Chapter 6

Study Designs for Biobank-Based Epidemiologic Research on Chronic Diseases

Esa Läärä

Abstract

A review is given on design options to be considered in epidemiologic studies on cancers or other chronic diseases in relation to risk factors, the measurement of which is based on stored specimens in large biobanks. The two major choices for valid and cost-efficient sampling of risk factor data from large biobank cohorts are provided by the *nested case–control* design, and the *case–cohort* design. The main features of both designs are outlined and their relative merits are compared. Special issues such as matching, stratification, and statistical analysis are also briefly discussed. It is concluded that the nested case–control design is better suited for studies involving biomarkers that can be influenced by analytic batch, long-term storage, and freeze-thaw cycles. The case–cohort design is useful, especially when several outcomes are of interest, given that the measurements on stored materials remain sufficiently stable during the study.

Key words: Nested case–control , Case–cohort, Matching, Stratification, Statistical analysis, Risk factors

1. Introduction

Epidemiologic studies of chronic diseases require large study populations and skillful planning on various aspects of study design, selection of the study subjects, measurements of the values of interesting risk factors and other variables, organization of the follow-up for identification of the study outcomes, and analysis of the results. Careful planning is even more demanding, when measurements are based on stored biological materials, such as tissue or blood specimens, considering the labor and costs associated with them.

In this paper, a review is presented on the choices of epidemiologic study designs to be considered in this kind of investigations. Our special focus is on the *nested case–control* (NCC)

Joakim Dillner (ed.), *Methods in Biobanking*, Methods in Molecular Biology, vol. 675,
DOI 10.1007/978-1-59745-423-0_6, © Springer Science+Business Media, LLC 2011

design and the *case–cohort* (CC) design. More detailed accounts on the various designs are given in many excellent textbooks, such as those of dos Santos Silva (1) and Rothman et al. (2). Important aspects of the two major designs from a more statistical perspective are concisely and quite untechnically treated, e.g., by Borgan and Samuelsen (3). Vineis et al. (4) provide an extensive discussion on the relative merits of the NCC and the CC designs with special reference to biobank-based studies, and they offer thoughtful guidelines for choosing between them.

As a concrete introduction to the theme, two representative examples of modern biobank-based epidemiologic research are briefly summarized.

Example 1. "Activation of maternal Epstein-Barr virus infection and risk of acute leukemia in the offspring" (5). The study population comprised a joint cohort of ca. 550,000 offspring, their mothers being identified from the Icelandic and the Finnish biobanks covering pregnant women. Serum samples were routinely taken from all these women in the first trimester of pregnancy, from 1975 to 1983 onwards in the two countries, respecively. Follow-up of the offspring began at birth and lasted until 1997. In the total of 7 million person-years of follow-up, 304 cases of acute lymphatic leukemia (ALL) and 39 cases of other leukemias (non-ALL) occurring in the offspring by 15 years of age were identified from the national cancer registries. Three or four control subjects for each case were sampled from the original cohorts by incidence density sampling. The control subjects were matched with the case on biobank/country, maternal age at serum sampling (±2 years), date of specimen collection (±2 months), as well as on gender, and date of birth (±2 months) of the offspring. The frozen sera from mothers of these cases and from 943 mothers of the control subjects were analyzed for antibodies to viral capsid antigen (VCA), early antigen, and EBV transactivator protein ZEBRA. One major result was that "EBV VCA IgM antibodies were associated with a statistically significant relative risk of childhood ALL (odds ratio = 1.9, 95% confidence interval: 1.2, 3.0)."

Example 2. "Risk alleles of *USF1*-gene predict cardiovascular disease" (6). The study population comprised two FINRISK cohorts in Finland, in total ca. 14,000 males and females, of initially 25–64 years of age. The cohorts were recruited in 1992 and 1997, respectively. The baseline measurements comprised a health examination and a structured questionnaire, and blood specimens were also taken at entry. A subcohort of 786 subjects was randomly sampled from the cohorts. The cohorts were followed-up from entry to 31 Dec 2001 and 31 Dec 2003, respectively. In the 112,000 years of total follow-up, 528 new cases of cardiovascular diseases (CVD) were identified in the cohorts, of which 72 were in the subcohort. The frozen blood specimens pertaining to the cases and the subchort members were genotyped. One of the main results was that "female carriers of a *USF1* risk haplotype had a twofold risk of a CVD event (hazard ratio (HR) 2.02; 95% confidence interval (CI) 1.16–3.53), after adjustment for conventional risk factors."

