eAppendix to: Reuse of Controls in Nested Case-control Studies

Nathalie C. Støer, Haakon E. Meyer and Sven Ove Samuelsen

INTRODUCTION

A frequently used study design within epidemiology is large cohort studies. In some situations additional information which is not included in the cohort, for instance biological material stored in biobanks, is required. The availability of such material is usually limited and might be very expensive to analyze for the entire cohort. A reasonable strategy is therefore to restrict the study sample to a subset of the original cohort. Two such designs are nested case-control[1,2] and case-cohort[3] studies. With a nested case-control design, *m* controls are sampled for each case. The controls are required to be alive and event free at the time the case experienced the event. Due to those requirements, we say that the controls are matched on time, or at risk status. With a case-cohort design, a subcohort is sampled at the outset of the study, and this subcohort is used as a reference population at all event times.

In many studies, more than one endpoint or type of event are of interest. Examples of this can be settings where two or more types of events ”compete”, e.g. death from cancer and death from cardiovascular diseases. Another example could be situations where one endpoint is a subset of another endpoint, e.g. incidence of prostate cancer and subsequent death from prostate cancer.

As the controls are matched to their respective cases in a nested case-control design, using them for other endpoints have traditionally not been considered possible. Thus, in the first example only the controls sampled for the endpoint in question can be used. In the second example, all controls sampled for incident cases that did not die from prostate cancer can not be included in the analysis towards death from prostate cancer. The case-cohort design, on the other hand, has often been the design of choice if it was to be carried out analyses towards multiple endpoints. Since the subcohort is a random sample from the cohort, it can be used as control population for all types of events.

In recent years, methods have been developed that allow for breaking the matching in nested case-control designs.[4–9] This opens up the possibility of reusing controls for other endpoints. Even though these methods have been around for some time, it does not seem that they have been picked up by epidemiologists yet. We will show that it is fairly easy to reuse controls also within the nested case-control design and that this in many situations can give large efficiency improvements.

Meyer et al.[10] showed in a recent paper that serum 25-hydroxyvitamin D (s-25(OH)D) was positively related to the risk of prostate cancer in a nested case-control study. We will re-analyze this study with inverse probability weighting (IPW).[4–8,11] The main endpoint in Meyer et al.[10] was incidence of prostate cancer. To investigate efficiency improvements with IPW and illustrate reuse of controls, we are also going to analyze the endpoint death from prostate cancer and subgroups of cases based on metastasis status.

DATA AND DESIGN

The cohort

The cohort consisted of participants in a collection of population based health surveys conducted in 17 of Norway’s 19 counties from 1981 to 1991. All men and women in selected birth cohorts were invited to participate, although in our context only male participants were of interest. The health surveys consisted of a health examination where a blood sample was drawn. Participants also completed a questionnaire including questions about physical activity during leisure time.

The data were linked to The Cancer Registry of Norway and to the Causes of Death Registry kept by Statistics Norway. Information on education was provided by Statistics Norway. The subjects were followed from time of health survey to incidence of prostate cancer, death or the end of December 2006. The primary endpoint was incidence of prostate cancer, and a secondary endpoint was death from prostate cancer (prostate cancer being the underlying cause of death).

The rationale for using death was that in recent years screening for prostate cancer has increased. This might increase the detection of the milder forms. By using death from prostate cancer as an endpoint instead of incidence, we may focus on the most serious cases. As an alternative approach, we divided the cases into metastatic and non-metastatic cancer, and analyzed these outcomes separately.

A primary case was defined as an individual experiencing prostate cancer and prior to that was free of all types of cancer. The individuals experiencing the secondary endpoint, death from prostate cancer, were thus a subset of the primary cases. The eFigure displays the different types of analyses that were carried out.

