Published in final edited form as:

Biometrics. 2010 December ; 66(4): 1220–1229. doi:10.1111/j.1541-0420.2009.01369.x.

Design and Analysis of Multiple Events Case-Control Studies

Wenguang Sun^{*},

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

Marshall M. Joffe^{*},

Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA

Jinbo Chen, and

Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA

Steven M. Brunelli

Renal-Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA 19104, USA

Summary

In case-control research where there are multiple case groups, standard analyses fail to make use of all available information. Multiple events case-control (MECC) studies provide a new approach to sampling from a cohort and are useful when it is desired to study multiple types of events in the cohort. In this design, subjects in the cohort who develop any event of interest are sampled, as well as a fraction of the remaining subjects. We show that a simple case-control analysis of data arising from MECC studies is biased and develop three general estimating-equation based approaches to analyzing data from these studies. We conduct simulation studies to compare the efficiency of the various MECC analyses with each other and with the corresponding conventional analyses. It is shown that the gain in efficiency by using the new design is substantial in many situations. We demonstrate the application of our approach to a nested case-control study of the effect of oral sodium phosphate use on chronic kidney injury with multiple case definitions.

Keywords

Multiple events case-control study; case-cohort study; nested case-control study; sampling from a cohort; semi-parametric efficient estimator

1. Introduction

In large-scale cohort studies, investigators are often interested in assessing risk factors for several clinical endpoints. For example, the female textile workers cohort study (Ray et al. 2007; Astrakianakis et al. 2007; Li et al. 2006) investigates occupational risk factors for several cancers observed in the cotton textile industry, including breast cancer, lung cancer, pancreatic cancer and liver cancer. The Nurses Health Study (NHS, Stampfer et al. 1985) investigates risk factors for several major chronic diseases in women including cancer,

---

^{*}sun@stat.ncsu.edu, mjoffe@mail.med.upenn.edu.

7. Supplementary Materials

Web Appendices and Tables referenced in Sections 3 and 4 are available under the Paper Information link at the Biometrics website: <http://www.biometrics.tibs.org>.

diabetes and cardiovascular disease. The chronic renal insufficiency cohort study (CRIC, Feldman et al 2003) investigates risk factors for the cardiovascular disease and the end-stage renal disease in a clinical population with chronic renal insufficiency. These endpoints will occur only in a minority of the study subjects over the course of follow-up. Due to the limited study budget, it is cost effective to obtain covariate histories on subjects who have developed outcomes of interest, and a subset of the remaining subjects.

In case-control research, investigators often wish to reuse a control group that has been used for studying associations with one outcome for studying associations with other outcomes. Reusing a single control group for multiple case groups can lead to savings in cost. However, a control group that is appropriate for one case group may not be appropriate for a different one. The case-cohort design circumvents this problem by selecting as a referent group a random sample of the entire underlying cohort; this random sample can then be used with any or multiple case groups (Kupper, McMichael and Spirtas 1975; Prentice 1986; Rothman, Greenland, and Lash 2008). However, when there are multiple case groups, standard analyses of case-cohort studies fail to make use of all available information. In particular, subjects who are not selected for the referent group but who develop one of the outcomes of interest are ignored in the analysis of other outcomes.

To gain further insight into design and analysis of these studies, it is productive to consider the above paradigm as a distinct design, where sampling from the cohort is based not only on the outcome of interest (case status) for a particular analysis but also on events/outcomes auxiliary for that analysis. We term these studies multiple-events case-control (MECC) studies. In the MECC design, there is a known and enumerated cohort, from which subjects are selected for additional measurements; thus, the sampling proportions are known or can be estimated from the data.

The work on sampling from a cohort (e.g., Prentice 1986; Langholz and Thomas 1990) demonstrates that proper design and analysis for case-control and case-cohort studies can yield consistent estimates of the same measure of association in the underlying cohort studies. The MECC study can be conceptualized in this framework as well. However, this framework by itself can fail to use information available on the subjects not sampled. The missing data perspective on these studies supplements the cohort sampling one and allows use of data on all subjects in the cohort. For exposition of analytic methods, we adopt this view, and so consider and develop estimators based on the semi-parametric estimation theory by Robins, Rotnitzky and Zhao (1994).

The main contributions of this paper include the following: a. We propose a new unified sampling and analysis framework when multiple event types are sampled from a cohort. b. The difficulty of implementing the semi-parametric estimation theory of Robins et al. (1994) is well known; we provide an accessible illustration on how to apply the theory to a new situation. The software is also made available. c. We compare our analytic methods taking appropriate account of the design with conventional analytic alternatives using simulations. We show that in many situations, the gain in efficiency and reduction in bias can be substantial by using the new design. d. We evaluate the performance of different estimators, including the weighted, weighted-augmented and semi-parametric efficient estimators, in the new design and give practical recommendations. e. The comparison (of different estimators) in Robins et al. (1994) considered a random sampling scheme from the underlying cohort, and only the efficiency for estimating the effect of the missing covariate was investigated. We consider a comparison using different sampling methods (nested case-control and MECC sampling from the underlying cohort), and extend the previous comparison of analytic approaches by including the always observed covariate as well.

The paper is organized as follows. In Section 2, we introduce the MECC design in the frameworks of sampling from a cohort and of missing data. Three estimating-equation based approaches to analyzing the MECC data are described in Section 3. Simulation studies are conducted in Section 4 to compare the efficiency of the three analytical approaches, as well as the efficiency of the MECC analyses versus conventional analyses. In Section 5, we illustrate the use of MECC analysis in a nested case-control study with multiple case definitions. The article concludes with a discussion of the results and some open problems. Derivations and additional simulation results are given in the web supplementary materials.

2. The Conceptual Framework for MECC Studies

Sampling from a cohort has long been applied in epidemiologic studies; the best known examples are case-control and case-cohort study designs (Breslow and Day 1980; Prentice 1986; Rothman, Greenland, and Lash 2008). These designs allow cost effective inference for most or all population-level parameters of interest. Our newly proposed MECC study design also fits into this framework. This section briefly reviews the literature on sampling from a cohort and the missing-data view of this problem, then formulates the MECC design within these frameworks. Our subsequent treatment assumes that all outcomes are binary.

Let Y be the outcome of interest and S the auxiliary outcome/event. We are interested in studying associations between a collection of covariates $X^* = (X, V)$ and outcome Y , where X is the collection of covariates that are only available for a subset of the cohort and V the collection of covariates that are measured for the entire cohort. Denote by $W = (V, Y, S)$ the data that are always observed.

