frac{\exp\{X_{i0}(t)\beta\}}{\sum_{j\in Rs_i}W_{ij}\exp\{X_{ij}(t)\beta\}} \quad (6)$$

which is the conditional likelihood for weighted conditional logistic regression. Note that the set Rs_i stays the same as (3) by keeping the matching in the design.

Since the full cohort from which the NCC design participants are selected is available, the full cohort data can be used to estimate the selection probability $P(S = 1 \mid Y = 0,X,Z)$ for those selected controls. We fit a logistic regression model on the factor of interest *X* and the matching factor Z for the estimation of the selection probability. If we have complete data on X and Z, the probability estimator is expected to be unbiased. However, X is most likely unavailable in the full cohort since an NCC design is utilized to avoid having to collect that in the full cohort. Also, as a part of Z, the risk-set matching for an NCC design needs to be translated to an individual level in the full cohort. Instead of X and the risk-set matching, we use proxy variables that can explain the selection in an NCC design. Our motivation to consider $P(S = 1| Y = 0,X,Z) \neq P(S = 1| Y = 0,Z)$ is when controls' sample availability is incorporated in the selection of controls for biomarker studies. Moreover, the size of risk-set, which can directly affect the probability of control selection, is mostly determined by the duration of follow-up. Thus, we propose to use those factors related to the study compliance or duration of follow-up as the proxy variables.

This inverse selection probability weighting approach is also useful when the characteristics for event-free subjects using the data from an NCC design are of interest. When matching factors other than risk-set were used in an NCC design, the characteristics in selected controls become similar to their cases, rather than those in event-free participants in the cohort. Hence, the controls included in an NCC design cannot be directly used to make inference on event-free population about the characteristics collected in an NCC design. In this context, this selection bias-corrected approach can also help make the inference.

4 | APPLICATION: TEDDY NESTED CASE-CONTROL DESIGN

TEDDY is a prospective cohort study across six participating clinical centres: the Pacific Northwest Diabetes Research Institute, Seattle, Washington; the Barbara Davis Center, Denver, Colorado; a combined Georgia/Florida site at the Medical College of Georgia, Augusta, Georgia and the University of Florida, Gainesville, Florida; University of Turku (Turku, Oulu, and Tampere, Finland); Lund University, Malmö, Sweden; and the Diabetes Research Institute, Munich, Germany.[14,15] TEDDY enrolled 8676 children before 4.5 months of age through newborn screening for high-risk HLA-DR-DQ genotypes and will follow them up until 15 years of age to identify genetic and environmental triggers of type 1 diabetes (T1D). The protocol was approved by Institutional Review Boards at participating centres, and all participants provided written informed consent before participation in the study.

In order to perform analyses across various biomarkers, TEDDY set up two NCC designs: one for islet autoimmunity (IA, the prediabetic endpoint) and the other for T1D. At close of the cohort for the NCC design (ie, sampling time), the median follow-up age was 40 months (first quartile = 25 and the third quartile = 60). Additional matching factors were having a first-degree relative with T1D (T1D family history), sex, and clinical centre located in the region where the participant was enrolled. TEDDY selected controls based on their sample availability in the six potential controls randomly selected from each risk-set.[16] This was not completely a selective selection, but the bias could still affect the analysis. For example, if three controls were randomly selected, through 100 bootstrap samples, the odds ratio for a factor can be expected to be 1.89 with 95% confidence interval (1.87, 1.91). But if the factor was analyzed in the 1 to 3 TEDDY NCC design,[8] the odds ratio estimate is 1.96.

TEDDY recently investigated whether plasma 25(OH)D concentration (nmol/L) throughout childhood is associated with development of IA in the 1 to 3 TEDDY NCC design.[8] The childhood 25(OH)D concentration was defined as the average of 25(OH)D measured up to each case's event time. The authors analyzed 376 matched sets including 1041 controls with at least one measure of 25(OH)D prior to each case's event time, using standard conditional logistic regression. There was a total of 1375 participants: 376 participants developed IA and 999 participants who were IA-free at sampling time. We used these data to illustrate our proposed selection bias-corrected analysis.