2. Validity and Efficiency of an Epidemiologic Study

An epidemiologic study is a measurement exercise (2). The object of measurement is some *parameter* of interest, such as the *hazard rate ratio* (HR or "relative risk") of a major coronary event between individuals with a high-risk and a low-risk haplotype, respectively. The result of this exercise is an *estimate* of the parameter, which is an empirical measure to be computed from the available data. Estimates of the HR include the *incidence rate ratio* (IR or *incidence density ratio*) obtained from a cohort study, or the *exposure odds ratio* (EOR) from a case–control study.

The estimation of a parameter is prone to error; we can express an estimate as a sum of three components:

$$\text{Estimate} = \text{true parameter value} + \text{bias} + \text{random error}.$$

Common sources of *bias* or *systematic error* include (a) confounding or non-comparability of the exposure groups, (b) measurement error and misclassification, (c) non-response, losses to follow-up, or otherwise incomplete data, and (d) sampling and selection of subjects to the study and to be measured. An educating presentation on various biases is given by Maclure and Schneeweiss (7). The main sources of *random error* are in turn (a) biological variation between and within individuals, (b) measurement variation, (c) sampling (whether random or non-random), and (d) division of exposure (whether properly randomized or non-randomized).

An epidemiologic study is said to be *valid*, when its design and methods would provide an *unbiased* estimate of the parameter (such as HR) of interest. Unbiased estimation means that the estimate (like IR or EOR) would equal the true parameter value (HR) if the study had no random error. For example, if the true HR on CVD events for high- vs. low-risk haplotype carriers was 2.5, this value would be exactly obtained by our estimate IR if we had unlimited amount of data and if our designs were valid. (NB. By exceptional luck, we could get an IR of 2.5 also with typical amount of data even with a biased design!)

The *precision* of an estimate means smallness of random error in estimation. Random error is measured by the variance or *standard error* (SE) of the estimate, or by the *confidence interval* (CI) of the parameter. The *efficiency* of a design means its ability to provide a precise estimate with given data. We say that design A is more efficient than design B if either (1) with the same amount of data, the estimate from A has a smaller random error than that from B, or (2) smaller amount of data is needed by design A to obtain the same precision as that obtained by B.

3. Cohort Studies

An outline of a typical cohort study or a full cohort design is as follows:

1. Subjects fulfilling the eligibility criteria are selected to the study cohort.
2. Risk factors of interest as well as relevant confounders and effect modifiers are measured in all cohort members.
3. New incident cases of outcome (e.g., cancer) are identified during the follow-up from the time of entry to until the time of exit from the follow-up.
4. Incidence rates = cases/person–time in the exposure groups, and the ratios (IRs) between them are computed.
5. Confounding and modification are controlled by stratification and Mantel–Haenszel methods, or nowadays more commonly by regression modeling: the Poisson regression, or the proportional hazards (Cox) model.

In both examples presented in the introduction, a full cohort design would imply that serologic assays for the EBV antibodies would have been performed on the sera of all the 550,000 mothers in Iceland and Finland, as well as genotyping for the USF1-gene would have been conducted for all the 14,000 members of the two FINRISK cohorts.

The principle of estimating the HRs of interest from a full cohort design is illustrated in the simplest possible setting: one single dichotomous risk factor. From the figures given in Table 1,

Table 1
Crude summary of follow-up results in a cohort study addressing the effect of a dichotomous risk factor ("exposed" vs. "unexposed") on the hazard of getting a given disease

	Exposed	Unexposed	Total
New cases	D_1	D_0	D
Person–time	Y_1	Y_0	Y
Incidence rate	$I_1 = D_1/Y_1$	$I_0 = D_0/Y_0$	$I = D/Y$

the target parameter, HR, is estimated by the ratio of the empirical incidence rates I_1 and I_0 in the two exposure groups.

$$\mathrm{IR} = \frac{I_1}{I_0} = \frac{D_1/\Upsilon_1}{D_0/\Upsilon_0} = \frac{D_1/D_0}{\Upsilon_1/\Upsilon_0}.$$

This crude estimation ignores the possible confounding caused by other risk factors of the outcome disease, but provides a convenient starting point to illustrate the precision and efficiency of different designs.