We consider here simple versions of the case-cohort and nested case-control designs (NCC, here we refer to the unmatched case-control study with controls selected from event-free subjects) in which the outcome is a binary indicator rather than a censored failure-time outcome. The case-cohort and nested case-control designs apply to both survival analysis and logistic regression analysis (see Rothman et al. 2008), although they were originally proposed for and are better known as methods for survival data.

In the case-cohort design, a sample or subcohort of the entire cohort is chosen to obtain information on X . Let $\delta_{CC} = 1$ if a subject is in the subcohort, and let $\pi_{V,CC}^* = P(\delta_{CC} = 1|V)$ denote the probability of being selected into the subcohort; this probability may be a function of the always-measured baseline covariates V , but not of the sometimes unmeasured covariate X , nor of post-treatment variables Y and S . In a case-cohort study, information on X is also obtained on all subjects with $Y = 1$. Let Δ_{CC} be an indicator of whether information on X is obtained in the simple case-cohort design. In this design, $\Delta_{CC} = I(\delta_{CC} = 1 \text{ or } Y = 1)$, and the probability of inclusion in the study

$\pi_{V,CC} = P(\Delta_{CC} = 1|X, V, S, Y) = P(\Delta_{CC} = 1|V, Y) = Y + (1 - Y)\pi_{V,CC}^*$. In nested case-control (NCC) studies, control subjects are selected from among the noncases; let Δ_{NCC} indicate whether covariates X are measured for a given subject. Here, we have

$\pi_{V,NCC} = P(\Delta_{NCC} = 1|X, V, S, Y) = P(\Delta_{NCC} = 1|V, Y) = Y + (1 - Y)\pi_{V,NCC}^*$, where $\pi_{V,NCC}^*$ is the sampling probability for the noncases. Thus, the sampling scheme is equivalent in the two designs (for simple binary outcomes). For binary outcomes, the main difference between the nested case-control design and the case-cohort design is that, in the case-control design, a random subcohort is selected in the case-cohort design while not in the nested case-control design. Two-phase designs (Breslow and Holubkov 1997; Scott and Wild 1997) generalize the above designs by allowing the probability of gathering information on X to depend on both Y and V , and to be less than 1 even when $Y = 1$.

Multiple events case-control (MECC) studies generalize the above designs by allowing sampling probabilities to depend on auxiliary events or outcomes S as well as on the event of interest Y and baseline covariates V . For binary S , we will typically choose to sample (measure X) for all subjects with $S = 1$, in addition to all subjects with $Y = 1$ and a subset of the remainder of the cohort. Let Δ_{MECC} be an indicator of whether covariates X are measured for a given subject in this design. Typically, we have

$$\pi_{V,Y,S,MECC} = P(\Delta_{MECC} = 1 | X, V, S, Y) = P(\Delta_{MECC} = 1 | V, S, Y) = I(Y=1 \text{ or } S=1) + I(Y=0 \text{ and } S=0)\pi_{V,MECC}^*,$$

where $\pi_{V,MECC}^*$ is the sampling probability for noncases without the auxiliary outcome.

Early analytic methods for the case-cohort and nested case-control designs adopted the view that these designs are based on sampling from a cohort. In particular, subjects for whom $\Delta = 1$ are sampled and so to be included in the analysis, whereas subjects for whom $\Delta = 0$ are not sampled and so not included. This view may stem from the view of these designs as variants of standard case-control studies, in which the sampling frame is not always fully specified and the analysis is confined to subjects with $\Delta = 1$.

An alternative view is that these are studies with data missing on X in the subset for whom $\Delta = 0$, and so methods for dealing with missing data are appropriate here. An advantage of such missing-data approaches is that they may use information on always-measured covariates for subjects with $\Delta = 1$ to obtain more efficient information about parameters of interest. Principled methods for dealing with missing data include multiple imputation, maximum likelihood, and estimating equations approaches. Likelihood-based approaches have been considered in this context in two-phase studies (e.g., Breslow and Cain (1988); Scott and Wild (1988; 1991; 1997); Breslow and Holubkov (1997)). Robins, Rotnitzky, and Zhao (1994) proposed an approach involving weighted and possibly augmented estimating equations, and similar ideas have been considered for both survival and binary outcomes (Fears and Brown 1986; Breslow and Chatterjee 1999; Lawless, Kalbfleisch and Wild 1999). Robins et al. (1994) explicitly considers the nested case-control sampling scheme and, in principle, also encompasses the MECC design outlined above. Even for the simpler case-cohort and NCC designs considered above where the nonparametric maximum likelihood approach (Scott and Wild, 1997) is applicable, an advantage of this approach is that it can be extended to continuous V . In this paper, we adopt this approach for the analysis of MECC design as well as for the simpler case-cohort and NCC analyses.

Instead of using the somewhat more complicated approach to MECC data that we will outline in Section 3, MECC data may be analyzed as case-cohort studies. This may be done simply if a random subcohort is selected and so subjects with $\delta_{CC} = 1$ are identified even when $Y = 0$ and $S = 1$. Even in the absence of a random subcohort identified a priori, this may also be done by randomly sampling from this subset with the same probability as the selection probability for the noncases in the same stratum; i.e., $P(\Delta = 1 | V, Y = 0, S = 1) = P(\Delta = 1 | V, Y = 0, S = 0)$. We will compare our methods for MECC analysis to the simpler analyses treating these as case-cohort studies (referred to as CC1) to consider the advantages of our approach over the simpler case-cohort approach.

Investigators contemplating a MECC study in which a single endpoint Y is the primary focus might want to consider a simpler case-cohort or NCC design in which the sampling is based on Y but not S . We might anticipate that such a design, while severely restricting our ability to study the associations of X and V with the auxiliary outcome S , might be more cost-efficient in studying the associations of these variables with the primary outcome Y . In such a design, we could increase the sampling fraction of subjects with $Y = 0$ from $\pi_{V,MECC}^*$ to $\pi_{V,CC2}^* = \pi_{V,MECC}^* + P(S=1, Y=0)(1 - \pi_{V,MECC}^*)$ without increasing the number of subjects from whom data on X are to be collected. We call a nested case-control study with sampling

fraction π_{CC2}^* the *CC2 alternative*. The comparison of MECC with CC2 would be of interest, for example, if one has a fixed budget and wants to consider the tradeoff between the MECC design for studying two outcomes and a nested case-control design for studying one outcome of primary interest. The numerical results for comparison of MECC versus CC1 and CC2 are provided in Section 4.

3. The Analytic Approaches to MECC Studies

In this section, we outline three unbiased estimating equation approaches to analyzing data from MECC studies, based on the theory of Robins et al. (1994): a simple approach using a weighted estimating function, a more complicated one in which the weighted estimating function is augmented, and the most complicated approach, using the optimal weighted and augmented function. In this approach, the observable and latent data (X, V, S, Y, Δ) are taken to be an i.i.d random vector. We consider linear logistic regression models for the outcome on covariates measured at baseline: $\text{logit}\{E(Y|X^*)\} = X^*\beta$. X^* can be replaced by $q(X^*)$, a known function of X^* , to allow for interactions or nonlinearities in X^* .