Nested case-control study

Our cohort consisted of 116,493 men. Among those there were 2,118 incident prostate cancer cases. A matched nested case-control design was chosen for the study.[10] For each incident case, $m = 1$ control, matched on age at serum sampling ±6 months, date of serum sampling ±2 months and county of residence, was sampled. In addition, it was required that the controls were alive and did not have any previous cancer diagnoses at the time of sampling.

We have a subsequent event setting; the cases were first diagnosed with prostate cancer, and later some cases died from this condition. However, the controls were only sampled for incident cases and when analyzing death of prostate cancer, the already sampled controls for the cases that died were used as controls in the traditional analysis.

STATISTICAL MODELS AND METHODS

For each individual we observed the pair (t_i, d_i) where d_i indicated whether subject i was a case or a control. The subjects were followed from age at health examination, l_i, to the event time age at prostate cancer or age at censoring, t_i, depending on the event indicator d_i. Age at health examination will be referred to as inclusion time. We used age as time scale instead of time on study, the subjects are thereby not followed from time zero and we have observations with delayed entries.

We considered Cox's proportional hazards model where the hazard function of events of type k was modeled as

$$h_{ki}(t_i|x_i, z_i) = h_{k0}(t_i)\exp(\beta_k' x_i + \gamma_k' z_i). \qquad (1)$$

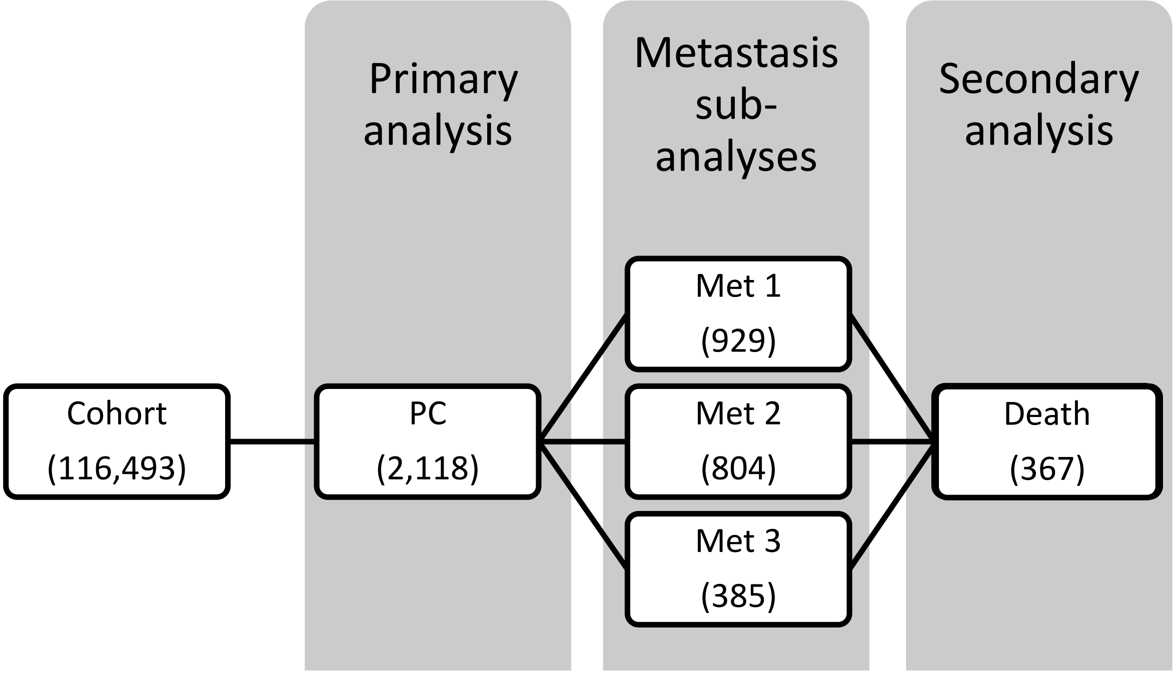
The $h_{ki}(t_i|x_i, z_i)$ is the general model for all types of events; when $k = 1$, $h_{1i}(t_i|x_i, z_i)$ is the model for incidence of prostate cancer. With $k = 2$, $h_{2i}(t_i|x_i, z_i)$ models death from prostate cancer whereas $k = 3, 4, 5$ will represent three metastasis groups; non-metastatic, metastatic and unknown cancer status. The β_k are the log-hazard ratios connected to the k-th endpoint. Further, $h_{k0}(t_i)$ is the k-th baseline hazard, the hazard when all covariates are equal to zero. The covariates x_i include the main exposure, s-25(OH)D together with the confounder education. The z_i are covariates used as matching variables with corresponding regression parameters γ_k. When the cases and controls are matched in such a way that they have the exact same value of the matching variables, as described in the next section, the γ_k's will cancel out in the traditional analysis. However, if the matching is broken, as with IPW, the case is compared to all subjects at risk and the matching variables should be adjusted for.

