eAppendix

Proposed formula of the sample size calculation for the case cohort design

We will start from the following well-known conventional formula of the sample size of the cohort study for a binary exposure variable.

$$N_{1full} = \frac{\left[Z_{\frac{\alpha}{2}} \sqrt{(1+\frac{1}{K})P_D(1-P_D)} + Z_\beta \sqrt{RR \cdot P_0(1-RR \cdot P_0) + \frac{P_0(1-P_0)}{K}} \right]^2}{[P_0(RR-1)]^2} \tag{1}$$

where $N_{1\text{full}}$ is the size of the exposed in the full cohort study, Z_c denotes $(1\text{-}c)$th standard normal quantile and RR is the relative risk or the ratio of the risk (incidence proportion) in the exposed (P_1) to that in the unexposed (P_0) (i.e., $RR=P_1/P_0$). In Eq.(1) K and P_D are defined by using the size of the unexposed, $N_{0\text{full}}$: K is defined as the ratio of the unexposed to the exposed or $K=N_{0\text{full}}/N_{1\text{full}}$ and P_D is the best common estimate of the incidence proportion under the null hypothesis defined as $P_D=(N_{1\text{full}}P_1+N_{0\text{full}}P_0)/N_{\text{full}} = P_0(RR+K)/(1+K)$ [1,2]. The total cohort size including both of the exposed and unexposed, N_{full}, can be expressed as

$$N_{full} = N_{1full}(1+K) \tag{2}$$

where $N_{1\text{full}}$ is given in (1).

In the article, we have proposed a sample size formula for the case-cohort study for a binary exposure variable as for (1) and (2). The exposed subjects (N_1) and total subjects (N) in the entire cohort for the case-cohort study with the same α, β, K, RR and P_0 as in (1) and (2) are formulated as

$$N_1 = N_{1full}(1+\frac{1}{m})$$

$$N = N_{1full}(1+K)(1+\frac{1}{m}) = N_{full}(1+\frac{1}{m}) \tag{3}$$

In (3), $N_{1\text{full}}$ and N_{full} are given in (1) and (2), respectively, and m is the ratio of the number of subjects in the subcohort to the expected number of cases in the entire cohort. The value of m should be assigned by a researcher who is planning the study. In most occasions, N_{full} in (2) rather than $N_{1\text{full}}$ in (1) in the full cohort study and N rather than N_1 in (3) in the case-cohort study may be more important as the exposure status is not normally known prior to the start of the study. The expected number of cases in the entire cohort for the case-cohort study is $P_1N_1+P_0N_0=P_0N_1(RR+K)=P_DN$ and the required size of the subcohort, n, is

$$n = mP_0N_1(RR+K) = mP_DN \tag{4}$$

where N_1 and N are given in (3).

For example, imagine that we are designing the study on multiple adverse drug reactions of a new statin, a drug used for patients with hypercholesterolemia. According to Jacobson[3.], for most statins, the muscle event characterized by the blood creatinine phosphokinase (CPK)>10ULN (where ULN means 'upper limit of normal range') may occur with the incidence proportion of 0.1 to 0.5% while the most serious type of muscle event, rhabdomyolysis occurs at most 15 in million users. In addition, statins may cause liver function abnormality and both of the increase of the blood alanine aminotrasferase (ALT) level>3ULN and the increase of the blood asparate aminotransferase (AST) level>3ULN occur in 0.1% or more of the users. Statins also precipitate renal events and the incidence proportion is more than 0.4% for proteinuria and more than 2% for hematuria. Imagine that a case-cohort study is designed to detect the increase of the incidence proportions for one or more of those adverse events except for rhabdomyolysis which may be judged to be too rare to estimate in this type of the study. The 5 target adverse events (CPK increase, ALT increase, AST increase, proteinuria and hematuria) have then the incidence proportion of 0.1% or higher. It is also assumed that the new statin is compared with the old statins as a whole and the ratio of the unexposed (those who use one of the old statins) to the exposed (those who use the new statin) (K) is, at the best guess, 3. When P_0=0.001 and K=3 are used, the required sample size of N_{full} in (2) to detect at least four times increase of the incidence proportion (i.e., RR=4) is 9986. The estimates of N in Eq.(3) for m=1, 2 and 5 will be then 19,972, 14,979 and 11,984, respectively where m is assigned by the researcher. The expected number of cases NP_D (where P_D=0.00175) for m=1, 2 and 5 will be 35, 27 and 21 so that the required size for the subcohort (n) will be estimated as 35, 54 and 105, respectively. As the event is relatively rare, most cases will occur outside the subcohort and in such a case, the expected number of n_{detail} will be 70, 81 and 126, for m=1, 2 and 5, respectively. Thus, the combination of the required sample size of the entire cohort (N) and those classified as a subcohort member or case (n_{detail}) expressed as (N, n_{detail}) will be (19,971, 70) for m=1, (14,979, 81) for m=2 and (11,983, 126) for m=5. From those 3 sets (or more sets if appropriate), the researcher may choose the best value of

m in terms of the available N and the cost needed to have the detailed information from n_{detail} subjects.

In the next section, we will show that Eq.(3) can be given as an approximation to a more accurate formula. The empirical power and type I empirical error of the formula in (3) is then compared with the nominal power and type I error by simulations in section 4. Before moving to the next 2 sections, however, we may emphasize that Eq.(3) is intuitively appealing as the case-cohort study can be regarded as a case-control study where controls are randomly selected from the non-cases at the beginning of the study. Indeed, Kim et al.[4] showed that the conventional sample size formula for the case-control study yields the empirical power similar to those by Cai and Zeng[5]. It is known that the ratio of the variance of an estimate for the log odds ratio from a case-control study to that from a cohort study yielding the same number of cases is given as $(1+1/m)$ [6]. This indicates that the variance of a full cohort study is equal to that of a case-control or case-cohort study conducted within the $(1+1/m)$ times larger entire cohort. Therefore, it may be intuitively understood that N is $(1+1/m)$ times larger than N_{full} as given in (3).

Theoretical consideration

Assume that from an entire cohort with N members consisting of N_1 exposed and N_0 unexposed subjects ($N=N_1+N_0$), n subjects are randomly selected as subcohort members at the beginning of the study of which n is the quantity equal to m times the number of the expected cases or $n=mNP_D$. When q is defined as the sampling fraction of the subcohort ($n=qN$), the relation between q and m may be given as $q=mP_D$. When the observed number of the exposed cases in the entire cohort is defined as a, that of the unexposed as b, the observed number of the exposed subcohort as n_1 and that of the unexposed subcohort as n_0, ($n_0=n-n_1$) the estimate of the risk (incidence proportion) in the exposed and that in the unexposed may be given as $\hat{P}_1 = a/n_1(n/N)$ and $\hat{P}_0 = b/n_0(n/N)$, respectively.

To estimate the variance of the risk difference $V[\hat{P}_1 - \hat{P}_0] = V[\hat{P}_1] + V[\hat{P}_0] - 2Cov[\hat{P}_0, \hat{P}_1]$, where $\hat{P}_1 = a/n_1(n/N)$ and $\hat{P}_0 = b/n_0(n/N)$, a, b and n_1 may be regarded as variables and n, N and N_1 as constants. It may be noted that a and n_1 (and b and n_0) may be considered to be independent of each other**,** as the procedure of selecting subcohort members done at the beginning of the study is independent of the event occurrence some time during the study. The distribution of a and b may be described by the binomial distribution and that of n_1 and n_0 may be described by the hypergeometric distribution. Therefore, E[a]= N_1P_1 , E[b]= N_0P_0 , V[a]= $N_1P_1(1-P_1)$, V[b]= $N_0P_0(1-P_0)$, E[n_1]= $n/(1+K)$, E[n_0]= $nK/(1+K)$ and V[n_1]= V[n_0]= $nK/(1+K)^2[(N-n)/(N-1)]$ where E[x] and V[x] denote the mean and variance of x, respectively. Using the delta method, V[$f(x)$]≈[$df(m_x)/dx$]^{2}V[x] where m_x = E[x], assuming x=n_1 for a function $f(x)=1/x$, V[$1/n_1$] may be expressed as $V[1/n_1] \approx [K(1+K)^2/n^3][(N-n)/(N-1)]$. Similarly, assuming x=n_0, V[$1/n_0$] may be expressed as $V[1/n_0] \approx [(1+K)^2/(n^3K^3)][(N-n)/(N-1)]$. Using the relationship, $V[x/y] \approx V[x]/E[y]^2 + E[x]^2V[1/y]$ where x is independent of y, $V[\hat{P}_1] = V[a/n_1](n/N)^2$, $V[\hat{P}_0] = V[b/n_0](n/N)^2$ and $(N-n)/(N-1) \approx 1-q$, $V[\hat{P}_1]$ and $V[\hat{P}_0]$ may be given, respectively, as