In case-control studies, a common approach is to perform a unweighted logistic regression analysis using subjects for whom full covariate information is available; i.e., $\text{logit}\{P(Y=1|X, V, \Delta_{CC}=1)\} = X^*\gamma_{CC}$. In these analyses, the estimate of the intercept is biased, but the estimates of the rest of the regression parameters are unbiased. However, in MECC studies, all estimates of the parameters γ_{MECC} in the regression model $\text{logit}\{P(Y=1|X, V, \Delta_{MECC}=1)\} = X^*\gamma_{MECC}$ are biased. This is illustrated in the simulation studies conducted in Section 4 (see Table 2).

Proper weighting methods may be applied through the estimating equation approach to eliminate the bias (e.g., Horvitz and Thompson 1952; Flanders and Greenland 1991; Zhao and Lipsitz 1992). Let $U_i(\beta)$ be the score function of the fully observable data for subject i and the unknown parameter β such that $E\{U_i(\beta)\} = 0$ when β is evaluated at its true value. In a logistic regression model, $U_i(\beta)$ is $\{Y_i - E(Y_i|X_i^*)\}X_i^*$. Denote by $\pi(W_i)$ the sampling probability for subject i . Then it is easy to show that the weighted estimating equation

$$D_{wt}(\beta) = \sum_i U_i(\beta) \Delta_i / \pi(W_i) = 0 \quad (1)$$

is unbiased. The estimator solving (1) is referred to as the weighted estimator (WTE). We summarize the weights used to estimate the regression parameters in MECC analysis in Table 1, where we also contrast the MECC weights (also the sampling weights) with weights used in case-control analysis. It was shown by Robins et al. (1994) that estimating the selection probabilities can improve efficiency even when these probabilities are known. We hence use estimated weights in estimating equations (1), (2) and (4) in subsequent simulation study and data analysis.

The information of the subjects who are not selected to the MECC study is ignored in the analysis using (1). The weighted approaches are not fully efficient for the estimation of model parameters, and the general approach to improving the efficiency involves augmenting the weighted estimating equation by additional terms. Let $\phi(W)$ be a function of W and $A(\phi) = \{\Delta - \pi(W)\}\phi(W)/\pi(W)$. We augment the estimating equation (1) to obtain

$$D_{wa}(\beta, \varphi) = \sum_i \{U_i(\beta)\Delta_i/\pi(W_i) - A_i(\varphi_i)\} = 0. \quad (2)$$

The estimator that solves (2) is referred to as the weighted-augmented estimator (WAE). The function φ does not depend on the missing covariates X , and for a given estimating function $U(\beta)$, the optimal choice of $\varphi(W)$ is $\varphi(W) = E\{U(\beta)|W\}$. The WAE is unbiased, and, by including an augmented term in the weighted estimating equation, we can reduce the variance of the weighted estimator by considering additional information in subjects with incomplete data. Our simulation studies in Section 4 show that the WAEs are, in general, more efficient than the WTEs.

In the presence of missing data, the most efficient approach involves choosing a new $U_i(\beta)$ and optimally augmenting the estimating function (Robins et al. 1994). Consider a general non-linear regression model

$$Y = g(X^*; \beta) + \varepsilon. \quad (3)$$

Let $h(X^*)$ be a function of all covariates X^* , $U(\beta, h) = h(X^*)\{Y - g(X^*; \beta)\}$ an unbiased estimating function, $\varphi(W)$ a function of W and $A(\varphi) = \{\Delta - \pi(W)\}\varphi(W)/\pi(W)$. A class of estimators $\beta(h, \varphi^h)$, including the WTE and WAE as special cases, solves the following estimating equation

$$D(\beta, h, \varphi^h) = \sum_i \{U_i(\beta, h)\Delta_i/\pi(W_i) - A_i(\varphi_i)\} = 0. \quad (4)$$

Denote by $\beta(h_E, \varphi_E)$ the optimal estimator in this class. For a logistic regression model with linear predictors, we have $h_E = X^*$ and $\varphi_E = 0$ when no data are missing. For fixed h , the optimal choice for φ is $\varphi^h = E[U(\beta, h)|W]$. Robins et al. (1994) showed that the optimal estimator $\beta(h_E, \varphi_E)$, also referred to as the semi-parametric efficient estimator (SEE), is unique in the class of estimators defined by (4), and achieves the semi-parametric efficiency bound. The WTE β_{wt} and WAE β_{wa} can be identified as special cases of $\beta(h, \varphi^h)$, that is, $\beta_{wt} = \beta(X^*, 0)$ and $\beta_{wa} = \beta(X^*, \varphi^{X^*})$. The derivation of the SEE is usually complicated, involving an iterative procedure to obtain h_E and φ_E . Illustration on how to adaptively obtain SEE in a simple logistic regression model is provided in the web supplementary materials.

4. Simulation

In this section, we conduct simulation studies to examine (i) the efficiency of various types of weighted, possibly augmented estimators (WTE, WAE, and SEE), both for MECC analyses and for CC analyses of MECC data, (ii) the efficiency gained by analyzing MECC data as MECC data rather than as case-cohort or nested case-control data (CC1), and (iii) the efficiency implications of using the MECC design instead of the single event case-control study with the same overall sampling proportion (CC2).

To allow our simulation model to be consistent with our model for analyzing $P(Y|X, V)$, the following factorization of the joint distribution is chosen to simulate X, V, Y and S :

$$P(X, V, Y, S) = P(X)P(V|X)P(Y|X, V)P(S|X, V, Y). \quad (5)$$

In our simulation, we shall consider both continuous and binary X (i.e., $X \sim N(0, 1)$ and $X \sim \text{Bernoulli}(0.3)$). Other variables are binary and generated according to the following logistic regression models:

$\text{logit}\{E(V|X)\} = \text{logit}(p_v) = X_v^* \beta_v$, $\text{logit}\{E(Y|X, V)\} = \text{logit}(p_y) = X_y^* \beta_y$, $\text{logit}\{E(S|X, V, Y)\} = \text{logit}(p_s) = X_s^* \beta_s$, where p_v , p_y and p_s are the conditional expectations of V , Y and S , and

$X_v^* = (1, X)$, $X_y^* = (1, X, V)$ and $X_s^* = (1, X, V, Y)$. The logistic regression model for the association between Y and $X^* = (X, V)$ is $\text{logit}[E(Y|X, V)] = \beta_0 + \beta_1 X + \beta_2 V$. The parameter of interest is $\beta_y = (\beta_0, \beta_1, \beta_2)$, and other parameters (β_v, β_s) are treated as nuisance parameters. Various settings are considered to investigate how the comparison results vary according to the correlation between (i) X and V , (ii) Y and X , (iii) S and X , and (iv) Y and S . The parameters for simulation are chosen such that the populations of cases ($Y = 1$) range from 4% to 15%. In all simulations, the cohort size is $N = 800$ and the number of replications is 1000. We noticed that for small cohort size and low event rate (cohort size=800, event rate<0.06), the SEE method fails to converge in about 5 percent out of the 1000 replications. The results from these replications have been excluded while producing our tables.