The precision in the estimation of the HR depends inversely on the numbers of cases. The estimated variance of the logarithm of the crude IR is, namely, expressed as

$$V = \frac{1}{D_1} + \frac{1}{D_0} = \frac{1}{\text{no. exposed cases}} + \frac{1}{\text{no. unexposed cases}}.$$

From this, we obtain the common approximate confidence limits for the hazard ratio:

$$\mathrm{IR} \times \exp(\pm 1.96 \times \sqrt{V}).$$

Note that the variance does not depend on the sizes of the exposure groups (or their person–times) as such, even if these were millions. However, for rare diseases with low rates, large cohorts are needed to obtain enough cases for adequate precision.

Collection and processing of data on exposure variables, confounders, and modifiers are very slow and expensive in large cohorts. It is relatively easy and cheap with data obtained by questionnaires or from readily available registers. However, it would be extremely costly and laborious for, e.g., measurements from biological specimens (like genotyping, antibody assays, etc.), dietary diaries, and occupational exposure histories in manual records. In our two example studies, the full cohort design would obviously be an imaginary possibility only.

Thus, a question arises whether we are able to obtain equally valid estimates of the interesting HRs with nearly as good precision as those obtained by some other, less costly strategies. The answer is "yes," and we shall justify this by first inspecting more closely the estimation of hazard ratios:

The crude IR in a cohort study can be expressed by

$$\begin{aligned}\mathrm{IR} = \frac{D_1/D_0}{\Upsilon_1/\Upsilon_0} &= \frac{\text{cases: exposed / unexposed}}{\text{person} - \text{times: exposed / unexposed}} \\ &= \frac{\text{exposure odds in cases}}{\text{exposure odds in person} - \text{times}} \\ &= \text{ exposure odds ratio (EOR)}\end{aligned}$$

In practical terms, this estimator relates the exposure distribution observed in the cases vs. the exposure distribution prevailing in the whole cohort. A suggestion is thus given for the search of more efficient designs:

1. To obtain information on the *numerators* of the incidence rates in the two exposure groups, one should aim at collecting exposure data on all possible cases of the outcome disease.
2. As to the *denominators* of the rates, one may estimate with high precision the division of person–times Y_1/Y_0 into the exposure groups by appropriate sampling of referent or "control" subjects, on whom exposure data will be measured and collected, from the members of the whole cohort at risk. This idea leads us to the case–control designs.

4. Case–control Studies

The general principle in the so-called case–control or case–base, or case–referent designs is the following: The selection of study subjects from a given study population is stratified by the outcome (disease) under study.

The study population comprises subjects who *would be* included as cases *if they got* the outcome disease during the study. Hence, this population may also be called as the source population of the cases (2).

In cohort-based case–control studies, the study population is a well-defined closed population, the membership being fixed by entry to the cohort and lasting forever. These kinds of case–control studies are the focus of this article, and the so-called hospital-based and register-based case–control studies are left aside (1).

In all types of case–control studies, the data on interesting risk factors are collected separately from

1. The *case group*, comprising all (or a high proportion of) the D subjects in the study population (total N subjects) encountering the outcome disease during follow-up
2. The referent or *control group*, which is a *random sample* of C subjects from the whole population (C much smaller than N), such that the eligible controls must be *at risk,* i.e., alive, under follow-up and still free from the outcome at specified time points

Depending on how these time points are actually specified, different sampling schemes or designs for the selection of control subjects are obtained. The major sampling schemes or designs are the following:

(a) *Traditional* design ("case–noncase" sampling): Controls are chosen from these $N - D$ cohort members who are still at risk (healthy) *at the end* of the follow-up. We do not consider this design any further, which is typically used in studies of acute diseases (outbreaks). It also presupposes complete follow-up (no losses) of the cohort over a fixed-length risk period, which is rarely realized with chronic diseases.

(b) *Incidence density sampling* (or concurrent sampling) design: Controls are drawn at different times *t during the follow-up* from these N_t subjects at risk. An important special case is the *nested case–control* design (NCC), in which a set of controls is sampled in a *time-matched* manner from the *risk set at each time t of diagnosis* of a new case.

(c) *Case–cohort* design (CC): The control group – *subcohort* – is a random sample of the whole cohort (N) *at the beginning of the follow-up*.