$$V[\hat{P}_1] \approx \frac{P_1(1-P_1)}{N_1} + \frac{KP_1^2}{n}\left[\frac{N-n}{N-1}\right] \approx \frac{P_1(1-P_1)}{N_1} + \frac{KP_1^2}{n}(1-q) \tag{5}$$

and

$$V[\hat{P}_0] \approx \frac{P_0(1-P_0)}{N_0} + \frac{P_0^2}{Kn}\left[\frac{N-n}{N-1}\right] \approx \frac{P_0(1-P_0)}{KN_1} + \frac{P_0^2}{Kn}(1-q) \tag{6}$$

From the relationship $m(N_0P_0+N_1P_1)=n$, Eqs (5) and (6) may be rewritten respectively as

$$V[\hat{P}_1] \approx \frac{P_1(1-P_1)}{N_1} + \frac{P_1K \cdot RR}{N_1m(K+RR)}(1-q) \tag{7}$$

and

$$V[\hat{P}_0] \approx \frac{P_0(1-P_0)}{KN_1} + \frac{P_0}{KN_1m(K+RR)}(1-q) \tag{8}$$

The estimate for the covariance for $\hat{P}_1$ and $\hat{P}_0$, $Cov[\hat{P}_1, \hat{P}_1]$, given as

$Cov[\hat{P}_1, \hat{P}_1] = E[a]E[b]\left(\frac{n}{N}\right)^2\left[E[\frac{1}{n_1 n_0}] - E[\frac{1}{n_1}]E[\frac{1}{n_0}]\right]$ may be obtained as follows. In general, using the first three terms of a Taylor's expansion of f(x) about the point E[x]= m_x, $f(x) = f(m_x) + (x-m_x)f^{(1)}(m_x) + [(x-m_x)^2/2]f^{(2)}(m_x) + \ldots$, where $f^{(1)}$ and $f^{(2)}$ are the first and second derivatives of $f(x)$ defined as $f(x)=1/x$, we may have the relationship E[$f(x)$] ≈ $(V[x]/m_x^2+1)/m_x$. Assuming $x=n_1$, we may have $E[1/n_1] \approx (K+1)/n[1+(1-q)K/n]$ and assuming $x=n_0$, we may have $E[1/n_0] \approx (K+1)/(nK)[1+(1-q)/(nK)]$. As $1/(n_1 n_0)$ can be expressed as $1/(n_1 n_0) = 1/(n_1(n-n_1)) = [(1/n_1)+(1/(n-n_1))]/n$, $E[1/(n_1 n_0)]$ is given as $E[1/(n_1 n_0)] = (E[1/n_1] + E[1/n_0])/n \approx (K+1)^2/(n^2 K) + (1-q)(K+1)(K^3+1)/(n^3 K^2)$. On the other hand, $E[1/n_1]E[1/n_0]$ is given as $E[1/n_1]E[1/n_0] \approx (K+1)^2/(n^2 K) + (1-q)(K+1)^2(K^2+1)/(n^3 K^2) + (1-q)^2(K+1)^2/(n^4 K)$, where the contribution of the last term $((1-q)^2(K+1)^2/(n^4 K))$ to this quantity is minor and may be ignored. We may then have the relationships, $E[1/(n_1 n_0)] - E(1/n_1)E(1/n_0) \approx -(1-q)(K+1)^2/(n^3 K)$. As $E[a]=N_1 P_1$ and $E[b]=N_0 P_0$, the following equation is derived.

$$Cov[\hat{P}_1, \hat{P}_0] = -\frac{(1-q)P_1}{N_1 m(K+RR)} \tag{9}$$

From the relation $V[\hat{P}_1 - \hat{P}_0] = V[\hat{P}_1] + V[\hat{P}_0] - 2Cov[\hat{P}_0, \hat{P}_1]$ and Eqs. (7) to (9), the variance of the risk difference $V[\hat{P}_1 - \hat{P}_0] = V[\hat{P}_1] + V[\hat{P}_0] - 2Cov[\hat{P}_1, \hat{P}_0]$ is given as:

$$V[\hat{P}_1 - \hat{P}_0] = \frac{1}{N_1}[P_1(1-P_1) + \frac{P_0(1-P_0)}{K}](1 + \frac{1}{m}f_1)$$

where

$$f_1 = \frac{(K \cdot RR + 1)^2(1-q)}{(K+RR)[K \cdot RR(1-P_1) + (1-P_0)]} \tag{10}$$

Under the null hypothesis, RR=1, the variance of the risk difference is given as:

$$V[\hat{P}_1 - \hat{P}_0] = \frac{1}{N_1}[P_D(1-P_D)(1+\frac{1}{K})](1+\frac{1}{m}f_0)$$

where

$$f_0 = \frac{1-q}{1-P_D} \tag{11}$$

Using the relationship, $\delta=SE_0Z_{\alpha/2}+SE_1Z_\beta$, where δ is the risk difference representing the worthwhile effect or $P_0(RR-1)$, SE_1 is $\sqrt{v_1}$, where v_1 is given as $V[\hat{P}_1-\hat{P}_0]$ in (10) and SE_0 is $\sqrt{v_0}$ where v_0 is given as $V[\hat{P}_1-\hat{P}_0]$ in (11), the following relation is derived.

$$N_1=\frac{\left[Z_{\frac{\alpha}{2}}\sqrt{(1+\frac{1}{K})P_D(1-P_D)(1+\frac{1}{m}f_0)}+Z_\beta\sqrt{[RR\cdot P_0(1-RR\cdot P_0)+\frac{P_0(1-P_0)}{K}](1+\frac{1}{m}f_1)}\right]^2}{[P_0(RR-1)]^2} \quad (12)$$

The value of f_0 in (11) does not depend on K or RR, and provided that the event is rare and both ($1-P_D$) and ($1-q$) are near 1, f_0 may be approximated by 1. Similarly, supposing a rare event occurrence, f_1 in (10) may be approximated by 1 provided that either K or RR is close to 1. The sample size in (3) can be obtained when replacing both of f_0 and f_1 in (12) by 1.

Monte Carlo simulation

For a number of sets of β, K, P_0, RR and m, the values of N_1 in (3) and (12) and N (estimated as $N_1(1+K)$) were calculated (α was fixed as 0.05 through the simulations) to estimate the empirical power. Using the same α, β, K and m, the simulation assuming $RR=1$ and $P_0=P_1=P_D$ was also performed to know Type I empirical error. The entire cohort with N subjects was assumed to consist of the N_0 unexposed (subject $1,2,\cdots N_0$) and the N_1 exposed (subject $N_0+1, N_0+2,\cdots N$). The size of the subcohort, n was calculated as an integer obtained by rounding up the quantity, mNP_D. In the first step of the simulation, n subjects were randomly selected from the entire cohort with N members. In the next step, N uniform random numbers between (0,1), $x_j (j=1,2,\cdots,N)$, were generated and the j-th subject was regarded as a case if $x_j<P_0 (j=1,2,\cdots,N_0)$ or $x_j<P_1 (j=N_0+1,N_0+2,\cdots,N)$ but otherwise as a non-case. For the cases, the time of the event occurrence, t, was defined as $t=-log(1-x_j)/\lambda_i$ where $\lambda_i=-log(1-P_i)$ (i=0,1) assuming that the constant hazard (i.e., exponential distribution) over the fixed observation period t_{obs}, which is set as unity ($t_{obs}=1$). In addition, to know whether the alteration of the assumption of the constant hazard affects the results, some simulations were made where the event was assumed to occur according to the Weibull distribution. With this assumption, the cumulative probability of the event occurrence till time, t, or $Pr(t)$ may be