The traditional nested case-control design with additional matching

The traditional way of analyzing data from a nested case-control design is a with a partial likelihood[1,12]

$$L(\beta_k) = \prod \frac{\exp(\beta_k' x_j)}{\sum_{i \in R_j} \exp(\beta_k' x_i)}. \qquad (2)$$

The product is over all cases of type k and indexed by j. The R_j is the sampled risk set at t_j, which constitutes the case at t_j together with its sampled controls. The estimation can be carried out by a stratified Cox-regression, where the stratification is with respect to risk sets. In a stratified Cox-regression the case is only compared to the controls within the same strata. Additionally matched data can therefore be analyzed in exactly the same manner as data with controls only matched on time. Note that under the assumption of equally or very closely matched cases and controls, the terms $\gamma_k' z_i$ in equation (1) are equal within the sampled risk sets and cancel out in equation (2). Separate analyses are carried out for each endpoint, hence one analysis for incidence, three analyses for metastasis status and one analysis for death from prostate cancer.



eFIGURE. Overview of analyses. Each box, except the first, represents the cases in each type of analysis. The numbers in parenthesis are the number of cases in each box. PC - prostate cancer, Met1 - local cancer, Met 2 - unknown status, Met 3 - advanced cancer

Inverse probability weighting

In the recent decade or so methods have been developed that allow for breaking the matching in nested case-control designs.[4–9] The methods can be divided into two groups; inverse probability weighting of the partial likelihood[4,5,7,8] and full likelihood methods.[6,9] We will only consider the former.

When we want to use our controls for a different endpoint and thereby break the matching, a naive approach would be to analyze our nested case-control data as if it was a cohort. However, we would then ignore the fact that the data is a biased sample from the cohort. The sample is biased with respect to the proportion of cases and controls. In our study we have 2,118 cases and the total size of the cohort is 116,493, hence the cases make up less than 2% of the cohort, while with the nested case-control design the cases make up 50% by design. In addition, the controls will generally be a biased sample with respect to the additional matching variables, and possibly the covariates.

The naive approach may give erroneous results because of the biased sample. However, inverse probability weighting can resolve this problem. The general idea of IPW is to reconstruct the full cohort by letting each sampled subject represent a number of subjects in the full cohort. Generally, the subjects that are at risk only for a short period of time, or are dissimilar to most cases on the matching variables, are less likely to be sampled as controls although they might comprise a considerable part of the cohort. To obtain a sample that is similar to the cohort, these subjects will receive a large weight since they are under-represented in the case-control sample. The subjects that are more likely to be sampled will correspondingly receive smaller weights. This is achieved by weighting with the inverse sampling probability. The cases are given weight 1, since usually all of them are included, hence sampled with probability 1.