It is worth mentioning that the term "nested case–control studies" has variable meanings. In biostatistical literature (3), it commonly refers to the most popular variant of density sampling, in which *time-matching* or *risk-set sampling* is employed: At each time t when a new case is found, a set of controls is sampled from the N_t members of the study population belonging to the *risk-set* at time t (see above). This is illustrated in Fig. 1. However, in some epidemiologic texts (1), the "nested case–control design" refers to any kind of control sampling when a study population is a well-defined cohort, covering thus also the traditional sampling as well as the case–cohort design. Here, the word "nested case–control design" is used in the first meaning, i.e., referring to the time-matched sampling of controls from risk sets (3).

Note that in this design, a control chosen at a time of some previous case can later on become a case, too.

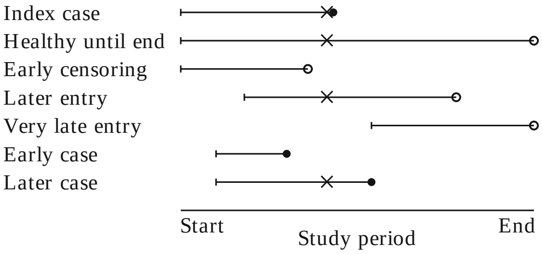

Fig. 1. Time-matched sampling from risk sets. Follow-up lines of seven subjects run vertically at different levels, and they may end either by the outcome event (*filled circle*) or censoring (*open circle*) due to deaths from other causes or emigration. The risk set from which controls are sampled at the time of diagnosis of the index case comprises subjects (marked by ×) who are alive, free from the outcome, and under follow-up at that time.

In order to guarantee a valid sampling frame for control selection from the relevant risk set at any time, it is very important to maintain accurate and complete follow-up also with respect to dates of deaths and emigrations occurring in the cohort, apart from the outcome events.

Example 1 in the introduction is a typical NCC study. Time-matched sampling of controls from the risk sets was employed, although not explicitly described in this paper that for each case, the chosen controls were alive, not censored, and free from leukemia at the date of diagnosis. Close time-matching was actually performed on the age scale, too, because the date of birth of each control was less than 2 months apart from that of the case. In addition to time and age, the selection of controls was matched on various other factors, too (more on this in Sect. 6).

Example 2 in the introduction is clearly a CC study. The subcohort, a random sample of 786 subjects from the whole cohort, selected at the outset, served as the control group for all subsequent cases. In this design, a subcohort member can become a case, too, as actually happened to 72 subjects.

The nested case–control variant of the density sampling design (b) is the most popular one in chronic disease epidemiology. The case–cohort design (c) is newer, but is gradually gaining in popularity. It is particularly recommended when several outcomes are of interest, and measurements of risk factors from any stored material are relatively stable.

5. Estimation, Precision, and Efficiency

Results from a case–control study are often summarized as in Table 2. From these four counts, the crude exposure odds ratio is computed:

$$\mathrm{EOR} = \frac{D_1 / D_0}{C_1 / C_0} = \frac{\text{cases: exposed / unexposed}}{\text{controls: exposed / unexposed}}.$$

A common but false doctrine, unfortunately still found in many elementary textbooks in epidemiology, is that the only parameter

Table 2
Crude summary of results in a case–control study with a dichotomous risk factor

	Exposed	Unexposed	Total
No. of cases	D_1	D_0	D
No. of controls	C_1	C_0	C

estimable from a case–control study is the odds ratio, meaning specifically the *risk odds ratio* (ROR)

$$\mathrm{ROR} = \frac{\text{odds of disease in the exposed}}{\text{odds of disease in the unexposed}}$$
$$= \frac{R_1 / (1 - R_1)}{R_0 / (1 - R_0)},$$

where R_1 and R_0 are the *risks* of disease over a fixed risk period in the two exposure groups. This holds indeed in the traditional "case–noncase" design. When the disease is "rare," the ROR is closely approximating the corresponding *risk ratio* RR = R_1/R_0 as well as the HR.

However, in case–control studies based on density sampling or case–cohort sampling, one can estimate directly the HR without any rare disease assumption. For the density sampling, the argument is simplified as follows (2): It can be shown that given certain assumptions, the exposure odds C_1/C_0 among the controls provide a statistically consistent estimate of the odds Y_1/Y_0 of person–times between the exposure groups in the whole cohort from which the cases and controls are sampled. Hence, EOR between cases and controls actually is a valid and efficient estimate of the unknown HR, which is the target of our interest.