4.1 | Selection probability estimation

Since cases are also potential controls until they develop the event of interest, the population for event-free subjects (ie, $Y = 0$) includes cases by their event time, as well as event-free subjects by their censored time at the time of design. A logistic regression model was used to estimate the selection probability for being included as a control in the NCC design for IA. We considered the factors related to retention in TEDDY as proxy variables. Previously, TEDDY identified such factors as country where the participant was enrolled, sex, illness experienced during the first year, maternal age, father's study participation, maternal lifestyle behaviours, and accuracy of the mother's risk perception.[17,18] Therefore, the logistic regression model considered the matching factors (T1D family history, child's sex, and clinical centre), the observed age (age of IA for IA cases and age of censoring at sampling for IA-free children), and those preidentified factors related to drop-outs in TEDDY. The final model included the factors with $P < 0.1$ (shown in Table 1). As expected, the matching factors were significantly associated with the control selection, along with the observed age showing older children being more often included. Participants with characteristics associated with higher compliance were more likely to be included as controls (positive father's study participation, older maternal age, and more reported illnesses within the first year). The selection probability was estimated from the final logistic regression model fit.

4.2 | Computation

For the selection bias-corrected analyses, the inverse of the selection probability estimate was applied as a weight for the regression parameter estimation. Taking into account the variability of the selection probability estimation, the jackknife variance was calculated and an approximation of the 95% confidence interval was obtained. Without weighting, the standard likelihood analysis was applied to obtain the regression parameter estimate and 95% confidence interval.

TABLE 1 Estimates of selection model by logistic regression from the TEDDY full cohort

		Odds Ratio	95% Confidence Interval	
			Lower	Upper
Observed age, mo		1.027	1.024	1.030
Clinical centre	Colorado	0.780	0.637	0.954
	Georgia	0.597	0.462	0.771
	Washington	0.576	0.457	0.725
	Finland	1.139	0.966	1.344
	Germany	0.928	0.714	1.207
	Sweden	1		
Sex	Girls	0.758	0.668	0.861
	Boys	1		
T1D family history	Yes	3.320	2.793	3.946
	No	1		
Father's participation	Yes	1.855	1.166	2.952
	No	1		
Maternal age, y		1.018	1.005	1.031
Number of illness in the first year		1.016	0.999	1.033

As an illustrative purpose, Cox regression was applied after adjusting for those additional matching factors, in order to examine the association between childhood 25(OH)D concentration and IA. The average of 25(OH)D was analyzed as a time-varying covariate by calculating it in each risk-set. Without a weighting, ignoring the NCC design, this produces a biased analysis since those subjects in the NCC design are handled as if they were the full TEDDY cohort. As shown in Table 2, the biased regression parameter estimate was −0.011 (95% confidence interval: −0.019, −0.003). The standard analysis using conditional logistic regression estimated the parameter −0.015 (−0.023, −0.007). Although this is supposed to be the best, it may be also biased due to the moderate selective control selection based on the sample availability in TEDDY. In applying weighted Cox regression adjusted for the matching factors, the parameter estimate became −0.013 (−0.026, 0.0004), with a slightly larger variation. When we applied the proposed weighted conditional likelihood, the estimate was −0.020 (−0.033, −0.007), showing a stronger negative effect size than the one using conditional logistic regression.

We also summarized the childhood 25(OH)D concentration by the case-control status (Table 3). By the nature of the design, the data for controls are available only up to the time of event of the cases to whom they were matched. If a case was also included as a control for another case, breaking the matching implies that the data as a control from the case are excluded from the analysis. On the other hand, our approach that keeps the matching includes the data as a control from the case, by the assumption that the matched sets are independent of each other by the design. The mean childhood 25(OH)D concentration was 51.33 nmol/L (standard deviation of 16.82) in the cases and 54.63 nmol/L (16.77) in the controls, respectively. When the proposed weighting was applied, the weighted mean in the controls was 55.04 nmol/L (17.21).

5 | SIMULATIONS

Based on the TEDDY data, a simulation study was conducted to assess the bias when a nonrandom control selection was ignored in an NCC design. The controls selected were determined by the 1 to 3 TEDDY NCC design. The prevalence model for IA given a covariate X was determined from the logistic regression model fit as $logitP(Y = 1| Z^a) = - 3.1533+g(Z^a)$ in the TEDDY cohort. When Z^a denotes the matching factors other than the risk-set, $g(Z^a) = - 0.0365 * Colorado - 0.3430 * Georgia - 0.4431 * Washington+0.4103 * Finland +0.0610 * Germany+1.0339 * FDR - 0.2423 * Girl$ in the TEDDY design. All variables are indicators; for example, $FDR = 1$ if the child has a T1D family history as defined in TEDDY. We assumed the prevalence model for IA given a covariate X as