The estimation is then carried out by a weighted Cox-regression

$$L_k(\beta_k,\gamma_k)=\prod\frac{\exp(\beta_k' x_j+\gamma_k' z_j)}{\sum_{i\in S_j}\exp(\beta_k' x_i+\gamma_k' z_i)w_i}, \qquad (3)$$

where the product is over all cases of type k. This likelihood is similar to the likelihood in equation (2). The main differences are the weights w_i, and that the sum is now over a set S_j, defined as all cases and all controls at risk at t_j. The non-weighted version of (3) corresponds to the naive approach. In a situation with no additional matching, there are no z_i's and the $\gamma_k' z_j$ part is omitted from (3) both in the nominator and in the denominator. With additional matching, the z_i may vary over S_j so $\gamma_k' z_i$ will generally not cancel.

The weights/sampling probabilities depend on the inclusion time, censoring time and matching variables. The sampling probabilities increase for increasing censoring time and decrease for increasing inclusion time; the longer a potential control has been included in the study the more opportunities it has had to be selected as a control.

The sampling probabilities $p_i = 1/w_i$, have to be estimated from the data and we will consider three estimation methods, first without taking the matching into account. Samuelsen[8] suggested an estimator with a similar form as the Kaplan-Meier (KM) estimator

$$p_i = 1 - \prod_{l_i < t_j < t_i} \left\{ 1 - \frac{m(t_j)}{n(t_j) - 1} \right\}, \qquad (4)$$

where $n(t_j)$ is the number at risk at the event time t_j. See also Suissa et al.[13] The expression in brackets is the probability for individual i of not being sampled as a control for the case experiencing the event at time t_j. Such probabilities are then multiplied each time individual i could have been sampled and 1 minus this product is the probability for individual i of ever being sampled as a control. It is only necessary to calculate the probabilities for controls, since non-sampled individuals are not included in the Cox-regression. We will refer to weights estimated by (4) as KM-weights. Note that for generality we use time varying number of controls $m(t_j)$ in (4) which means that cases can have different number of controls. However, in our situation $m(t_j) = m = 1$.

A more model based estimation procedure is to use the fitted values from a logistic regression model where a sampling indicator is the outcome and the variables that the sampling probabilities depend on are covariates.[6,11,14] Without additional matching, this will be the censoring time, and the inclusion time if the time scale is different from time on study. In our study, the covariates in the logistic regression would be age at health examination, which is the inclusion time, and age at censoring. Note that the estimation is carried out on the cohort with the cases excluded, since by design the cases are sampled with probability 1.

A third option for estimating the weights is by using a generalized additive model[15] (GAM) with the censoring time and inclusion time as covariates and the sampling indicator as outcome. This model is more flexible than the ordinary logistic regression model, however in our experience there is usually little difference between logistic regression and GAM, with respect to final hazard ratios and standard errors.

Since our subjects are used a number of times in the estimation, the variance estimation is somewhat more involved with IPW than it is with the traditional estimator. Robust variances[16] are often a good alternative, although somewhat conservative in some situations.[14] There exists explicit variance formulas for the Kaplan-Meier type of weights in situations both with and without additional matching,[7,14] we have however used the robust variances in the following analyses.

Additional matching

To adjust for confounding and to increase efficiency, the controls in a nested case-control design are often matched on additional factors. We will distinguish between two "types" of matching criteria; caliper matching and category matching. We make the distinction because the sampling probabilities will not be estimated in exactly the same manner with the two criteria. With caliper matching, the value of the matching variable for a potential control must fall within an interval around the value of the matching variable for the case. With category matching, the value of the matching variable must match the value of the case exactly. In our study we have two caliper matching criteria; age at serum sampling ±6 months and date of serum sampling ±2 months, and one category matching criterion; county of residence.

The sampling probabilities will be affected when the controls are matched on additional factors and the estimation procedures may need to take this into account. Salim et al.[7] and Cai and Zheng[4] suggested a generalization of the Kaplan-Meier type of weights with caliper matching, which can be expressed as

$$p_i = 1 - \prod_j \left\{ 1 - \frac{m(t_j)}{n_j(t_j) - 1} I(\text{Control } i \text{ could be sampled for case } j) \right\}. \qquad (5)$$

The $n_j(t_j)$ is the number at risk at time t_j who meet the matching criteria of case j. The $I(\cdot)$

is an indicator function, hence the product is only over cases that subject i could have been sampled as a control for. The only differences between (5) and (4) are that the product is now only over subjects that meet the matching criteria, and similarly the denominator only consists of the number of subjects at risk that meet the matching criteria.