4.1 Comparison of Estimators in MECC Analysis

One naive MECC approach is to treat the MECC data as if they arose from a case-control study. In case-control studies, an unadjusted logistic regression analysis (or the maximum likelihood approach) produces unbiased estimates of covariate effects (but not of the intercept), as is noted by Prentice and Pyke (1979). However, the maximum likelihood approach for the MECC model $P(Y|X, V, \Delta_{MECC} = 1)$ is biased since the controls from MECC studies, unlike those from conventional case-control studies, do not represent a random sample from the noncases. The estimate obtained using this biased approach is denoted by MLE_{MECC} . Other estimators being considered for comparison include WTE, WAE and SEE.

We consider comparisons for both binary and continuous X . The simulation results for binary X are summarized in Table 2. Additional comparisons of MLE, WTE and WAE for continuous X are provided in Table 3 (and Table 2 in supplementary materials). We can see from Table 2 that the MLE_{MECC} is biased, whereas the WTE, WAE and SEE (of the MECC analysis) are asymptotically unbiased. Relative efficiencies of different MECC estimators are also shown, with the WTE of CC1 analysis as the referent method. The following observations can be made based on the simulation results from Table 2:

1. The WAE improves the WTE on estimation of β_0 and β_2 but not on estimation of β_1 . In fact, the WAE and WTE are asymptotically equivalent in estimating the effect of the incomplete covariate (β_1) when estimated selection probabilities are used in the estimating equation, as is indicated by Corollary 6.1 in Robins et al. (1994).
2. The efficiency for estimating the complete covariate (β_2) is not discussed in Robins et al. (1994). Table 2 shows that the efficiency gain of WAE over WTE in estimating β_2 is large when (a) X and V are weakly correlated, and (b) X and Y are strongly associated (Setting 1–3).
3. The SEE further improves the WAE in the estimation of all three parameters (including β_1). The efficiency gain of SEE over WAE is large when (a) Y and X are correlated (Setting 1–3, 6–8), and (b) S and X are uncorrelated (Setting 3–5).

4.2 Comparison of MECC design and analysis with alternatives

We report here comparisons of MECC analysis with corresponding CC1 analysis of the same data and with case-control analysis of the CC2 design. MECC versus CC1 is the primary comparison in this work.

The results for comparison are shown in Table 3 and Table 4, respectively for binary X and continuous X (for which only WTE and WAE are considered). We consider a situation where V and X are strongly correlated. Simulation results for weakly correlated V and X are provided Table 1 and Table 2 in the supplementary materials. We mainly discuss the results in Table 3, and similar conclusions can be drawn from other tables. Table entries are the relative efficiencies for estimating β_y with the WTE of CC1 as the referent method. For each comparison, we also list the average proportion of subjects with $\Delta_j = 1$, where j may be MECC, CC1, or CC2. These proportions are denoted by p_{CC1} , p_{MECC} and p_{CC2} , respectively for corresponding designs and analyses.

The following observations can be made based on the simulation results:

1. MECC analysis is more efficient than CC1 analysis in estimating β_1 and β_2 for the corresponding analytic approaches (WTE, WAE and SEE). The gain in efficiency is large when (a) Y and X are moderately or strongly correlated (Setting 1–3, Table 3), and (b) S and X are strongly correlated (Setting 6–8, Table 3)
2. If the goal is solely to estimate associations with the main outcome Y , CC2 analysis is more efficient than MECC analysis in most situations. The gain in efficiency is large when (a) Y and X are uncorrelated or weakly correlated (Setting 1–3, Table 3), (b) Y and S are moderately or strongly correlated, and (c) S and X are uncorrelated or weakly correlated.
3. It is interesting that MECC can be more efficient than CC2 in situations when both the associations between (a) Y and X , and (b) S and X are strong (Setting 3, Table 3; Setting 2–5, Table 4). This is because the subjects for whom $\Delta_{CC1} = 0$ and $S = 1$ provide more information about β_1 than the same number of subjects randomly sampled from the noncases ($Y = 0$).

The confidence intervals of all estimators (obtained based on the estimated robust variance-covariance matrices), reported in Table 3 and 4 in the web supplementary materials, show that most confidence intervals have slightly higher coverage probabilities than the nominal level 95%. We also reported in Table 5 in the web supplementary materials for comparison of the empirical variance vs. estimated variance. The results show that the asymptotic approximation (used in the derivation of SEE) is improved slightly with increased sample sizes. For the simulation settings being considered (the numbers of events vary from 40 to 300), the asymptotic approximations are good.

5. MECC Analysis of a Nested Case-Control Study

So far we have discussed the use of MECC design in a cohort study with *multiple* types of events. However, “multiple” can be understood in a broader sense. In this section we discuss an interesting application of MECC design in a nested case-control study where only one type of event (chronic kidney injury) is of interest.

Colonoscopy is the diagnostic test of choice for many diseases of the lower gastrointestinal tract, and is among the preferred screening modalities for colorectal cancer. Prior to colonoscopy, patients need to undergo bowel preparation, and the sensitivity of colonoscopy is dependent upon the quality of this preparation. Bowel purging is most often achieved using either oral sodium phosphate (OSP)-based or polyethylene glycol (PEG)-based agents.

The former is often preferred due to its greater efficacy, cost-effectiveness and patient tolerability. Recent case reports and series have suggested a potential association between OSP preparations and chronic kidney disease. Given the number of patients undergoing colonoscopy (14 million annually) and the clinical stakes of missed lesions, clarification of this potential risk is of great clinical importance.