given as $Pr(t)=1-exp[-(\lambda_i t)^s]$ where λ_i is the reciprocal of the scale parameter for the unexposed (i=0) and that for the exposed (i=1) and s is the shape parameter where s in the exposed is assumed to be the same as that in the unexposed. Using t_{obs}=1, λ_i may be given as $\lambda_i = -[\log(1-P_i)]^{\frac{1}{s}}$. For the cases, using the relationship $x_j=Pr(t)$, t may be given as $t = [\log(1-x_j)/\log(1-P_i)]^{\frac{1}{s}}$. When s=1 the hazard is constant (and the distribution is exponential) while when s<1 the event rate decreases over time and when s>1 the rate increases over time. When P_0 or P_1 is in the range between 0.001 and 0.3, as employed in the current simulation studies, 80% or more of the events are expected to occur during the first half of the observation period when s=0.3 while 80% or more will occur during the last half of the period when s=2.5. To compare the empirical power and type I empirical error of the current article with that by Cai and Zeng[5], their equation (11) $\tilde{n} = nBP_D/(n-B(1-P_D))$ where $\tilde{n}$ and n correspond to n and N in this study, respectively, has been converted, using the relationship $\tilde{n} = mnP_D$ (or $\tilde{n} = mNP_D$ by the symbols in this article), to:

$$N = \frac{B[1+m(1-P_D)]}{m}$$

where

$$B = \frac{(Z_{\alpha/2}+Z_\beta)^2}{\theta^2 p_1 p_2 P_D} \tag{13}$$

In the above equation, $Z_{\alpha/2}$ was used for the two-sided test instead of Z_α employed in the paper by Cai and Zeng[5]. Using the ratio of the unexposed to exposed (K) defined in this study, the quantities p_1 and p_2, the proportion of the exposed and that of the unexposed in the entire cohort, are given as $p_1=1/(1+K)$ and $p_2=K/(1+K)$, respectively, and θ is defined as $\theta=\log(\Lambda_0/\Lambda_1)$, where Λ_1 and Λ_0 are the cumulative hazard in the exposed and unexposed, respectively, and given as $\Lambda_i=-log(1-P_i)$ by using the cumulative incidence proportion, P_1 and P_0 in the entire cohort defined in this study. It may be noted that in equation (11) in Cai and Zeng[5], the size of the subcohort is formulated as a function of the given size of the entire cohort while in Eq.(13) above, both of the entire cohort size (N) and subcohort size, n (given as $n=mNP_D$), are formulated explicitly as a function of m which is assigned by the researcher as in (3) and (4). In addition, to compare the empirical power and type I empirical error with those by Kim et al.[4], the formula for the power

$$Power = \Phi\{[Z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})(1/n_D+1/n_C)}+(p_{E|D}-p_{E|C})]/[\sqrt{p_{E|D}(1-p_{E|D})/n_D+p_{E|C}(1-p_{E|C})/n_C}]\}$$

given as their equation (2) has been converted, using the relationship $n_D=P_Dn/q=P_DN$, $n_C=n(1-P_D)=mn_D(1-P_D)= mP_DN(1-P_D)$, to:

$$N = \frac{\{Z_{\alpha/2}\sqrt{\bar{p}(1-\bar{p})(1+1/[m(1-P_D)]}+Z_\beta\sqrt{p_{E|D}(1-p_{E|D})+p_{E|C}(1-p_{E|C})/[m(1-P_D)]}\}^2}{P_D(p_{E|D}-p_{E|C})^2}$$

where

$$p_{E|C} = 1/(1+K)$$
$$p_{E|D} = (RRp_{E|C})/(1+p_{E|C}(RR-1))$$

and

$$\bar{p} = (n_Dp_{E|D}+n_Cp_{E|C})/(n_D+n_C) = (p_{E|D}+m(1-P_D)p_{E|C})/(1+m(1-P_D)) \quad (14)$$

N_1, N_0 and n are estimated as $N_1=N/(1+K)$, $N_0=KN_1$ and $n=NmP_D$ using N estimated in (14). It may be noted that in equation (2) in Kim et al.[4], the power is estimated for a subcohort sampled from a given size of the entire cohort with a sampling fraction of q. In Eq. (14) above, however, both of the size of the entire cohort, N, and that of the subcohort, n, are formulated explicitly as a function of m as in (3) and (4). The simulation data were analyzed by the Cox model as in the paper by Self and Prentice[7] using SAS 9.1 (SAS, North Carolina) with the program obtained through Statlib (http://lib.stat.cmu.edu/general/robphreg)[8].

Results of simulation

eTable 1 shows the values of N_1 estimated by (3) and (12) as well as empirical power levels estimated as the fraction of the trials where the null hypothesis (RR=1) was rejected in 10,000 iterations for each combination of α (fixed as 0.05, two-sided), β(0.1 or 0.2), K (0.25, 0.5, 1, 2 or 4), P_0 (0.001, 0.01 or 0.01) and m (1 or 5 [for P_0=0.001 and 0.01] and 1 or 3 [for P_0=0.1]) for RR=2. eTable 2 shows the estimates for N_1 and empirical power levels for RR=3 using the same combination of α, β, P_0, K and m as in eTable 1. The estimates of N_1 (dotted lines) and N (solid lines) in Eq.(3) for RR=3 (as in eTable 2) are shown in eFigure 1 and the empirical power in eTable 2 are shown in eFigure 2. As shown in eFigure 1, for the particular combination of α, β, P_0, RR and m, N is the smallest when K=1. When m=1, the sample size estimated by (3) is larger than that by (12) when K<1 (open versus closed circles (for β=0.1) and open versus closed triangles (for β=0.2) in eFigure 1). However, the estimate by (3) is smaller than that by (12) when $K>1$. The estimate by (3) also differs from that by (12) when $m=5$ (for $P_0=0.001$ and 0.01) or $m=3$ (for $P_0=0.1$), though the difference is less remarkable. As shown in eTables 1 and 2 and eFigure 2, Eq.(12) (closed circles ($\beta=0.1$) and closed triangles ($\beta=0.2$) in eFigure 2) gives the empirical power that is relatively stable and near to the nominal power, ($1-\beta$). The empirical power obtained by Eq.(3) (open circles ($\beta=0.1$) and open triangles ($\beta=0.2$) in eFigure 2) tends to decrease with the increase of K. The empirical power by Eq.(3) is however larger than or near to the nominal power in general as long as the parameters are in the range examined in eTables 1 and 2 and eFigure 2. In eTable 3, results for some additional simulations are shown for a selected combination of parameters where α, β, RR and m are fixed as $\alpha=0.05$, $\beta=0.2$, $RR=3$ and $m=1$. As shown in eTable 3, the Type I empirical error is near to the nominal error (0.05) irrespective of the method used. When the Weibull distribution is assumed for the occurrence of events estimated by the model in (12), the empirical power for $s=0.3$ and $s=2.5$ (s is the shape parameter) is similar to that for $s=1$ where the hazard is constant. Eq.(13) derived from the formula from Cai and Zeng[5] tends to give an empirical power less than nominal power (i.e., underpower) when $K<1$ but show the opposite (overpower) when $K>1$. Eq. (14) derived from the formula for a case-control study in Kim et al.[4] provides the empirical power close to the nominal power when $P_0=0.01$ and 0.001 but tends to yield too large estimate of N when $P_0=0.1$.