If the matching criterion is narrow, hence only a few or perhaps only one individual can be matched to each case, the sampling probabilities will be close to, or equal to 1. In such situations, the idea of ”reconstructing” the cohort by giving larger weight to the controls than to the cases, will collapse. When the ”true” sampling probabilities are close to 1, a more model based approach with logistic regression or GAM, where the matching variables, in addition to inclusion time and event time are used as covariates, might work better. Such an idea has also been mentioned by Saarela et al.[6]

There are some difficulties by using logistic regression or GAM regarding how the matching variables should be included in the regression. With category matching and not too many categories, a natural approach is to include the matching variables as categorical covariates. With many categories, collapsing small groups into bigger groups or including the matching variables as continuous covariates are reasonable. With caliper matching there may be several ways to adjust for the matching variables. In our analysis we have used age at diagnosis as a continuous covariate while date of serum sampling is categorized into month of serum sampling and used as a categorical covariate. We only included month of serum sampling since the amount of sun exposure is assumed to have a yearly variation. Since there is not any standardized recipe of how to adjust the weights for the matching variables, it is important to compare the hazard ratio from IPW with the hazard ratio from the traditional estimator.

One reason for matching is to adjust for confounding, however after we have broken the matching the cases are no longer only compared to their matched controls, but to all cases and controls at risk at the given time. This means that the confounding is no longer ”matched away” and the matching variables should be adjusted for in the Cox-regression. The need for such adjustment can also be seen as an advantage of IPW since the necessity of the particular matching can be evaluated. This is impossible for category matching with the traditional estimator since all subjects within each matched set will have the same value of the matching variables.

An alternative to adjusting for the matching variables is to stratify on them. In effect this is what is done with the traditional estimator. However, from simulations (not presented), adjusting for matching variables seems to be somewhat more efficient than stratification.

RESULTS

The results from the IPW analyses and the traditional analyses can be found in eTable 1-3. All hazard ratios correspond to a 30 nmol/l increase in s-25(OH)D concentrations and all IPW analyses are adjusted for the matching variables.

The risk of incident prostate cancer increased with increasing s-25(OH)D concentrations (eTable 1, left side) and adjusting for education did not alter this (not shown). The HR from IPW was similar to the HR from the traditional estimator while the standard errors from IPW were somewhat smaller than the standard errors from the traditional estimator, resulting in larger efficiencies for IPW.

Efficiency is calculated as the variance from the traditional estimator divided by the variances from the IPW estimators. When the efficiency is larger than 1, the traditional estimator needs more case-control pairs to obtain the same accuracy as IPW. The efficiency with logistic regression weights was 1.38, hence the traditional estimator requires 38% more case-control pairs to obtain the same level of accuracy as IPW.

The right side of eTable 1 displays the results from the secondary analysis with death from prostate cancer as endpoint. There was no association between s-25(OH)D and death from prostate cancer and adjusting for education did not change this (not shown). The HRs from IPW were similar to the traditional HR. The standard errors were considerably smaller with IPW, as all available controls could be used towards the subset of cases, thereby increasing efficiency to about 2.

eTable 2 displays the result of the analyses with localized cancer, advanced cancer and unknown metastasis status as endpoints. Out of the 2,118 cases, 929 had localized cancer, 385 had advanced cancer and 804 had unknown metastasis status. The hazard ratios from IPW for localized cancer and

eTABLE 1. Comparison of the traditional estimation method and IPW. Incidence of prostate cancer and death from prostate cancer.