Based on the need for manual abstraction of exposure data (OSP versus PEG), study of this association took the form of a nested case-control study (Brunelli et al. 2007). Unfortunately, there is no consensus definition of incident chronic kidney disease in the clinical literature, necessitating that case status be adjudicated according to a clinically “reasonable” definition. In evaluating the association between OSP use and acute kidney injury, Hurst et al. (2007) determined that a 50% rise in serum creatinine (“strict case definition”) would represent a significant loss of kidney function. In contrast, Brunelli et al. (2007) determined cases to be a 25% rise in serum creatinine (“loose case definition”) based on: (i) a limited number of cases existed under the “strict case” definition, (ii) the consensus among study investigators that this still represented a clinically significant reduction in kidney function, and (iii) some precedent for this definition in the clinical literature. This “loose case” definition is hence used in selecting the subjects and subsequent association analysis (Brunelli et al. 2007). Further investigations were undertaken to compare Brunelli’s study with Hurst’s study (Brunelli et al. 2008). Various potential sources of discrepancy between the two studies were considered in the sensitivity analysis of Brunelli’s data. In particular, the investigators are interested in the impact of changing the case definition on subsequent analysis results.

Here we briefly describe the original sampling scheme and analytical methods using the “loose case” definition and an MECC reanalysis of data using the “strict case” definition. The nested case-control study conducted by Brunelli et al. (2007) enrolled 2237 subjects in the source cohort. The loose case definition was met in 141 instances; of these, bowel preparation data were available for 116. Among the 2096 colonoscopies for which the loose case definition was not met, 398 are randomly sampled and covariate data are available for 349. Patients for whom exposure data were not available did not differ from those for whom they were, see Table 1 in Brunelli et al. (2007). Among the $n = 465$ subjects for whom the covariate data are available, the strict case definition is again met in 26. In our new analysis, the outcome of interest Y is the “strict case” and the auxiliary outcome S is the “loose case”. Although this is not a typical “multiple events” study (since $Y = 1$ implies $S = 1$), it naturally falls into the category of our MECC design, where information on OSP use is obtained on subjects with either the main outcome or the auxiliary outcome, as well as a random subset of subjects without either outcome.

Table 2 in Brunelli et al. (2007) presents a comparison of baseline characteristics of cases and controls, showing significant differences for gender and exposure to diuretics. Hence we consider the following logistic regression model:

$$\text{logit}\{E(Y=1|\text{OSP}, \text{Diuretic}, \text{Gender})\} = \beta_0 + \beta_1 \text{OSP} + \beta_2 \text{Diuretic} + \beta_3 \text{Gender}. \quad (6)$$

Let $X = (\text{diuretic}, \text{OSP})$ be the incomplete covariate (the exposure information of a patient to OSP or diuretic is only available for subjects who are selected) and $V = \text{Gender}$, the covariate available for all subjects.

In reanalysis with strict case definition, a naive approach is to fit a logistic regression model using all the $n = 465$ subjects collected for studying the loose case association (MLE_{MECC}). However, this analysis is biased for studying the strict case association. Alternately, we may

randomly select from the 90 loose cases (for whom $S = 1$ and $Y = 0$) with probability $398/2096$ (10 subjects were finally selected), and add them to the original 349 controls to recreate a random subset of the noncases (NCC/CC1). The NCC approach is unbiased but not efficient. An MECC analysis increases the efficiency by including in the study additional loose cases (80 subjects) who were not sampled in NCC. We implemented the weighted, weighted-augmented approaches in both loose and strict case analyses. The standard errors of the point estimates are obtained based on the robust variance-covariance estimates (Web Appendix D in the supplementary materials).

The results for all analyses are summarized in Table 5. We can see that the standard errors of the estimators in the MECC analyses are slightly smaller than those in the corresponding NCC analyses. In addition, the weighted-augmented approach significantly improves the weighted approach in estimating all regression parameters. The new analysis indicates no association between OSP use and CKD, as is originally reported by Brunelli et al. (2007) using the loose case definition. However, the significantly reduced number of cases (from $n = 116$ to $n = 26$) may have limited the power of detecting the association.

6. Discussion

The MECC design is useful when it is desirable to study more than one outcome in the cohort. It extends the traditional case-cohort design in that it not only uses the same subgroup of controls for studying several outcomes, but also uses subjects with one outcome to study other outcomes. A more general version of the design allows the sampling probabilities to depend on (V, Y, S) . The case-control follow-up study (Weiss and Lazovich 1996; Joffe 2003), where sampling is based on cancer diagnosis (an auxiliary outcome) but the outcome of interest is cancer mortality, may be viewed as a variant of the MECC study in which the sampling fraction may be unknown.

Our simulation results show that the new design, equipped with the analytical approaches in Robins et al. (1994), can eliminate the bias of some traditional methods and improve the efficiency of the CC1 analysis of the same data. If the goal is solely to estimate associations with the main outcome Y , the MECC design is in general inferior compared to CC2 design unless the incomplete covariate (X) is strongly associated with both outcomes $(Y$ and $S)$. In addition, our simulation results indicate that augmenting the estimating equations is beneficial, requiring relatively little additional effort, whereas the semi-parametric efficient estimator, which requires substantial additional effort to obtain, sometimes provides little additional efficiency gain. The augmented approach makes use of information on the subjects with incomplete data, and can substantially improve the precision in estimating the effect of the always observable variables. The implementation of this approach is simple especially when W is discrete. However, this approach has scarcely been used in nested case-control research despite its advantages.

The MECC design is a sort of two-phase design (e.g., Breslow and Holubkov 1997; Scott and Wild 1997; White 1982; Breslow and Cain 1988; and Flanders and Greenland 1991), in which sampling depends on outcome Y and other variables. Other methods, including likelihood and pseudolikelihood-based approaches have been proposed for some such designs (e.g., Breslow and Holubkov 1997; Scott and Wild 1997). These approaches produce asymptotically efficient estimators and the implementations are not so complicated as the SEE. However, these methods require that the variables stratified for sampling (except the outcome variable) are independent of the disease outcome variable conditional on the covariates. This condition will typically fail to hold in the MECC design since the auxiliary outcome variable could still be correlated with the outcome variable of interest even after adjusting for other covariates (e.g., the example discussed in Section 5). It would be of

interest for future research to implement these likelihood approaches to model sampling scheme that also depends on auxiliary outcomes.

Nested case-control and case-cohort studies often take place where a failure-time outcome rather than a simple binary event is of interest, and inference is often performed for parameters in semiparametric proportional hazards models. The simple weighting approach discussed here extends naturally to this setting. However, extensions of the more efficient augmented approaches will require further consideration.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the Associate Editor and the referee for detailed and constructive comments, which have greatly helped to improve the presentation of the paper.