$$logitP(Y=1|X,Z^a)=\beta 0+g(Z^a)+\beta^*X \tag{7}$$

Based on the invariance property of the odds ratio, we assumed the covariate model for X as $logitP(X = 1| Y,Z^a) = g(Z^a)+\beta * Y$, resulting in

$$P(X=1|Y,Z^a)=1/\{1+\exp(-g(Z^a)-\beta^*Y)\} \tag{8}$$

Assuming the control selection from event-free subjects in the cohort also depended on X and Z^a, the selection model can be written

TABLE 2 Association between childhood 25(OH)D concentration (average by event time, nmol/L) and islet autoimmunity (IA) in the TEDDY 25(OH)D analysis

Approach	Selection Bias Correction	Likelihood	Regression Parameter Estimate	95% Confidence Interval	
				Lower	Upper
Keeping the matching[a]	Without	Conditional[c]	−0.015	−0.023	−0.007
	With	Weighted conditional[d]	−0.020	−0.033	−0.007
Breaking the matching[b]	Without	Partial[c]	−0.011	−0.019	−0.003
	With	Weighted partial[d]	−0.013	−0.026	0.0004

[a] Conditional logistic regression was used. Childhood 25(OH)D concentration was calculated with the measures by the case's age of IA for each matched set.

[b] Cox regression adjusted for clinical centre, sex, and T1D family history was used. Childhood 25(OH)D concentration was calculated at each risk-set to be analyzed as a time-dependent covariate.

[c] Likelihood variance estimation.

[d] Jackknife variance estimation.

TABLE 3 The mean 25(OH)D concentration (nmol/L) at the status of IA free in the TEDDY 25(OH)D analysis

			N	Mean(Standard Deviation)
		Characteristics of	**376**	**51.33 (16.82)**
Cases	Selection bias correction			
Controls (keeping the matching)	Without	Controls	1041	54.63 (16.77)
	With	Event-free subjects in the cohort	1041	55.04 (17.21)
Event-free subjects (breaking the matching)	Without	Selective event-free subjects at the time of the design	999	54.83 (16.74)
	With	Event-free subjects at the time of the design, by the cases' event time	999	55.11 (17.24)

as $logitP(S = 1|\, Y = 0, Z^a, X) = r(Z^a) + \alpha * X$, where $r(Z^a)$ is a linear function of Z^a and α is the selection parameter for the dependency between S and X. Then, we can assume $logitP(X = 1|\, Y = 0, Z^a, S) = s(Z^a) + \alpha * S$, resulting in

$$logitP(X = 1|Y = 0, Z^a, S = 1) - logitP(X = 1|Y = 0, Z^a, S = 0) = \alpha \tag{9}$$

Using (8), we first generated X for the cases (Y = 1), given β (effect size). For event-free subjects (Y = 0), using (8) and (9), X was generated for those selected as controls (S = 1) and those not selected in the cohort (S = 0), respectively, given β and α.

Based on the randomly generated X and the given factor Z, we estimated $P(S = 1|\, Y = 0, Z, X)$ by fitting a logistic regression model and obtained the estimate of β using the standard conditional logistic regression ignoring the nonrandom selection, as well as the proposed conditional logistic regression weighted by the inverse selection probability. Then, the relative bias was obtained as the difference from the estimate of β by fitting the likelihood (6) in the simulated cohort. Two selection probabilities were considered for Z: (1) the matching factors other than risk-set (ie, Z^a) and (2) in addition to (1), the proxy variables for the risk-set matching, which are the observed age, father's study participation, maternal age, and illness within the first year. This process was repeated 100 times, and the mean and standard deviation of the relative bias are reported in Table 4.

Without the correction, the estimate of β tends to be greater than the true β (ie, positive relative biases). A stronger selection parameter showed greater bias when the nonrandom selection was ignored. With the correction, bias was reduced but still remained. We suspect that this is because the simulated biases were generated without reflecting the risk-set matching when the controls selected were based on that. The bias reduction varied depending on the combination of effect size and selective parameter, but it was generally improved when the proxy variables for the risk-set matching were considered in the selection probability estimation.