Method	HR[a]	se(β)[b]	95% CI	Eff.[c]	HR[a]	se(β)[b]	95% CI	Eff.[c]
	Incidence				Death			
Trad.[d]	1.17	0.05	1.06 - 1.29	1	0.98	0.12	0.77 - 1.25	1
KM[e]	1.12	0.05	1.03 - 1.23	1.15	1.09	0.09	0.91 - 1.29	1.93
GAM[f]	1.13	0.04	1.04 - 1.22	1.37	1.06	0.09	0.89 - 1.27	1.95
GLM[f]	1.11	0.04	1.02 - 1.21	1.38	1.06	0.09	0.89 - 1.27	2.00

[a] HR correspond to a 30 nmol/l increase in s-25(OH)D [b] Standard error of log hazard ratio
[c] Variance from traditional estimator divided by variance from IPW
[d] The traditional estimator, equation(2) [e] IPW with Kaplan-Meier type of weights
[f] GAM/GLM - IPW with weights estimated with GAM/logistic regression

unknown cancer status are reasonably close to the traditional estimator, whereas the differences are somewhat larger for advanced cancer.

With IPW, we could use all sampled controls when analyzing all three endpoints, e.g. for advanced cancer all 2,118 controls were used as controls for the 385 cases. In contrast, the traditional estimator could only use the 385 controls sampled for the advanced cases. The extra controls available for IPW resulted in larger efficiencies, being close to or above 2 for all cancer types. Hence, the traditional analyses would need twice as many case-control pairs to obtain the same level of accuracy as IPW.

In the previous analyses there were some discrepancies between the traditional estimator and the IPW estimators. A possible partial explanation for this is that while the traditional estimator only uses the case-control pairs in the estimation, IPW methods make use of all available controls. Thus in the metastasis sub analyses and analysis of death from prostate cancer, the IPW estimation is based on a number of subjects not included in the traditional estimation and somewhat different results are not unexpected.

Since our cohort is a collection of health surveys, a number of variables are known for all subjects, for instance physical activity (PA). We could then investigate the discrepancy between the estimators by analyzing the association between PA and incidence of prostate cancer with the traditional estimator and IPW, and compare the results to a complete cohort analysis. The cohort estimates were fitted with Cox-regression on the entire cohort and adjusted for the matching variables in the same way as IPW. The original PA variable has four levels, however we recoded it into three categories where 1 corresponds to sedentary activity, 2 to moderate activity at least four hours a week and 3 corresponds to regularly intermediate to intensive physical activity.

eTABLE 2. Comparison of the traditional estimation method and IPW. Metastasis sub-analysis.

Method	HR[a]	se(β)[b]	95% CI	Eff.[c]
		Localized cancer		
Trad.[d]	1.08	0.07	0.94 - 1.25	1
KM[e]	1.13	0.06	1.01 - 1.26	1.71
GAM[f]	1.14	0.05	1.02 - 1.26	1.81
GLM[f]	1.12	0.05	1.01 - 1.25	1.83
		Advanced cancer		
Trad.[d]	1.23	0.12	0.98 - 1.54	1
KM[e]	1.08	0.07	0.92 - 1.25	2.18
GAM[f]	1.07	0.08	0.92 - 1.25	2.12
GLM[f]	1.05	0.08	0.90 - 1.23	2.14
		Unknown cancer status		
Trad.[d]	1.23	0.08	1.05 - 1.45	1
KM[e]	1.17	0.06	1.05 - 1.32	1.96
GAM[f]	1.17	0.06	1.05 - 1.31	2.08
GLM[f]	1.15	0.06	1.03 - 1.29	2.09

[a] HR correspond to a 30 nmol/l increase in s-25(OH)D
[b] Standard error of log hazard ratio
[c] Variance from traditional estimator divided by variance from IPW
[d] The traditional estimator, equation(2)
[e] IPW with Kaplan-Meier type of weights
[f] GAM/GLM - IPW with weights estimated with GAM/logistic regression

eTABLE 3. The traditional estimator and IPW compared to the cohortanalysis. Estimates of physical activity on incidence of prostate cancer; HR (95% CI).