References

- Astrakianakis G, Seixas N, Ray R, Camp J, Gao D, Feng Z, Li W, Wernli K, Fitzgibbons E, Thomas D, Checkoway H. Lung cancer risk among female textile workers exposed to endotoxin. *Journal of the National Cancer Institute*. 2007; 99:357–64. [PubMed: 17341727]
- Breslow, N.; Day, N. *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*. Lyon, France: International Agency for Research on Cancer; 1980.
- Breslow N, Cain K. Logistic Regression for Two-Stage Case-Control data. *Biometrika*. 1988; 75:11–20.
- Breslow N, Chatterjee N. Design and Analysis of Two-phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis. *Journal of the Royal Statistical Society, Series C*. 1999; 48:457–468.
- Breslow N, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B*. 1997; 59:447–461.
- Brunelli SM, Lewis JD, Gupta M, Latif SM, Weiner MG, Feldman HI. Risk of kidney injury following oral phosphosoda bowel preparations. *Journal of the American Society of Nephrology*. 2007; 18:3199–3205. [PubMed: 17978309]
- Brunelli, SM.; Lewis, JD.; Lynch, KE.; Joffe, MM.; Gupta, M.; Latif, SM.; Weiner, MG.; Feldman, HI. Technical Report. Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania; 2008. Further investigation of the association between oral sodium phosphate use and kidney injury.
- Feldman H, Appel L, Chertow G, Cifelli D, Cizman B, Daugirdas J, Fink J, Franklin-Becker E, Go A, Hamm L, He J, Hostetter T, Hsu C, Jamerson K, Joffe M, Kusek W, Landis J, Lash J, Miller E, Mohler E, Muntner P, Ojo A, Rahman M, Townsend R, Wright J. The Chronic Renal Insufficiency Cohort (CRIC) Study: Design and Methods. *Journal of the American Society of Nephrology*. 2003; 14:148–153. [PubMed: 12506147]
- Fears T, Brown C. Logistic Regression Methods for Retrospective Case-Control Studies Using Complex Sampling Procedures. *Biometrics*. 1986; 42:955–960. [PubMed: 3814735]
- Flanders W, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*. 1991; 10:739–47.
- Horvitz D, Thompson D. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
- Hurst FP, Bohen EM, Osgard EM, Oliver DK, Das NP, Gao SW, Abbott KC. Association of oral sodium phosphate purgative use with acute kidney injury. *Journal of the American Society of Nephrology*. 2007; 18:3192–3198. [PubMed: 17978311]

- Joffe M. A case-control follow-up study for disease-specific mortality. *Biometrics*. 2003; 59:115–125. [PubMed: 12762448]
- Kupper L, McMichael A, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. *Journal of the American Statistical Association*. 1975; 70:524–528.
- Langholz B, Thomas D. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology*. 1990; 131:169–176. [PubMed: 2403467]
- Lawless J, Kalbfleisch J, Wild C. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B*. 1999; 61:413–438.
- Li W, Ray M, Gao D, Fitzgibbons E, Seixas N, Camp J, Wernli K, Astrakianakis G, Feng Z, Thomas D, Checkoway H. Occupational risk factors for pancreatic cancer among female textile workers in Shanghai, China. *Occupational and Environmental Medicine*. 2006; 63:788–93. [PubMed: 16847032]
- Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–411.
- Ray M, Gao L, Li W, Wernli K, Astrakianakis G, Seixas S, Camp J, Fitzgibbons E, Feng Z, Thomas B, Checkoway H. Occupational exposures and breast cancer among women textile workers in Shanghai. *Epidemiology*. 2007; 18:383–92. [PubMed: 17435449]
- Robins J, Rotnitzky A, Zhao L. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Rothman, K.; Greenland, S.; Lash, T. *Modern Epidemiology*. 3. Lippincott-Raven: Philadelphia; 2008.
- Scott A, Wild C. Hypothesis Testing in Case-Control Studies. *Biometrika*. 1989; 76:806–808.
- Scott A, Wild C. Fitting logistic models in stratified case-control studies. *Biometrics*. 1991; 47:497–510.
- Scott A, Wild C. Fitting regression models to case-control data by maximum likelihood. *Biometrika*. 1997; 84:57–71.
- Stampfer M, Willett W, Colditz G, Rosner B, Speizer F, Hennekens C. A prospective study of postmenopausal estrogen therapy and coronary heart disease. *New England Journal of Medicine*. 1985; 313:1044–1049. [PubMed: 4047106]
- Weiss N, Lazovich D. Case-control studies of screening efficacy: the use of persons newly diagnosed with cancer who later sustain an unfavorable outcome. *American Journal of Epidemiology*. 1996; 143:319–322. [PubMed: 8633615]
- White J. A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*. 1982; 115:119–128. [PubMed: 7055123]
- Zhao L, Lipsitz S. Designs and analysis of two-stage studies. *Statistics in Medicine*. 1992; 11:769–82.

Table 1

Weights used in the MECC and NCC sampling and analysis

Analysis Type		Nested Case-Control			MECC		
Type of Subject		Sampling Prob.	Δ_{CC}	Weights	Sampling Prob.	Δ_{MECC}	Weights
$Y = 1$		1	1	1	1	1	1
	$Y = 0, S = 1$	π_{CC}	1	$1/\pi_{CC}$	1	1	1
$Y = 0, S = 0$			0	0		1	1
	$Y = 0, S = 0$	π_{CC}	1	$1/\pi_{CC}$	π_{MECC}	1	$1/\pi_{MECC}$
			0	0		0	0

Table 2

Comparison of different estimators in MECC analyses for binary X : $X \sim \text{Bernoulli}(0.3)$. The MLE is biased, whereas the WTE, WAE and SEE are asymptotically unbiased. The WTE is improved by the WAE, which is again improved by the SEE. pCC1 (PMECC) is the average proportion of subjects with $\Delta_{CC1} = 1$ ($\Delta_{MECC} = 1$).