Discussion and conclusion

In Eq.(3) or (12) m but not q (sampling fraction) is assigned by the researcher as the key quantity though it is q that is often mentioned in the case-cohort study such as 'a 10% random sample of subjects was selected from the entire cohort to serve as the subcohort'[9]. Though q and m are simply related as $q=mP_D$, the use of m may make the researcher realize that the particular value of q (e.g., 10%) can be either too large, optimal or too small depending on P_D. There may be no single answer about what is the best value of m. If the estimation for all or some of co-variates is quite costly, (e.g., expensive laboratory test is required), the value of n_{detail} may be minimized. For a single event, the expected number of cases and the size of the subcohort are given as $(1+1/m)N_{\text{full}}P_D$ and $m(1+1/m)N_{full}P_D$, respectively. The expected number of cases who have been selected as a subcohort member is $m(1+1/m)N_{full}P_D^{\,2}$. The expected value of n_{detail} may be therefore given as $(1+1/m)(m+1-P_D)N_{\text{full}}P_D$ and the smallest when $m=1/\sqrt{1-P_D}$. For multiple events, values of N_{full} may be calculated for each kind of the events of interest and the largest N_{full} (defind as N_{fullmax}) may be adopted. When P_D for the event used to estimate N_{fullmax} is defined as $P_{D\min}$, the expected number of the cases for that event may be given as $(1+1/m)N_{\text{fullmax}}P_{D\min}$. If the total number of the cases for all kinds of the events of interest is r times larger than this quantity, or $r(1+1/m)N_{\text{fullmax}}P_{D\min}$, n_{detail} may be given as $(1+1/m)(m-mrP_{D\min}+r)N_{\text{fullmax}}P_{D\min}$ which will be the smallest when $m=\sqrt{r/(1-rP_{D\min})}$ where n_{detail} is $(\sqrt{r}+\sqrt{1-rP_{D\min}})^2 N_{full\max}P_{D\min}$. For the scenario of the statin study (see the paragraph following Eq.(4)), the incidence proportion in the unexposed is 0.1% for one of the 3 muscle or hepatic events (the increase of CPK>10ULN, that of AST>3ULN, and that of ALT>3ULN), 0.4% for proteinuria and 2% for hematuria. If the actual risk is increased by a factor of 4 for all of these 5 target events, r may be given as 27 (3 for the 3 events with the incidence of 0.1% plus 4 for the event with the incidence of 0.4% plus 20 for the incidence of 2%) and $\sqrt{r/(1-rP_{D\min})}$ may be around 5. The incidence of the event in the real study is however difficult to predict in advance. For instance, the incidence may not be in fact affected by the exposure as opposed to the (alternative) hypothesis. In the imaginary statin study, the possible range of P_0 is 0.1-0.5% for the muscle event and 0.1% or more for the two hepatic events. The incidence in the actual study can be higher (e.g., 0.2% or more) than that used in the sample size estimation because the smallest value of the possible range of the incidence (i.e., 0.1%) may be used for the sample size estimation. We believe however that in many occasions m=3 to 5 may be chosen as an optimal value of m. This is because the entire cohort size is just 1.2 to 1.3 (the value of 1+1/m when m=3 to 5) times larger than its minimum value (the size of the full cohort) while the ratio of n_{detail} to its smallest quantity $(1+1/m)(m-mrP_{D\min}+r)/(\sqrt{r}+\sqrt{1-rP_{D\min}})^2$ is less than 2 irrespective of the value of r provided that r≥1. For an existing cohort with a specific sample size defined as $N_{\text{available}}$ (which should be greater than N_{full}), the possible smallest number of m defined as m_{min} (which is not necessarily an integer) is given as $m_{\min}=N_{\text{full}}/(N_{\text{available}}-N_{\text{full}})$ because the relationship $N_{\text{available}}=(1+1/m_{\min})N_{\text{full}}$ should hold. If $N_{\text{available}}$ is much larger than N_{full} and $m_{\min}<1$, the use of m_{min} may result in too many cases and too small subcohort. For example, if a large cohort of half million subjects is available for our imaginary statin study where N_{full}=9,986, m_{min} is calculated as 0.02. The resultant case-cohort study is obviously inefficient as the expected number of cases (NP_D where N=500,000 and P_D=0.00175) is 875 and the required sample size of the subcohort ($Nm_{min}P_D$) is 18. One possible option for such occasion is to use a fraction of the available cohort as the entire cohort of which the size is estimated as in the earlier parts of this manuscript. As in eTable 3, several different approaches for sample size estimation can yield similar results as noted by Kim et al.[4] The best approach may depend on the circumstance where the sample size formula is used. For the full cohort study, for instance, instead of the conventional formula of the sample size (Eq.(1)), simplified formulae such as that shown by Schulz and Grimes[10] or that by Torgerson and Miles[11] may be useful when no guidance from a statistician is available. In such a circumstance, Eq.(3) may be advantageous particularly when a case-cohort study is a candidate among others (such as a full cohort study or nested case-control study) because the comparison between different designs is straightforward. Eq.(3) provides a satisfactory estimate in general though Eq.(3) may somewhat underestimate the sample size when both K and RR are larger than 1.

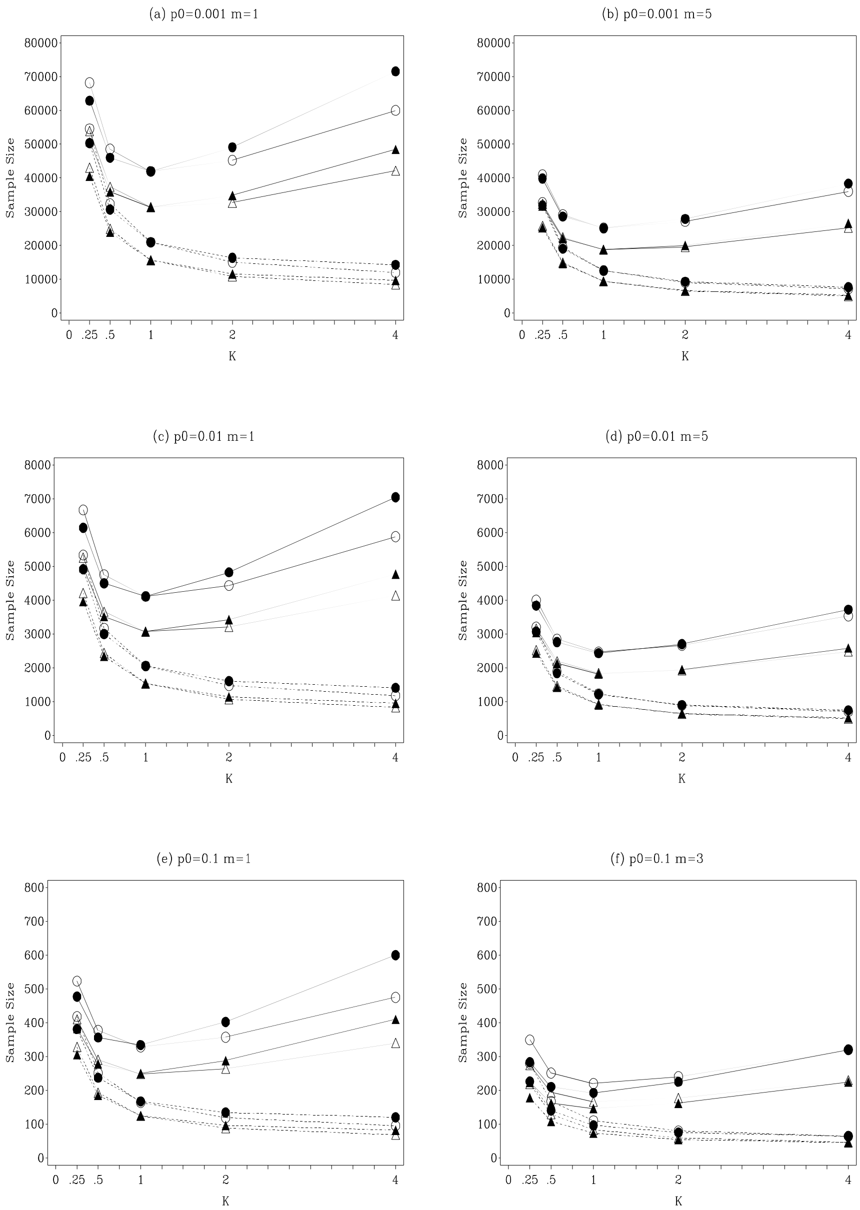
References

Clinical Epidemiology. 1999; **52**: 1165-1172.