Method	PA 1[a]	PA 2[a]	PA 3[a]
Cohort	1	1.07 (0.95 - 1.22)	1.18 (1.04 - 1.35)
Trad.[b]	1	1.09 (0.92 - 1.29)	1.31 (1.08 - 1.58)
KM[c]	1	1.01 (0.85 - 1.21)	1.15 (0.95 - 1.39)
GAM[d]	1	1.08 (0.91 - 1.28)	1.23 (1.02 - 1.48)
GLM[d]	1	1.07 (0.90 - 1.26)	1.23 (1.02 - 1.48)

[a] PA - Physical activity, PA 1: sedentary activity,
PA 2: moderate activity
PA 3: intermediate to intensive activity
[b] The traditional estimator, equation (2)
[c] IPW with Kaplan-Meier type of weights
[d] IPW with weights estimated with GAM/logistic regression

eTable 3 displays the results from the analyses. The reference group is PA 1. For PA 2, all point estimates were close to the cohort estimate. However, the KM estimate is somewhat smaller than the other estimates. For PA 3, the estimates from IPW were closer to the cohort than the traditional estimator. The KM estimate was closest to the cohort estimate, however also here smaller than the other estimates.

By applying IPW to a study of serum 25-hydroxyvitamin D and prostate cancer, we have shown that inverse probability weighting can increase the efficiency of NCC-designs quite substantially in some situations. The IPW analyses require an additional estimation step, which is estimation of weights. However, this can be carried out fairly easily e.g. with logistic regression.

REFERENCES

1. Thomas DC. Addendum to ”Methods of cohort analysis: Appraisal by application to asbestos mining” by Liddell FDK, McDonald JC and Thomas DC. *J Roy Stat Soc Ser A*. 1977; 140:469–491.
2. Langholz B, Goldstein L. Risk set sampling in epidemiologic cohort studies. *Stat Sci*. 1996; 11:35–53.
3. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
4. Cai T, Zheng Y. Evaluating prognostic accuray of biomarkers in nested case-control studies. *Biostatistics*. 2012; 13:89–100.
5. Chen KN. Generalized case-cohort sampling. *J Roy Statist Soc Ser B*. 2001; 63:791–809.
6. Saarela O, Kulathinal S, Arjas E, Läärä E. Nested case-control data utilized for multiple outcomes: A likelihood approach and alternatives. *Stat Med*. 2008; 27:5991–6008.
7. Salim A, Hultman C, Sparén P, Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. *Biostatistics*. 2009; 10:70–79.
8. Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*. 1997; 84:379–394.
9. Scheike TH, Juul A. Maximum likelihood estimation for Cox’s regression model under nested case-control sampling. *Biostatistics*. 2004; 5:193–206.
10. Meyer HE, Robsahm TE, Bjørge T,Brustad M, Blomhoff R. Vitamin D, season and the risk of prostate cancer. A nested case-control study within Norwegian health studies. *Am J Clin Nutr*. 2013; 97:147-154.
11. Støer NC, Samuelsen SO. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Anal*. 2012; 18:261–283.
12. Borgan Ø, Goldstein L, Langholz B. Methods for the analysis of samled cohort data in the Cox proportional hazards model. *Ann Stat*. 1995; 23:1749–1778.

13. Suissa S, Edwardes MDD, Boivin JF. External comparisons from nested case-control designs. *Epidemiol*. 1998; 9:72–78.

14. Samuelsen SO, Ånestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. *Scand J Stat*. 2007; 34:103–119.

15. Hastie LJ, Tibshirani RJ. Additive models, trees and related methods. In: *The elements of statistical learning*. 2nd ed. London: Chapman & Hall; 2009:295–336.

16. Barlow WE. Robust variance-estimation for the case-cohort design. *Biometrics*. 1994; 50:1064–1072.