6 | DISCUSSION

NCC studies are particularly advantageous for longitudinal biomarker studies as they can reduce the high cost and labour associated with collecting complete data in prospective cohort studies. The choice of this design for biomarker studies is growing, not only because it requires a small selection of noncases but also because the design can be used with greater flexibility to match on longitudinal variables such as the sample availability/compliance. As the NCC studies become more popular and more flexibly designed, the importance of how well the choice of statistical tool fully respects the way the study is constructed will be vital to produce valid findings from the study.

A key aspect of an NCC design is the selection of a control to pair with a case at a specific time based on the case's event. The control is selected among the event-free subjects at the specific time unique to each case (ie, the risk-set matching). The chance of the selection must be independent of when the subjects drop-out of the study or later become a case themselves in the full cohort (ie, between risk-set independence). In practice, often a desire is to avoid selecting any controls that become eventually cases in the closed cohort at the time of the design. However, this modifies the risk-sets and violates the between risk-set independence. Then, the design becomes neither an NCC design nor a case-control design, and no standard statistical methods for either design will produce valid analyses. If the implementation of an NCC design maintained the between risk-set independence, the choice of analytical tool should be one of those methods conditioning on the matching. When the matching is ignored (ie, broken), no statistical modelling will be sufficient to remove the bias given the complexity of longitudinal matching nested within the subject level of matching. For this reason, breaking the matching should be the last choice in the NCC data analysis.

In this paper, we considered when controls were selectively chosen within a risk-set, in order to avoid selecting controls without necessary data for the implementation of an NCC design. We proposed an inverse probability weighting within the matching strata and analyzed the NCC data with weighted conditional logistic regression. Although weighted Cox regression has been available for nonrandom NCC design, this technique requires the matching to be broken and considers those included in the design as a subcohort. This application fails to support the choice of an NCC design to begin with. In order to estimate the selection probability of controls, we used a logistic regression model with the factors related to drop-out and compliance.

We illustrated our approach using the TEDDY data analysis. However, the TEDDY NCC design was not completely selective

TABLE 4 Simulation results from 100 replications: relative bias (empirical standard deviation)

		Conditional Logistic Regression		
			With Selection Bias Correction	
True Effect Size β	Selection Parameter α	Without Selection Bias Correction	Selection Probability Estimation on the Matching Factors Other than Risk-Set	Selection Probability Estimation on the Matching Factors Other than Risk-set + TEDDY Compliance Factors Including the Observed Age
−2.0	−1.25	0.972 (0.065)	−0.200 (0.054)	−0.174 (0.065)
	−0.75	0.592 (0.069)	−0.360 (0.071)	−0.351 (0.079)
−1.5	−1.25	0.984 (0.061)	−0.075 (0.048)	−0.038 (0.063)
	−0.75	0.596 (0.059)	−0.224 (0.058)	−0.213 (0.067)
−1.0	−1.25	0.995 (0.061)	0.083 (0.055)	0.135 (0.071)
	−0.75	0.602 (0.056)	−0.080 (0.050)	−0.059 (0.062)
−0.02	−1.25	1.012 (0.070)	0.663 (0.150)	0.726 (0.148)
	−0.75	0.608 (0.061)	0.295 (0.074)	0.342 (0.084)

since six potential controls were randomly selected first, from which three were selected based on availability of samples. Therefore, the difference we presented between with and without weighting in the conditional logistic regression analysis may not be greater than that if the design was completely selective. In our simulation study, we kept the status of TEDDY case-control and considered two types of selection probability estimation with and without proxy variables for the risk-set matching. We showed the bias in the analysis without weighting and the bias reduction in weighted conditional logistic regression with both types of weighting. The weighting that considered those factors for the risk-set matching performed better in general but still failed to remove the bias completely. It is likely because the simulated biases did not reflect the risk-set matching when the TEDDY control status was used. Nevertheless, performance may be improved with better estimates of the selection in a future study.

ACKNOWLEDGEMENTS

A complete list of the members of the TEDDY Study Group can be found in the online supplemental appendix.

The TEDDY Study is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and Contract No. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082).

CONFLICT OF INTEREST

No other potential conflicts of interest relevant to this article were reported.

AUTHOR CONTRIBUTIONS

HL and JPK conceptualized the study. HL performed statistical analysis and wrote the manuscript. JPK and KFL acquired data, reviewed, and contributed to discussion. All authors approved the final version of the manuscript.