In the case–cohort design, the principle is the same but the estimation of the hazard ratio is more complicated. Nevertheless, the argument above illustrates the true role of the controls: They are NOT representing the population of "non-cases," i.e., those who would remain healthy; instead, they are providing data on the distribution of exposures in the whole cohort.

As an aside, another common but misleading textbook wisdom says that absolute levels of incidence rates or risks cannot be estimated from a case–control study. This statement holds only for studies based on an ill-defined source population of cases, such as hospital-based case–control studies in USA. Suppose, however, that (1) a well-defined cohort is followed up for Y total person-years, (2) $D = D_1 + D_0$ cases plus $C = C_1 + C_0$ controls are drawn from it, and (3) their exposure assessed. In these circumstances, the person-years and the crude absolute incidence rates in the two exposure groups $k = 0$, 1 would be estimated in a straightforward way:

$$\tilde{Y}_k = \frac{C_k}{C} \times Y, \quad \tilde{I}_k = \frac{D_k}{\tilde{Y}_k}.$$

These crude computations are, however, not useful in real-life studies with variable follow-up times over a wide age range. More refined methods for absolute risk estimation are available, though, as presented by Langholz and Borgan (8).

Consider next the precision and efficiency of the estimation of "relative risk" in case–control studies. In density sampling, or the NCC design, the estimated variance of the logarithm of the crude exposure odds ratio may be expressed as

$$\begin{aligned} V_{\mathrm{NCC}} &= \frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0} \\ &= \text{cohort variance} + \text{sampling variance}. \end{aligned}$$

The variance depends thus basically on the numbers of exposed and unexposed cases, whenever the numbers of controls C_1 and C_0 are clearly bigger than the numbers of cases. Hence, the variance is not much bigger than that in a full cohort study with the same number of cases. Usually, the gain to be obtained with more than four or five controls per case is marginal. This shows that the case–control design is very cost-efficient!

Some results from Example 1 are summarized in Table 3. Ignoring matching for the sake of illustration only, the crude estimate of the HR between the antibody positives and the antibody negatives is

$$\mathrm{EOR} = \frac{30/274}{47/815} = 1.9$$

Even though one should not be content with reporting a crude estimate when really analyzing matched data, we note that this value happens to be numerically the same as the HR estimate (or "odds ratio," as the authors called it) reported in the original article, which was adjusted for matching factors and for some other covariates by conditional logistic regression model (see Sect. 7).

The estimated variance of log(EOR) is

$$V = \left(\frac{1}{30} + \frac{1}{274}\right) + \left(\frac{1}{47} + \frac{1}{815}\right) = 0.0370 + 0.0225 = 0.0595,$$

and the 95% confidence interval ranges from 1.2 to 3.1, these crude limits being again close to the reported ones. Thus, the variance

Table 3
Maternal IgM antibodies to the EBV VCA and the acute lymphatic leukemia (ALL) in the offspring. Numbers of antibody positive and negative cases and controls

	Maternal antibody status		
	Positive	**Negative**	**Total**
No. of cases of ALL	30	274	304
No. of controls	47	815	862

in the EOR estimation was increased only by 0.0225/0.037 = 61%, when antibody status was assessed in less than 900 controls, compared to the theoretically conceivable full cohort design, which would have required altogether 550,000 antibody assays.

6. Matching and Other Forms of Stratified Sampling

Matching is a procedure typically applied in nested case–control studies. It means stratified sampling of controls, such that for each individual case, the controls are chosen from, e.g., the same region, sex, and age group, etc., as the case.

The main reason for matching is that it creates similar distributions in controls and cases for the factors used as matching criteria, which leads to more balanced comparisons. Hence it tends to increase precision and efficiency in HR estimation, but only if the matching factors are (1) strong risk factors of the disease and (2) correlated with the exposure.

In addition, confounding due to observable but not quantifiable factors (like sibship, neighborhood, etc.) can be removed by close matching, but the bias is removed only if the data are properly analyzed. Especially in biobank studies matching the controls with each case on the storage time, freeze-thaw cycle and analytic batch improve comparability of measurements from frozen biological material (4).