	Bias			Relative Efficiency				Proportion	
	MLE _{MEC}	WTE	WAE	SEE	WTE _{CC1}	WTE	WAE	SEE	pCC1 pMECC
<i>X and V are strongly correlated: $\beta_v = (-2, 3)$</i>									
<i>Setting 1: $\beta_y = (-3, 0, 0), \beta_s = (-3, 0, 1, 0)$</i>									
β_0	2.34	-0.01	-0.03	-0.03	1.00	1.01	1.20	1.20	
β_1	0.00	0.00	0.00	-0.01	1.00	1.13	1.09	1.31	0.09 0.16
β_2	-0.55	0.01	-0.04	-0.04	1.00	1.13	1.28	1.56	
<i>Setting 2: $\beta_y = (-3, 1, 0), \beta_s = (-3, 0, 1, 0)$</i>									
β_0	2.13	-0.01	-0.02	-0.02	1.00	0.99	1.12	1.14	
β_1	0.03	0.08	0.08	0.04	1.00	1.09	1.09	1.31	0.14 0.20
β_2	-0.48	-0.06	-0.09	-0.06	1.00	1.11	1.26	1.60	
<i>Setting 3: $\beta_y = (-3, 2, 0), \beta_s = (-3, 0, 1, 0)$</i>									
β_0	1.76	-0.01	-0.02	-0.02	1.00	0.98	1.06	1.08	
β_1	0.02	0.05	0.05	0.02	1.00	1.04	1.03	1.17	0.23 0.28
β_2	-0.33	-0.03	-0.05	-0.03	1.00	1.07	1.33	1.67	
<i>Setting 4: $\beta_y = (-3, 2, 0), \beta_s = (-3, 1, 1, 0)$</i>									
β_0	1.77	-0.02	-0.03	-0.03	1.00	1.02	1.06	1.06	
β_1	-0.33	0.06	0.07	0.04	1.00	1.10	1.10	1.21	0.23 0.30
β_2	-0.41	-0.03	-0.05	-0.02	1.00	1.16	1.50	1.67	
<i>X and V are weakly correlated: $\beta_v = (-0.5, 1)$</i>									
<i>Setting 5: $\beta_y = (-3, 0, 0), \beta_s = (-3, 0, 1, 0)$</i>									
β_0	2.35	-0.01	-0.04	-0.05	1.00	1.00	1.23	1.23	
β_1	0.01	0.02	0.02	0.01	1.00	1.06	1.03	1.14	0.09 0.17
β_2	-0.58	-0.03	-0.03	-0.02	1.00	1.05	1.68	1.84	
<i>Setting 6: $\beta_y = (-3, 1, 0), \beta_s = (-3, 0, 1, 0)$</i>									
β_0	2.12	-0.02	-0.04	-0.05	1.00	1.02	1.15	1.17	0.14 0.21

	Bias			Relative Efficiency					Proportion	
	MLE _{MEC}	WTE	WAE	SEE	WTE _{CC1}	WTE	WAE	SEE	p _{CC1}	p _{MECC}
β_1	0.03	0.05	0.05	0.03	1.00	1.08	1.06	1.19		
β_2	-0.46	-0.02	-0.01	0.01	1.00	1.10	1.60	2.04		
Setting 7: $\beta_y = (-3, 2, 0), \beta_s = (-3, 0, 1, 0)$										
β_0	1.76	-0.02	-0.03	-0.02	1.00	1.03	1.16	1.23		
β_1	0.04	0.05	0.02	0.02	1.00	1.05	1.04	1.09	0.23	0.29
β_2	-0.32	-0.01	-0.01	-0.01	1.00	1.06	1.40	1.74		

Table 3

Comparison of different study designs/analyses for discrete X: $X \sim \text{Bernoulli}(0.3)$. X and V are strongly correlated: $\beta_V = (= 2, 3)$. p_{CC1} (p_{CC2} , p_{MECC}) is the average proportion of subjects with $\Delta_{CC1} = 1$ ($\Delta_{CC2} = 1$, $\Delta_{MECC} = 1$). Table entries are the relative efficiencies of different estimators.

	MLE			WTE			WAE			SEE		
	CC1	CC2	CC1	MEC	CC1	MEC	CC2	CC1	MEC	CC2	CC1	MEC
<i>Setting 1: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 0, 1, 0)$, $p_{CC1} = 0.09$, $p_{MEC} = p_{CC2} = 0.16$</i>												
β_0	0.01	0.01	1.00	1.01	1.13	1.20	1.20	1.26	1.20	1.20	1.20	1.26
β_1	1.20	1.91	1.00	1.13	1.86	0.98	1.09	1.85	1.29	1.31	1.94	1.94
β_2	1.21	1.79	1.00	1.13	1.75	1.14	1.28	2.00	1.39	1.56	2.11	2.11
<i>Setting 2: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 3, 0, 0)$, $p_{CC1} = 0.14$, $p_{MEC} = p_{CC2} = 0.20$</i>												
β_0	0.01	0.01	1.00	0.99	1.11	1.12	1.12	1.16	1.14	1.14	1.17	1.17
β_1	1.22	1.83	1.00	1.09	1.77	1.00	1.09	1.77	1.29	1.31	1.85	1.85
β_2	1.23	1.68	1.00	1.11	1.60	1.12	1.26	1.88	1.52	1.60	2.02	2.02
<i>Setting 3: $\beta_Y = (-3, 2, 0)$, $\beta_S = (-3, 3, 0, 0)$, $p_{CC1} = 0.23$, $p_{MEC} = p_{CC2} = 0.35$</i>												
β_0	0.01	0.02	1.00	1.00	1.07	1.07	1.07	1.10	1.08	1.08	1.11	1.11
β_1	1.13	1.82	1.00	1.53	1.77	1.00	1.53	1.77	1.15	1.56	1.82	1.82
β_2	1.19	1.99	1.00	1.55	1.90	1.26	1.94	2.24	1.64	2.10	2.38	2.38
<i>Setting 4: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 3, 0, 1)$, $p_{CC1} = 0.21$, $p_{MEC} = p_{CC2} = 0.39$</i>												
β_0	0.01	0.03	1.00	1.02	1.18	1.18	1.16	1.22	1.19	1.16	1.22	1.22
β_1	1.13	1.96	1.00	1.44	1.96	0.99	1.42	1.96	1.17	1.49	1.96	1.96
β_2	1.14	2.06	1.00	1.50	2.06	1.21	1.85	2.27	1.40	1.97	2.27	2.27
<i>Setting 5: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 3, 0, 2)$, $p_{CC1} = 0.14$, $p_{MEC} = p_{CC2} = 0.29$</i>												
β_0	0.01	0.02	1.00	1.02	1.16	1.13	1.14	1.20	1.16	1.16	1.21	1.21
β_1	1.14	2.01	1.00	1.45	2.03	0.99	1.46	2.03	1.21	1.51	2.03	2.03
β_2	1.17	2.00	1.00	1.45	1.99	1.19	1.80	2.22	1.43	1.91	2.24	2.24
<i>Setting 6: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 2, 0, 0)$, $p_{CC1} = 0.14$, $p_{MEC} = p_{CC2} = 0.23$</i>												
β_0	0.01	0.02	1.00	1.01	1.16	1.20	1.22	1.26	1.20	1.23	1.27	1.27
β_1	1.09	1.73	1.00	1.22	1.73	0.99	1.20	1.73	1.13	1.24	1.74	1.74
β_2	1.11	1.69	1.00	1.16	1.68	1.56	1.74	2.06	1.87	1.91	2.09	2.09
<i>Setting 7: $\beta_Y = (-3, 1, 0)$, $\beta_S = (-3, 1, 0, 0)$, $p_{CC1} = 0.14$, $p_{MEC} = p_{CC2} = 0.21$</i>												

	MLE		WTE		WAE		SEE	
	CC1	CC2	CC1	MEC	CC1	MEC	CC1	CC2
β_0	0.01	0.02	1.00	1.01	1.16	1.22	1.25	1.30
β_1	1.07	1.55	1.00	1.06	1.53	1.04	1.10	1.56
β_2	1.10	1.56	1.00	1.05	1.52	1.68	2.02	2.14

Table 4

Comparison of different study designs/analyses for continuous X : $X \sim N(0, 1)$. X and V are strongly correlated: $\beta_V = (-2, 3)$. p_{CC1} (p_{CC2} , p_{MECC}) is the average proportion of subjects with $\Delta_{CC1} = 1$ ($\Delta_{CC2} = 1$, $\Delta_{MECC} = 1$). Columns 2–7 are the relative efficiencies of different estimators.