ORCID

Hye-Seung Lee https://orcid.org/0000-0002-5194-7101
Kristian F. Lynch https://orcid.org/0000-0002-4310-4163
Jeffrey P. Krischer https://orcid.org/0000-0003-4526-888X

REFERENCES

1. Thomas D. Addendum to 'Methods of cohort analysis: Appraisal by application to asbestos mining'. In: Liddell, FDK.; McDonald, JC.; Thomas, DC., editors. *J R Stat Soc*. 1977;140:469-491.
2. Wacholder J. Practical considerations in choosing between the case-cohort and NCC designs. *Epidemiology*. 1991;2(2):155-158.
3. Baker SG. Improving the biomarker pipeline to develop and evaluate cancer screening tests. *J Natl Cancer Inst*. 2009;101(16):1116-1119. Epub 2009/07/04
4. Rundle A, Ahsan H, Vineis P. Better cancer biomarker discovery through better study design. *Eur J Clin Invest*. 2012;42(12):1350-1359.
5. Rundle AG, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. *Cancer Epidemiol Biomarkers Prev*. 2005;14(8):1899-1907. Epub 2005/08/17
6. Rothman K. *Modern Epidemiology*. Boston: Little, Brown and Company; 1986.
7. Lin I-F, Paik MC. Matched case-control data analysis with selection bias. *Biometrics*. 2001;57(4):1106-1112.
8. Norris J, Lee H-S, Frederiksen B, et al. Plasma 25-hydroxyvitamin D concentration and risk of islet autoimmunity. *Diabetes*. 2017 Oct 23;67(1):146-154 pii: db170802. https://doi.org/10.2337/db17-0802 [Epub ahead of print]

9. Samuelsen SO. A pseudo-likelihood approach to analysis of nested case-control studies. *Biometrika*. 1997;84(2):379-394.
10. Stoer NC, Samuelsen SO. Inverse probability weighting in nested case-control studies with additional matching—a simulation study. *Stat Med*. 2013;32(30):5328-5339.
11. Chen K. Generalized case-cohort estimation. *J R Stat Soc Ser B*. 2001;63(4):791-809.
12. Borgan O, Keogh R. Nested case-control studies: should one break the matching? *Lifetime Data Anal*. 2015;21(4):517-541.
13. Kim RS, Kaplan RC. Analysis of secondary outcomes in nested case-control designs. *Stat Med*. 2014;33(24):4215-4226.
14. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr Diabetes*. 2007;8(5):286-298.
15. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Ann N Y Acad Sci*. 2008;1150:1-13.
16. Lee H-S, Burkhardt BR, McLeod W, et al. and the TEDDY study group. Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab Res Rev*. 2014;30(5):424-434.
17. Johnson SB, Lee H-S, Baxter J, Lernmark B, Roth R. Simell T for the TEDDY study group. The Environmental Determinants of Diabetes in the Young (TEDDY) study: predictors of early study withdrawal among participants with no family history of type 1 diabetes. *Pediatr Diabetes*. 2011;12(3):165-171.
18. Johnson SB, Lynch KF, Baxter J, et al. Predicting later study withdrawal in participants active in a longitudinal birth cohort study for 1 year: The TEDDY study. *J Pediatr Psychol*. 2016;41(3):373-383.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Lee H-S, Lynch KF, Krischer JP, the TEDDY Study Group. Nested case-control data analysis using weighted conditional logistic regression in The Environmental Determinants of Diabetes in the Young (TEDDY) study: A novel approach. *Diabetes Metab Res Rev*. 2019;e3204. https://doi.org/10.1002/dmrr.3204

The TEDDY Study Group

Colorado Clinical Center: Marian Rewers, M.D., Ph.D., PI[1,4,5,6,10,11], Kimberly Bautista[12], Judith Baxter[9,12,15], Daniel Felipe-Morales, Kimberly Driscoll, Ph.D.[9], Brigitte I. Frohnert, M.D.[2,14], Marisa Gallant, M.D.[13], Patricia Gesualdo[2,6,12,14,15], Michelle Hoffman[12,13,14], Rachel Karban[12], Edwin Liu, M.D.[13], Jill Norris, Ph.D.[2,3,12], Andrea Steck, M.D.[3,14], Kathleen Waugh[6,7,12,15]. University of Colorado, Anschutz Medical Campus, Barbara Davis Center for Childhood Diabetes.