As noted above, in Example 1, the control subjects were matched with the cases on time of diagnosis and age. Moreover, the controls were drawn from the same biobank/country and the same gender group, and the differences in maternal ages were less than 2 years compared to that in the cases. In addition, the dates of specimen collection were within ±2 months. Hence, matching on storage time was realized. It was not mentioned in the paper, whether the sera of each case and the matched controls were assayed in the same run, and whether they were matched on the freeze-thaw cycle, too.

Matching must always be accounted for in the statistical analysis of data either using simple Mantel–Haenszel estimators or by conditional logistic regression modeling (2).

A word of warning about *overmatching* should be said at this point. Matching a case with a control subject, namely, is a very different issue from matching an unexposed subject to an exposed one, e.g., in a randomized block experiment or in an observational cohort study – and is much trickier (2).

First, if one employs matching on an *intermediate* variable between exposure and outcome, a bias in effect estimation will be introduced. Second, matching on a *surrogate* or *correlate* of exposure, which is not a true risk factor of the outcome, would lead to loss of efficiency in estimation.

From the latter fact arises the idea of *counter-matching* (9): Choose a control which *is not similar* to the case with regard to the easily measured surrogate, which is strongly correlated with the exposure. This procedure tends to increase the statistical efficiency of the design, but necessitates a somewhat more complicated statistical analysis.

In CC studies, the efficiency may sometimes be improved by selecting the subcohort from the whole cohort at entry using stratified sampling, instead of simple random sampling (10). Useful stratification is based on a variable U, which is (a) surrogate of the main risk factor Z of interest, and (b) easy and cheap to measure, and available for the whole cohort. Stratification by U with few strata, the most informative of them getting the greatest sampling fractions, tends to increase efficiency in estimating the HRs associated with Z. Note, however, that this stratification may not be efficient for other risk factors.

7. Statistical Analysis of Case–control Data

In previous sections, we presented for illustrative purposes only very simple formulas used in crude estimation of the interesting hazard ratios. However, when analyzing case–control data arising from whatever design, more refined approaches are needed in order to propely allow for the specific sampling design used, including possible stratification or matching, as well as for confounding and effect-modification due to other relevant risk factors.

The most popular approach for statistical analysis is based on fitting the proportional hazards (PH) model, also known as the Cox model (3). In this model, the hazard (i.e., the theoretical incidence rate) of the outcome event at time (often age) t for a cohort member i possessing a risk factor profile $x_i = (x_{i1}, \ldots, x_{ip})$ is expressed as

$$\lambda_i\left(t, x_i; \beta\right) = \lambda_0(t)\exp\left(x_{i1}\beta_1 + \cdots + x_{ip}\beta_p\right).$$

In this model, $\lambda_0(t)$ is the baseline hazard depending on the basic time variable t. The parameters $\beta_1 \ldots, \beta_p$ are regression coefficients with the following interpretation. For each quantitative or binary explanatory variable (risk factor) X_j, the regression coefficient β_j is interpreted to be the logarithm of the hazard ratio (HR_j) associated with a unit change of the value of X_j. The hazard ratio itself is obtained as the antilogarithm: $\text{HR}_j = \exp(\beta_j)$.

In the estimation of these parameters, the typical method for nested case–control studies is based on maximizing the partial likelihood function, which is equivalent to fitting the equivalent conditional logistic regression model (3). This can nowadays be easily done by appropriate procedures found in many statistical programs (like R, SAS, S-Plus, and Stata). In case–cohort studies, the estimation is based on an analogous weighted pseudo-likelihood. The computational tools for the partial likelihood mentioned above can be used here, too, but they must be supplemented by certain additional calculations in order to obtain valid standard errors and confidence intervals, which take into account the special features of this design. See Samuelsen et al. (10) for details of such computations using the R environment.

Estimation of "absolute" risks is also feasible by proper weighting, as shown by Langholz and Borgan (8).

Full-likelihood solutions have also been recently developed, but they tend to be computationally quite challenging (using methods such as, e.g., EM algorithm, and MCMC simulation for Bayesian data augmentation).

8. Concluding Remarks

The properties of NCC and CC designs are now briefly compared on a few selected dimensions, based on more detailed discussions found, e.g., in references (3, 4).