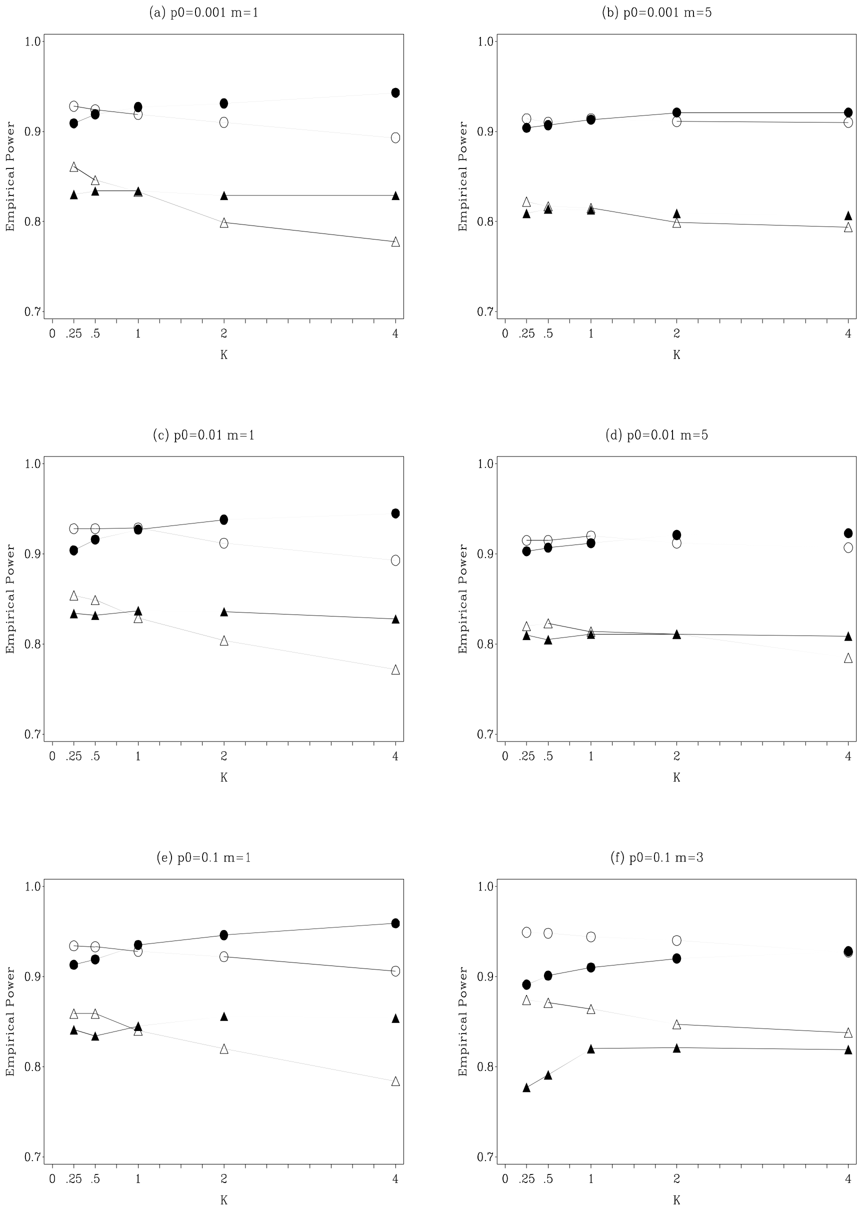
1. McNeil D. . Sample size. In: *Epidemiological research methods.* New York: John Wiley & Sons.; 1996.
2. Strom BL. Sample size considerations for pharmacoepidemiology studies. In: *Pharmacoepidemiology*. 4th ed. West Sussex: John Wiley & Sons. 2005; 29-36.
3. Jacobson TA. Statin safety: Lessons from new drug application for marketted statins. *Am. J. Cardiol.* 2006; **97** (suppl.): 44C-51C.
4. Kim MY, Xue X, Du Y. Approaches for calculating power for case-cohort studies. *Biometrics.* 2006; **62**: 929-933.
5. Cai J, Zeng D. Sample size/power calculation for case-cohort studies. *Biometrics.* 2004; **60**: 1015-1024.
6. Clayton D, Hill M. Likelihood for the odds ratio. In: *Statistical models in epidemiology*. Oxford: Oxford University Press; 1993: 166-174
7. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*. 1988; **16**:64-81.
8. Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics.* 1994;

50: 1064-72.

9. Strom B, Schinnar R, Bilker WB, Feldman H, Farrar JT, Carson JL. Gastrointestinal tract bleeding associated with naproxen sodium vs ibuprofen. *Archives of Internal Medicine.* 1997; **157**: 2026-2031.
10. Schulz KF, Grimes DA. Sample size calculations in randomized trials: mandatory and mystical. *Lancet.* 2005; **365**: 1348-1353.
11. Torgerson DJ, Miles JNV. Simple sample size calculation. *Journal of Evaluation in Clinical Practice*. 2007; **13**:952-953.



eFigure 1. Size of the exposed subjects (dotted lines) and that of the entire cohort (solid lines). Open symbols (circles and triangles) indicate estimates by Eq.(3) while closed symbols indicate estimates by Eq.(12). Open (Eq.3) and closed (Eq.12) circles are estimates for = 0.1 and open (Eq.3) and closed (Eq.12) triangles are estimates for = 0.2. The figures are shown for six combinations of P_0 and m. Other parameter values are set as = 0.05 and RR=3.



eFigure 2. Empirical powers of Eq.(3) and Eq.(12). Open symbols (circles and triangles) indicate empirical powers by Eq.(3) while closed symbols indicate those by Eq.(12). Open (Eq.3) and closed (Eq.12) circles are empirical powers for = 0.1 and open (Eq.3) and closed (Eq.12) triangles are those for = 0.2. The figures are shown for six combinations of P_0 and *m*. Other parameter values are set as = 0.05 and *RR* = 3.

eTable 1. Empirical power of the case-cohort study (RR=2)

				K=0.25		*K*=0.5		*K*=1		*K*=2		*K*=4	
P_0	β	*m*	*method*	N_1	*Power*	N_1	*Power*	N_1	*Power*	N_1	*Power*	N_1	*Power*
0.001	0.2	1	Eq.3	125921	0.834	73470	0.825	47021	0.818	33612	0.793	26795	0.776
			Eq.12	120283	0.818	71391	0.816	47022	0.818	34950	0.811	29024	0.806
		5	Eq.3	75553	0.819	44082	0.818	28213	0.804	20167	0.796	16077	0.795
			Eq.12	74368	0.805	43626	0.809	28185	0.814	20419	0.806	16519	0.799
	0.1	1	Eq.3	162420	0.921	96331	0.912	62946	0.909	45974	0.908	37317	0.900
			Eq.12	152707	0.900	92715	0.905	62948	0.910	48365	0.915	41345	0.924
		5	Eq.3	97452	0.906	57799	0.907	37768	0.909	27584	0.903	22390	0.905
			Eq.12	95445	0.897	57024	0.900	37730	0.907	28041	0.910	23191	0.914
0.01	0.2	1	Eq.3	12402	0.831	7240	0.825	4637	0.816	3317	0.795	2645	0.778
			Eq.12	11841	0.819	7034	0.818	4638	0.817	3451	0.815	2869	0.813
		5	Eq.3	7441	0.821	4344	0.816	2782	0.803	1990	0.804	1587	0.800
			Eq.12	7249	0.808	4257	0.809	2754	0.814	1998	0.805	1618	0.806
	0.1	1	Eq.3	16003	0.920	9494	0.916	6206	0.914	4534	0.905	3681	0.894
			Eq.12	15036	0.901	9135	0.900	6208	0.916	4775	0.919	4085	0.924
		5	Eq.3	9602	0.903	5697	0.911	3724	0.909	2721	0.910	2209	0.908
			Eq.12	9309	0.897	5566	0.904	3687	0.905	2742	0.910	2270	0.911
0.1	0.2	1	Eq.3	1050	0.850	617	0.840	398	0.826	287	0.809	230	0.788
			Eq.12	997	0.831	598	0.831	400	0.833	301	0.831	253	0.831
		3	Eq.3	700	0.854	412	0.852	266	0.842	192	0.835	154	0.827
			Eq.12	614	0.801	367	0.813	243	0.807	180	0.813	149	0.813
	0.1	1	Eq.3	1361	0.932	810	0.931	532	0.918	390	0.913	318	0.904
			Eq.12	1269	0.907	777	0.914	534	0.924	416	0.929	359	0.934
		3	Eq.3	907	0.934	540	0.932	355	0.935	260	0.927	212	0.922
			Eq.12	791	0.895	480	0.901	324	0.911	246	0.914	208	0.920

Empirical power calculated as the proportion of the trials with hazard ratio estimated by the Cox regression analysis is significantly different from 1 (α=0.05, two-sided) in 10,000 iterations shown for each combination of *K, β,* P_0, and *m*. The data for *RR*=2 are shown. Nominal power is (1-β).