Finland Clinical Center: Jorma Toppari, M.D., Ph.D., PI[¥^1,4,11,14], Olli G. Simell, M.D., Ph.D., Annika Adamsson, Ph.D.[^12], Suvi Ahonen*[±§], Mari Åkerlund*[±§], Anne Hekkala, M.D.[µ¤], Henna Holappa[µ¤], Heikki Hyöty, M.D., Ph.D.*[±6], Anni Ikonen[µ¤], Jorma Ilonen, M.D., Ph.D.[¥¶3], Sinikka Jäminki*[±], Sanna Jokipuu[^12], Leena Karlsson[^], Miia Kähönen[µ¤12,14], Mikael Knip, M.D., Ph.D.*[±5], Minna-Liisa Koivikko[µ¤], Mirva Koreasalo*[±§2], Kalle Kurppa, M.D., Ph.D.*[±13], Jarita Kytölä*[±], Tiina Latva-aho[µ¤], Katri Lindfors, Ph.D.*[13], Maria Lönnrot, M.D., Ph.D.*[±6], Elina Mäntymäki[^], Markus Mattila*, Katja Multasuo[µ¤], Teija Mykkänen[µ¤], Tiina Niininen[±]*[12], Sari Niinistö[±§2], Mia Nyblom*[±], Sami Oikarinen, Ph.D.*[±], Paula Ollikainen[µ¤] , Sirpa Pohjola [µ¤], Petra Rajala[^], Jenna Rautanen[±§], Anne Riikonen*[±§], Minna Romo[^], Suvi Ruohonen[^], Satu Simell, M.D., Ph.D.[¥13], Maija Sjöberg[¥^12], Aino Stenius[µ¤12], Päivi Tossavainen, M.D.[µ¤], Mari Vähä-Mäkilä[^], Sini Vainionpää[^12], Eeva Varjonen[¥^12], Riitta Veijola, M.D., Ph.D.[µ¤14], Irene Viinikangas[µ¤], Suvi M. Virtanen, M.D., Ph.D.*[±§2]. [¥]University of Turku, *University of Tampere, [µ]University of Oulu, [^]Turku University Hospital, Hospital District of Southwest Finland, [±]Tampere University Hospital, [¤]Oulu University Hospital, §National Institute for Health and Welfare, Finland, [¶]University of Kuopio.

Georgia/Florida Clinical Center: Jin-Xiong She, Ph.D., PI[1,3,4,11], Desmond Schatz, M.D.*[4,5,7,8], Diane Hopkins[12], Leigh Steed[12,13,14,15], Jennifer Bryant[12], Katherine Silvis[2], Michael Haller, M.D.*[14], Melissa Gardiner[12], Richard McIndoe, Ph.D., Ashok Sharma, Stephen W. Anderson, M.D.[^], Laura Jacobsen, M.D.*[14], John Marks, DHSc.*[14], P.D. Towe*. Center for Biotechnology and Genomic Medicine, Augusta University. *University of Florida, [^]Pediatric Endocrine Associates, Atlanta.

Germany Clinical Center: Anette G. Ziegler, M.D., PI[1,3,4,11], Ezio Bonifacio Ph.D.*[5], Miryam D'Angelo, Anita Gavrisan, Cigdem Gezginci, Anja Heublein, Verena Hoffmann, Ph.D.[2], Sandra Hummel, Ph.D.[2], Andrea Keimer[¥2], Annette Knopff[7], Charlotte Koch, Sibylle Koletzko, M.D.[¶13], Claudia Ramminger[12], Roswith Roth, Ph.D.[9], Marlon Scholz, Joanna Stock[9,12,14], Katharina Warncke, M.D.[14], Lorena Wendel, Christiane Winkler, Ph.D.[2,12,15]. Forschergruppe Diabetes e.V. and Institute of Diabetes Research, Helmholtz Zentrum München, Forschergruppe Diabetes, and Klinikum rechts der Isar, Technische Universität München. *Center for Regenerative Therapies, TU Dresden, [¶]Dr. von Hauner Children's Hospital, Department of Gastroenterology, Ludwig Maximillians University Munich, [¥]University of Bonn, Department of Nutritional Epidemiology.