The statistical efficiency in the two designs is roughly similar with the same amount of cases and controls, apart from some exceptional circumstances. Statistical analysis and inference in NCC studies are fairly straightforward with widely available software fitting conditional logistic regression or PH models. In CC studies, the analysis is somewhat more complicated, although software for PH models can be used when augmented with additional tricks to get valid SE, etc.

In the NCC design, only the time scale used in the definition of risk sets can be the time variable t in the baseline hazard of the PH model. However, in the CC design, the analysis of outcome rates based on the PH model is possible to conduct on different time scales (e.g., age, time since first exposure, or time since entry), because the subcohort members are not time-matched to the cases.

Missing data on risk factors may induce bias and inefficiency in the estimation of interesting parameters. In a NCC study, whenever very close matching was employed, a whole matched case–control set would be lost if the case had data missing on the risk factor(s) of interest. In CC studies, missingness of a few data items is less serious.

Quality and comparability of biological measurements based on frozen biological material are a serious concern in biobank-based studies. The NCC design allows each case and its own controls to be matched for analytic batch, storage time, and freeze-thaw cycle. This has the virtue that differential misclassification (1, 2) of exposure may be removed. In CC studies, the measurements for the subcohort members are performed at different times – typically earlier – than for the cases. This may more easily lead to differential misclassification and bias with unpredictable direction.

The possibility of investigating many diseases using the same control group for each group of cases is complicated (11) in the NCC study, and even impossible with too refined matching. In CC design, the same control group can easily serve for several diseases, because when no matching (on time or any other factor) is employed, no subcohort member is "tied" with any case.

In conclusion, cost-efficient sampling designs based on "case–controlling" are available and widely used in large-scale epidemiologic studies based on biobank cohorts. The NCC design is better suited for studies involving biomarkers that can be influenced by analytic batch, long-term storage, and freeze-thaw cycles. The CC design is useful especially when several outcomes are of interest, given that the measurements on stored materials remain sufficiently stable during the study. Finally, proper application of these designs requires well-organized follow-up systems for accurate identification of cases, deaths, and migrations occurring in the study cohort, as well as adequate statistical expertise in both planning and analysis of specific studies.

References

1. dos Santos Silva, I. (1999). *Cancer Epidemiology: Principles and Methods.* International Agency for Research on Cancer, Lyon.
2. Rothman, KJ., Greenland, S., and Lash, TL. (2008). *Modern Epidemiology, 3rd ed.* Lippincott Williams and Wilkins, Philadelphia, PA.
3. Borgan, Ø and Samuelssen, S.-O. (2003). A review of cohort sampling designs for Cox's regression model: Potentials for epidemiology. *Norsk Epidemiologi* **13**, 239–248. http://www.medisin.ntnu.no/ism/nofe/norepid/2003(2)%2008-Borgan.pdf
4. Rundle, A.G., Vineis, P. and Ahsan, H. (2005). Design Options for Molecular Epidemiology Research within Cohort Studies. *Cancer Epidemiology, Biomarkers and Prevention* **14**, 1899–1907.
5. Tedeschi, R., Bloigu, A., Ögmundsdottir, H.M. *et al.* (2007). Activation of Maternal Epstein-Barr Virus Infection and Risk of Acute Leukemia in the Offspring. *American Journal of Epidemioloy* **165**, 134–137.
6. Komulainen, K., Alanne, M., Auro, K. *et al.* (2006). Risk Alleles of *USF1*-Gene Predict Cardiovascular Disease of Women in Two Prospective Studies. *PLoS Genetics* **2**, e69.
7. Maclure, M. and Schneeweiss, S. (2001). Causation of Bias: The Episcope. *Epidemiology* **12**, 114–122.
8. Langholz, B. and Borgan, Ø. (1997). Estimation of Absolute Risk from Nested Case–control Data. *Biometrics* **53**, 768–775.
9. Langholz B and Borgan Ø. (1995). Counter-Matching: A Stratified Nested Case–control Sampling Method. *Biometrika* **82**, 69–79.
10. Samuelsen, S.-O., Ånestad, H. and Skrondal, A. (2007). Stratified Case–cohort Analysis of General Cohort Sampling Designs. *Scandinavian Journal of Statistics* **34**, 103–119.
11. Saarela, O., Kulathinal, S., Arjas, E. and Läärä, E. (2008). Nested Case–control Data Utilized for Multiple Outcomes: A Likelihood Approach and Alternatives. *Statistics in Medicine* **27**, 5991–6008.