Note; *K*= ratio of the unexposed to exposed; N_1= size of the exposed in the entire cohort; *β*= value of β in Eqs. (3) and (12); P_0=probability of failure in the unexposed; *RR*=relative risk (incidence proportion ratio) in the exposed to unexposed; *m*=ratio of the subcohort to the expected number of cases in the entire cohort;

eTable 2. Empirical power of the case-cohort study (RR=3)

				K=0.25		*K*=0.5		*K*=1		*K*=2		*K*=4	
P_0	β	*m*	*method*	N_1	*Power*	N_1	*Power*	N_1	*Power*	N_1	*Power*	N_1	*Power*
0.001	0.2	1	Eq.3	43080	0.861	24918	0.846	15664	0.833	10899	0.799	8430	0.778
			Eq.12	40577	0.830	23943	0.834	15666	0.834	11614	0.829	9692	0.829
		5	Eq.3	25848	0.822	14951	0.817	9399	0.815	6540	0.799	5058	0.794
			Eq.12	25324	0.809	14738	0.814	9387	0.813	6677	0.809	5314	0.807
	0.1	1	Eq.3	54550	0.928	32337	0.924	20969	0.919	15074	0.910	11996	0.893
			Eq.12	50283	0.909	30652	0.919	20971	0.927	16360	0.931	14310	0.943
		5	Eq.3	32730	0.914	19403	0.910	12582	0.914	9045	0.911	7198	0.910
			Eq.12	31854	0.904	19044	0.907	12565	0.913	9294	0.921	7666	0.921
0.01	0.2	1	Eq.3	4214	0.854	2441	0.849	1537	0.829	1071	0.804	829	0.772
			Eq.12	3967	0.834	2345	0.832	1538	0.837	1143	0.836	956	0.828
		5	Eq.3	2529	0.820	1465	0.823	922	0.814	643	0.811	497	0.785
			Eq.12	2441	0.810	1424	0.805	910	0.811	649	0.811	518	0.809
	0.1	1	Eq.3	5339	0.928	3168	0.928	2056	0.929	1479	0.912	1177	0.893
			Eq.12	4917	0.904	3002	0.916	2058	0.927	1609	0.938	1410	0.945
		5	Eq.3	3204	0.915	1901	0.915	1234	0.920	888	0.912	707	0.907
			Eq.12	3074	0.903	1841	0.907	1217	0.912	902	0.921	745	0.923
0.1	0.2	1	Eq.3	328	0.859	193	0.859	124	0.840	88	0.820	68	0.784
			Eq.12	306	0.841	185	0.834	125	0.845	96	0.856	82	0.854
		3	Eq.3	219	0.874	129	0.871	83	0.864	59	0.847	46	0.838
			Eq.12	178	0.777	108	0.791	73	0.820	54	0.821	45	0.819
	0.1	1	Eq.3	418	0.934	251	0.933	164	0.928	119	0.922	95	0.906
			Eq.12	381	0.913	237	0.919	167	0.935	134	0.946	120	0.959
		3	Eq.3	279	0.949	167	0.948	110	0.944	80	0.940	64	0.927
			Eq.12	226	0.891	140	0.901	96	0.910	75	0.920	64	0.928

Empirical power calculated as in eTable 1. The data for RR=3 are shown.
Note; see footnotes to eTable 1 for an explanation of the abbreviations.

eTable 3 Empirical power and Type I error of the case-cohort study (RR=3)

				K=0.25			K=1			K=4		
P_0	β	m	Method	N	Power	Type I error	N	Power	Type I error	N	Power	Type I error
0.001	0.2	1	Eq.3 (s=1)	53850	0.852	0.045	31328	0.826	0.047	42150	0.775	0.044
			Eq.12 (s=0.3)	50720	0.824	0.046	31330	0.824	0.051	48460	0.814	0.045
			Eq.12 (s=1)	50720	0.834	0.042	31330	0.830	0.050	48460	0.821	0.045
			Eq.12 (s=2.5)	50720	0.835	0.045	31330	0.818	0.050	48460	0.817	0.043
			Eq.13 (s=1)	31168	0.609	0.041	25938	0.764	0.049	57915	0.881	0.048
			Eq.14 (s=1)	47143	0.805	0.043	28864	0.805	0.045	45435	0.805	0.044
0.01	0.2	1	Eq.3 (s=1)	5268	0.857	0.046	3072	0.833	0.047	4140	0.763	0.047
			Eq.12 (s=0.3)	4958	0.815	0.045	3074	0.818	0.044	4775	0.811	0.044
			Eq.12 (s=1)	4958	0.838	0.044	3074	0.831	0.045	4775	0.827	0.043
			Eq.12 (s=2.5)	4958	0.833	0.045	3074	0.826	0.045	4775	0.819	0.045
			Eq.13 (s=1)	3030	0.606	0.040	2528	0.755	0.045	5660	0.884	0.046
			Eq.14 (s=1)	4763	0.826	0.043	2912	0.814	0.049	4575	0.808	0.040
0.1	0.2	1	Eq.3 (s=1)	409	0.865	0.043	248	0.847	0.050	340	0.787	0.041
			Eq.12 (s=0.3)	381	0.805	0.044	248	0.824	0.048	405	0.837	0.044
			Eq.12 (s=1)	381	0.844	0.044	248	0.848	0.049	405	0.850	0.046
			Eq.12 (s=2.5)	381	0.834	0.044	248	0.853	0.048	405	0.849	0.046
			Eq.13 (s=1)	221	0.584	0.041	190	0.742	0.045	440	0.881	0.047
			Eq.14 (s=1)	541	0.941	0.044	322	0.930	0.048	495	0.916	0.047