Sweden Clinical Center: Åke Lernmark, Ph.D., PI[1,3,4,5,6,8,10,11,15], Daniel Agardh, M.D., Ph.D.[6,13], Carin Andrén Aronsson, Ph.D.[2,12,13], Maria Ask, Jenny Bremer, Corrado Cilio, Ph.D., M.D.[5,6], Emelie Ericson-Hallström, Annika Fors, Lina Fransson, Thomas Gard, Rasmus Bennet, Monika Hansen, Susanne Hyberg, Hanna Jisser, Fredrik Johansen, Berglind Jonsdottir, M.D., Ph.D.[12], Silvija Jovic, Helena Elding Larsson, M.D., Ph.D.[6,14], Marielle Lindström, Markus Lundgren, M.D., Ph.D.[14], Maria Månsson-Martinez, Maria Markan, Jessica Melin[12], Zeliha Mestan, Caroline Nilsson, Karin Ottosson, Kobra Rahmati, Anita Ramelius, Falastin Salami, Anette Sjöberg, Birgitta Sjöberg, Carina Törn, Ph.D.[3,15], Anne Wallin, Åsa Wimar[14], Sofie Åberg. Lund University.

Washington Clinical Center: William A. Hagopian, M.D., Ph.D., PI[1,3,4,5,6,7,11,13,14], Michael Killian[6,7,12,13], Claire Cowen Crouch[12,14,15], Jennifer Skidmore[2], Ashley Akramoff, Masumeh Chavoshi, Kayleen Dunson, Rachel Hervey, Rachel Lyons, Arlene Meyer, Denise Mulenga[12], Jared Radtke, Matei Romancik, Davey Schmitt, Julie Schwabe, Sarah Zink. Pacific Northwest Research Institute.

Pennsylvania Satellite Center: Dorothy Becker, M.D., Margaret Franciscus, MaryEllen Dalmagro-Elias Smith[2], Ashi Daftary, M.D., Mary Beth Klein, Chrystal Yates. Children's Hospital of Pittsburgh of UPMC.

Data Coordinating Center: Jeffrey P. Krischer, Ph.D.,PI[1,4,5,10,11], Sarah Austin-Gonzalez, Maryouri Avendano, Sandra Baethke, Rasheedah Brown[12,15], Brant Burkhardt, Ph.D.[5,6], Martha Butterworth[2], Joanna Clasen, David Cuthbertson, Christopher Eberhard, Steven Fiske[9], Jennifer Garmeson, Veena Gowda, Kathleen Heyman, Belinda Hsiao, Christina Karges, Francisco Perez Laras, Hye-Seung Lee, Ph.D.[1,2,3,13,15], Qian Li[2,3], Shu Liu, Xiang Liu, Ph.D.[2,3], Kristian Lynch, Ph.D. [5,6,9,15], Colleen Maguire, Jamie Malloy, Cristina McCarthy[12,15], Aubrie Merrell, Steven Meulemans, Hemang Parikh, Ph.D.[3], Ryan Quigley, Cassandra Remedios, Chris Shaffer, Laura Smith, Ph.D.[9,12], Susan Smith[12,15], Noah Sulman, Ph.D., Roy Tamura, Ph.D.[1,2,12,13,14], Dena Tewey, Michael Toth, Ulla Uusitalo, Ph.D.[2,15], Kendra Vehik, Ph.D.[4,5,6,9,14,15], Ponni Vijayakandipan, Keith Wood, Jimin Yang, Ph.D., R.D.[2,15]. *Past staff: Michael Abbondondolo, Lori Ballard, David Hadley, Ph.D., Wendy McLeod.* University of South Florida.

Project scientist: Beena Akolkar, Ph.D.[1,3,4,5,6,7,10,11]. National Institutes of Diabetes and Digestive and Kidney Diseases.

Other contributors: Kasia Bourcier, Ph.D.[5], National Institutes of Allergy and Infectious Diseases. Thomas Briese, Ph.D.[6,15], Columbia University. Suzanne Bennett Johnson, Ph.D.[9,12], Florida State University. Eric Triplett, Ph.D.[6], University of Florida.

Committees:
[1]Ancillary Studies, [2]Diet, [3]Genetics, [4]Human Subjects/Publicity/Publications, [5]Immune Markers, [6]Infectious Agents, [7]Laboratory Implementation, [8]Maternal Studies, [9]Psychosocial, [10]Quality Assurance, [11]Steering, [12]Study Coordinators, [13]Celiac Disease, [14]Clinical Implementation, [15]Quality Assurance Subcommittee on Data Quality.