Designs	WTE			WAE			Sampling Proportion		
	CC1	MECC	CC2	CC1	MECC	CC2	p_{CC1}	p_{CC2}	$p_{MECC} \approx p_{CC2}$
Setting 1: $\beta_y = (-2.5, 0, 0)$, $\beta_s = (-2.5, 3, 0, 0)$									
β_0	1.00	1.06	1.21	1.12	1.23	1.30			
β_1	1.00	1.24	1.78	1.00	1.26	1.78	0.24		0.42
β_2	1.00	1.18	1.61	1.26	1.53	1.76			
Setting 2: $\beta_y = (-2.5, 1, 0)$, $\beta_s = (-2.5, 3, 0, 0)$									
β_0	1.00	1.18	1.21	1.11	1.23	1.31			
β_1	1.00	2.38	1.92	0.97	2.45	1.91	0.21		0.38
β_2	1.00	1.56	1.62	1.27	1.91	1.84			
Setting 3: $\beta_y = (-2.5, 2, 0)$, $\beta_s = (-2.5, 3, 0, 0)$									
β_0	1.00	1.26	1.12	1.03	1.25	1.23			
β_1	1.00	1.90	1.37	0.98	1.97	1.36	0.34		0.45
β_2	1.00	1.51	1.40	1.24	1.78	1.64			
Setting 4: $\beta_y = (-2.5, 1, 0)$, $\beta_s = (-2.5, 3, 0, 1)$									
β_0	1.00	1.23	1.17	1.09	1.27	1.30			
β_1	1.00	2.05	1.64	0.97	2.12	1.64	0.21		0.37
β_2	1.00	1.42	1.48	1.27	1.80	1.74			
Setting 5: $\beta_y = (-2.5, 1, 0)$, $\beta_s = (-2.5, 3, 0, 2)$									
β_0	1.00	1.13	1.05	1.04	1.14	1.15			
β_1	1.00	2.15	1.78	0.99	2.21	1.77	0.21		0.37
β_2	1.00	1.43	1.49	1.21	1.73	1.71			
Setting 6: $\beta_y = (-2.5, 1, 0)$, $\beta_s = (-2.5, 2, 0, 0)$									
β_0	1.00	1.19	1.03	1.10	1.26	1.23			
β_1	1.00	2.08	1.65	0.99	2.09	1.64	0.21		0.33
β_2	1.00	1.30	1.42	1.31	1.70	1.75			
Setting 7: $\beta_y = (-2.5, 1, 0)$, $\beta_s = (-2.5, 1, 0, 0)$									

Designs	WTE		WAE		Sampling Proportion		
	CC1	MECC	CC2	CC1	MECC	CC2	p_{CC1} $p_{MECC} \approx p_{CC2}$
β_0	1.00	1.15	1.12	1.12	1.26	1.33	
β_1	1.00	1.41	1.58	0.99	1.39	1.56	0.29
β_2	1.00	1.20	1.52	1.38	1.65	1.91	

Table 5

The analysis of the data from a nested case-control study of risk factors for chronic kidney injury.

Parameters	Point Estimate	St. Err.	95% C.I.	p-value
<i>Loose Case Definition: Weighted NCC Analysis</i>				
<i>Intercept</i>	-3.07	0.23	-3.52, -2.62	< 0.001
<i>OSP</i>	-0.43	0.23	-0.87, 0.02	0.06
<i>Diuretic</i>	0.39	0.24	-0.07, 0.86	0.097
<i>Gender</i>	0.64	0.24	0.17, 1.11	0.007
<i>Loose Case Definition: Augmented NCC Analysis</i>				
<i>Intercept</i>	-2.87	0.20	-3.26, -2.48	< 0.001
<i>OSP</i>	-0.43	0.21	-0.84, -0.01	0.04
<i>Diuretic</i>	0.39	0.22	-0.04, 0.82	0.07
<i>Gender</i>	0.52	0.20	0.12, 0.91	0.01
<i>Strict Case Definition: MLE_{MECC} (Biased)</i>				
<i>Intercept</i>	-3.04	0.45	-3.92, -2.16	< 0.001
<i>OSP</i>	-0.47	0.41	-1.28, 0.33	0.25
<i>Diuretic</i>	0.21	0.44	-0.65, 1.08	0.63
<i>Gender</i>	0.65	0.46	-0.25, 1.56	0.16
<i>Strict Case Definition: Weighted NCC Analysis</i>				
<i>Intercept</i>	-4.80	0.45	-5.68, -3.92	< 0.001
<i>OSP</i>	-0.58	0.42	-1.40, 0.25	0.17
<i>Diuretic</i>	0.29	0.45	-0.59, 1.17	0.52
<i>Gender</i>	0.78	0.47	-0.14, 1.70	0.10
<i>Strict Case Definition: Augmented NCC Analysis</i>				
<i>Intercept</i>	-4.47	0.34	-5.14, -3.80	< 0.001
<i>OSP</i>	-0.57	0.32	-1.21, 0.06	0.08
<i>Diuretic</i>	0.28	0.35	-0.40, 0.96	0.42
<i>Gender</i>	0.75	0.36	0.04, 1.46	0.04
<i>Strict Case Definition: Weighted MECC Analysis</i>				
<i>Intercept</i>	-4.59	0.45	-5.47, -3.71	< 0.001
<i>OSP</i>	-0.55	0.42	-1.38, 0.27	0.19
<i>Diuretic</i>	0.28	0.45	-0.60, 1.16	0.54
<i>Gender</i>	0.76	0.46	-0.16, 1.68	0.10
<i>Strict Case Definition: Augmented MECC Analysis</i>				
<i>Intercept</i>	-4.49	0.33	-5.14, -3.82	< 0.001
<i>OSP</i>	-0.55	0.31	-1.17, 0.06	0.08
<i>Diuretic</i>	0.26	0.33	-0.41, 0.93	0.41
<i>Gender</i>	0.74	0.35	0.04, 1.44	0.04