Empirical power calculated as in eTable 1. The data for RR=3 are shown. Eq.3=Eq. (3); Eq.12=Eq.(12); Eq.13=Eq.(13) converted from the equation by Cai and Zeng (2004); Eq.14=Eq.(14) converted from the equation by Kim et al. (2006);s=shape parameter in the Weibull distribution. *Note*; see footnotes to eTable 1 for an explanation for other abbreviations.

```
/***********************************************************************
* Simulation program for sample size for case-cohort design            *
*                                                                      *
* distribution for time to event data: Exponential                     *
*                                                                      *
*                                                          2009/09/30 *
***********************************************************************/

/* option ------------------------------------------------------------ */
  options nodate nocenter pageno=1 /* nonotes */;title;

  %macro new_sim_ex(lotnum);

/* Kubota and Wakana simple formula ---------------------------------- */
  data d2(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;
    n1=round(((probit(1-alpha/2)*sqrt((1+1/k)*pd*(1-pd))
      +probit(1-beta)*sqrt(p1*(1-p1)+p0*(1-p0)/k))**2)/((p0-p1)**2)*(1+1/m));
    n0=round(n1*k);n=n0+n1;
    formula=1;
  run;

/* Kubota and Wakana hypergeometric formula -------------------------- */
  data d3(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;q=m*pD;
    f0=(1-q)/(1-pd);f1=((k*rr+1)**2)*(1-q)/((k+rr)*(k*rr*(1-p1)+(1-p0)));
    n1=round(((probit(1-alpha/2)*sqrt((1+1/k)*pd*(1-pd)*(1+1/m*f0))
      +probit(1-beta)*sqrt((p1*(1-p1)+p0*(1-p0)/k)*(1+1/m*f1)))**2)/((p0-p1)**2));
    n0=round(n1*k);n=n0+n1;
    formula=2;
  run;

/* Cai and Zeng ------------------------------------------------------ */
```

```
  data d4(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;pr1=1/(1+k);pr2=k/(1+k);
    lamda1=-log(1-p1);lamda2=-log(1-p0);
    theta=log(lamda2/lamda1);
    b=((probit(1-alpha/2)+probit(1-beta))**2)/((theta**2)*pr1*pr2*pd);
    n1=round(b*(1+m*(1-pd))/(m*(1+k)));n0=round(n1*k);n=n0+n1;
    formula=3;
  run;

/* Kim et al. -------------------------------------------------------- */
  data d5(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;
    pec=1/(1+k);ped=(rr*pec)/(1+pec*(rr-1));pbar=(ped+m*(1-pd)*pec)/(1+m*(1-pd));
    ft=sqrt(pbar*(1-pbar)*(1+1/(m*(1-pd))));st=sqrt(ped*(1-ped)+pec*(1-pec)/(m*(1-pd)));
    n1=round(((probit(1-alpha/2)*sqrt(pbar*(1-pbar)*(1+1/(m*(1-pd))))
      +probit(1-beta)*sqrt(ped*(1-ped)+pec*(1-pec)/(m*(1-pd))))**2)/(pd*(ped-pec)**2*(1+k)));
    n0=round(n1*k);n=n0+n1;
    formula=4;
  run;

  data d6;set d2 d3 d4 d5;run;
  proc sort data=d6 out=d6;by type formula;run;
  proc means data=d6 noprint;var n;output out=d7 max=nall;run;
  data d8;set d7;call symput('nallmax',compress(nall));run;

/* subcohort members ------------------------------------------------- */
  data d9(keep=type formula lm ite id scr nite);set d6;
    lm=round((n1*p1+n0*p0)*m+.5,1);do ite=1 to nite;do id=1 to n;scr=ranuni(0);output;end;end;
  run;
  proc sort data=d9 out=d9;by type formula lm nite ite id;run;
  proc rank data=d9 out=d9;by type formula lm nite ite;var scr;ranks scr_r;run;
  data d9(keep=type formula nite lm ite id scr);set d9;by type formula nite lm ite id;
    if (.<scr_r<=lm) then scr=1;else scr=0;
```

```
  run;
  proc transpose data=d9 out=d10 prefix=sc;by type formula nite lm ite;var scr;run;
  data d11;merge d10 d6;by type formula;run;
  proc sort data=d11 out=d11;by type formula ite;run;

/* Exponential distribution ----------------------------------------- */
  data d12(keep=type formula nite ite event startime survtime lwgt id exps);set d11;by type formula ite;
    array sc{*} sc1-sc&nallmax;
    obs_t=1;lambda0=-log(1-p0)/obs_t;lambda1=-log(1-p1)/obs_t;
    exps=1;do id=1 to n1;
      x=ranuni(0);time=round(-log(1-x)/lambda1,.0001);
      if ((x<=p1)&(sc{id}=1)) then do;     event=1;startime=time-0.00005;survtime=time;        lwgt=0;  output;
                                           event=0;startime=0;           survtime=time-0.00005;lwgt=0;
output;end;
      else if ((x<=p1)&(sc{id}=0)) then do;event=1;startime=time-0.00005;survtime=time;        lwgt=-10;output;end;
      else if ((x>p1)&(sc{id}=1))  then do;event=0;startime=0;           survtime=9;           lwgt=0;  output;end;
    end;
    exps=0;do id=n1+1 to n;
      x=ranuni(0);time=round(-log(1-x)/lambda0,.0001);
      if ((x<=p0)&(sc{id}=1)) then do;     event=1;startime=time-0.00005;survtime=time;        lwgt=0;  output;
                                           event=0;startime=0;           survtime=time-0.00005;lwgt=0;
output;end;
      else if ((x<=p0)&(sc{id}=0)) then do;event=1;startime=time-0.00005;survtime=time;        lwgt=-10;output;end;
      else if ((x>p0)&(sc{id}=1))  then do;event=0;startime=0;           survtime=9;           lwgt=0;  output;end;
    end;
  run;

/* proportional hazard model (Self and Prentice) --------------------- */
  proc phreg data=d12 outest=d13 covsandwich(aggregate) covout noprint;by type formula ite;
    model (startime,survtime)*event(0)=exps /ties=breslow offset=lwgt;id id;run;
  data d13(keep=type formula ite est vari);set d13;by type formula ite;
    if (first.ite) then do;est=.;vari=.;end;retain est vari;
    select;
```

```
      when(lowcase(_type_)='parms') est=exps;
      when((lowcase(_type_)='cov')&(lowcase(_name_)='exps')) vari=exps;
      otherwise;
    end;
    if (last.ite) then output;
  run;
  data d14;merge d13 d6;by type formula;run;
  data d14(keep=type formula n0 n1 p0 p1 rr m alpha beta nite k lot ite sig est vari);set d14;by type formula ite;
    p=2*(1-probnorm(abs(est/sqrt(vari))));if (.<p<=alpha) then sig=1;else sig=0;lot=&lotnum;
  run;
  data d15;infile fname1;input type formula n0 n1 p0 p1 rr m alpha beta nite k lot ite sig est vari;run;
  data d16;set d15 d14;run;
  data _null_;set d16;
    file fname1 noprint;
    put @1 type 6.0 @11 formula 1.0 @13 n0 8.0 @23 n1 8.0 @33 p0 6.4 @40 p1 6.4 @47 rr 6.4 @54 m 2.0
        @57 alpha 5.3 @63 beta 5.3 @70 nite 7.0 @80 k 5.3 @86 lot 3.0 @90 ite 7.0 @99 sig 1.0 @101 est 8.3 @111 vari
8.3;
  run;

  proc datasets;delete d2-d15;run;quit;

  %mend new_sim_ex;

  %macro summary;

/* compute power mean percentile ------------------------------------ */
  proc sort data=d16 out=d16;by type formula n0 n1 p0 p1 rr m alpha beta k;run;
  proc univariate data=d16 noprint;by type formula n0 n1 p0 p1 rr m alpha beta k;
    var sig est vari;output out=d17 n=total_n xxx_n xxx_nn
                 mean=power est_m vari_m median=xxx est_md vari_md
                 pctls pctlpts=2.5 97.5 pctlpre=xxx_ est_ vari_;
  run;
  data d17;set d17;by type formula n0 n1 p0 p1 rr m alpha beta k;
```

```
    ex_est_m=exp(est_m);ex_est_md=exp(est_md);ex_est_2_5=exp(est_2_5);ex_est_97_5=exp(est_97_5);
  run;
  data _null_;set d17;
    file fname2 noprint;
    put @1 type 6.0 @11 formula 1.0 @13 n0 8.0 @23 n1 8.0 @33 p0 6.4 @40 p1 6.4 @47 rr 6.4 @54 m 2.0
        @57 alpha 5.3 @63 beta 5.3 @70 total_n 7.0 @80 k 5.3 @87 power 5.3 @93 ex_est_m 8.3 @102 ex_est_2_5 8.3
        @111 ex_est_md 8.3 @120 ex_est_97_5 8.3 @129 vari_m 8.3;
  run;

  proc datasets;delete d16-d17;run;quit;

  %mend summary;

/* output file ------------------------------------------------------- */
  filename fname1 'C:¥sp_each_ex.out';
  filename fname2 'C:¥sp_power_ex.out';

/* Note: there will be an error message on the log window at the first
 running of this SAS code since the output file is appended to an
 existing file which has the same file name. If the file is not prepared,
 a new file is created. You can disregard the error message.           */

/* input simulation condition parameters ----------------------------- */
 data d1;
    input type k m p0 rr alpha beta nite;
    cards;
       1  4  1 0.001 3 0.05 0.2 1000
       2  4  5 0.001 3 0.05 0.2 1000
       3  4  1 0.01  3 0.05 0.2 1000
       4  4  5 0.01  3 0.05 0.2 1000
       5  4  1 0.1   3 0.05 0.2 1000
       6  4  3 0.1   3 0.05 0.2 1000
    ;
```

```
run;

%new_sim_ex(1);
%summary;
```

```
/**********************************************************************
* Simulation program for sample size for case-cohort design           *
*                                                                     *
* distribution for time to event data: Weibull                        *
*                                                                     *
*                                                          2009/09/29 *
**********************************************************************/

/* option ------------------------------------------------------------ */
  options nodate nocenter pageno=1 /* nonotes */;title;

/* macro ------------------------------------------------------------- */

  %macro new_sim_we(lotnum);

/* Kubota and Wakana simple formula ---------------------------------- */
  data d2(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite gamma);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;
    n1=round(((probit(1-alpha/2)*sqrt((1+1/k)*pd*(1-pd))
      +probit(1-beta)*sqrt(p1*(1-p1)+p0*(1-p0)/k))**2)/((p0-p1)**2)*(1+1/m));
    n0=round(n1*k);n=n0+n1;
    formula=1;
  run;

/* Kubota and Wakana hypergeometric formula -------------------------- */
  data d3(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite gamma);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;q=m*pD;
    f0=(1-q)/(1-pd);f1=((k*rr+1)**2)*(1-q)/((k+rr)*(k*rr*(1-p1)+(1-p0)));
    n1=round(((probit(1-alpha/2)*sqrt((1+1/k)*pd*(1-pd)*(1+1/m*f0))
      +probit(1-beta)*sqrt((p1*(1-p1)+p0*(1-p0)/k)*(1+1/m*f1)))**2)/((p0-p1)**2));
    n0=round(n1*k);n=n0+n1;
    formula=2;
  run;
```

```
/* Cai and Zeng ------------------------------------------------------- */
  data d4(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite gamma);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;pr1=1/(1+k);pr2=k/(1+k);
    lamda1=-log(1-p1);lamda2=-log(1-p0);theta=log(lamda2/lamda1);
    b=((probit(1-alpha/2)+probit(1-beta))**2)/((theta**2)*pr1*pr2*pd);
    n1=round(b*(1+m*(1-pd))/(m*(1+k)));n0=round(n1*k);n=n0+n1;
    formula=3;
  run;

/* Kim et al. --------------------------------------------------------- */
  data d5(keep=type formula n0 n1 n k m p0 p1 rr alpha beta nite gamma);set d1;
    pd=p0*(rr+k)/(1+k);p1=p0*rr;pec=1/(1+k);ped=(rr*pec)/(1+pec*(rr-1));
    pbar=(ped+m*(1-pd)*pec)/(1+m*(1-pd));ft=sqrt(pbar*(1-pbar)*(1+1/(m*(1-pd))));
    st=sqrt(ped*(1-ped)+pec*(1-pec)/(m*(1-pd)));
    n1=round(((probit(1-alpha/2)*sqrt(pbar*(1-pbar)*(1+1/(m*(1-pd))))
      +probit(1-beta)*sqrt(ped*(1-ped)+pec*(1-pec)/(m*(1-pd))))**2)/(pd*(ped-pec)**2*(1+k)));
    n0=round(n1*k);n=n0+n1;
    formula=4;
  run;

  data d6;set d2 d3 d4 d5;run;proc sort data=d6 out=d6;by type formula;run;
  proc means data=d6 noprint;var n;output out=d7 max=nall;run;
  data d8;set d7;call symput('nallmax',compress(nall));run;

/* subcohort members -------------------------------------------------- */
  data d9(keep=type formula lm ite id scr nite);set d6;
    lm=round((n1*p1+n0*p0)*m+.5,1);do ite=1 to nite;do id=1 to n;scr=ranuni(0);output;end;end;
  run;
  proc sort data=d9 out=d9;by type formula lm nite ite id;run;
  proc rank data=d9 out=d9;by type formula lm nite ite;var scr;ranks scr_r;run;
  data d9(keep=type formula nite lm ite id scr);set d9;by type formula nite lm ite id;
    if (.<scr_r<=lm) then scr=1;else scr=0;
```

```
  run;
  proc transpose data=d9 out=d10 prefix=sc;by type formula nite lm ite;var scr;run;
  data d11;merge d10 d6;by type formula;run;
  proc sort data=d11 out=d11;by type formula ite;run;

/* Weibull distribution ---------------------------------------------- */
  data d12(keep=type formula nite ite event startime survtime lwgt id exps);set d11;by type formula ite;
    array sc{*} sc1-sc&nallmax;
    obs_t=1;lambda0=(-log(1-p0)/obs_t)**(1/gamma);lambda1=(-log(1-p1)/obs_t)**(1/gamma);
    exps=1;do id=1 to n1;
      x=ranuni(0);time=round(((-log(1-x))**(1/gamma))/lambda1,.0001);
      if ((x<=p1)&(sc{id}=1)) then do;     event=1;startime=time-0.00005;survtime=time;        lwgt=0;  output;
                                           event=0;startime=0;          survtime=time-0.00005;lwgt=0;
output;end;
      else if ((x<=p1)&(sc{id}=0)) then do;event=1;startime=time-0.00005;survtime=time;        lwgt=-10;output;end;
      else if ((x>p1)&(sc{id}=1))  then do;event=0;startime=0;          survtime=9;           lwgt=0;  output;end;
    end;
    exps=0;do id=n1+1 to n;
      x=ranuni(0);time=round(((-log(1-x))**(1/gamma))/lambda0,.0001);
      if ((x<=p0)&(sc{id}=1)) then do;     event=1;startime=time-0.00005;survtime=time;        lwgt=0;  output;
                                           event=0;startime=0;          survtime=time-0.00005;lwgt=0;
output;end;
      else if ((x<=p0)&(sc{id}=0)) then do;event=1;startime=time-0.00005;survtime=time;        lwgt=-10;output;end;
      else if ((x>p0)&(sc{id}=1))  then do;event=0;startime=0;          survtime=9;           lwgt=0;  output;end;
    end;
  run;

/* proportional hazard model (Self and Prentice) --------------------- */
  proc phreg data=d12 outest=d13 covsandwich(aggregate) covout noprint;by type formula ite;
    model (startime,survtime)*event(0)=exps /ties=breslow offset=lwgt;id id;run;
  data d13(keep=type formula ite est vari);set d13;by type formula ite;
    if (first.ite) then do;est=.;vari=.;end;retain est vari;
    select;
```

```
      when(lowcase(_type_)='parms') est=exps;
      when((lowcase(_type_)='cov')&(lowcase(_name_)='exps')) vari=exps;
      otherwise;
    end;
    if (last.ite) then output;
  run;
  data d14;merge d13 d6;by type formula;run;
  data d14(keep=type formula n0 n1 p0 p1 rr m alpha beta nite k lot ite sig est vari gamma);set d14;by type formula
ite;
    p=2*(1-probnorm(abs(est/sqrt(vari))));
    if (.<p<=alpha) then sig=1;else sig=0;
    lot=&lotnum;
  run;
  data d15;infile fname1;input type formula n0 n1 p0 p1 rr m alpha beta nite k lot ite sig est vari gamma;run;
  data d16;set d15 d14;run;
  data _null_;set d16;
    file fname1 noprint;
    put @1 type 6.0 @11 formula 1.0 @13 n0 8.0 @23 n1 8.0 @33 p0 6.4 @40 p1 6.4 @47 rr 6.4 @54 m 2.0
        @57 alpha 5.3 @63 beta 5.3 @70 nite 7.0 @80 k 5.3 @86 lot 3.0 @90 ite 7.0 @99 sig 1.0
        @101 est 8.3 @111 vari 8.3 @121 gamma 5.3;
  run;

  proc datasets;delete d2-d15;run;quit;

  %mend new_sim_we;

  %macro summary;

/* compute power mean percentile ------------------------------------- */
  proc sort data=d16 out=d16;by type formula n0 n1 p0 p1 rr m alpha beta k gamma;run;
  proc univariate data=d16 noprint;by type formula n0 n1 p0 p1 rr m alpha beta k gamma;
    var sig est vari;output out=d17 n=total_n xxx_n xxx_nn
               mean=power est_m vari_m median=xxx est_md vari_md
```

```
                 pctls pctlpts=2.5 97.5 pctlpre=xxx_ est_ vari_;
  run;
  data d17;set d17;by type formula n0 n1 p0 p1 rr m alpha beta k gamma;
    ex_est_m=exp(est_m);ex_est_md=exp(est_md);ex_est_2_5=exp(est_2_5);ex_est_97_5=exp(est_97_5);
  run;
  data _null_;set d17;
    file fname2 noprint;
    put @1 type 6.0 @11 formula 1.0 @13 n0 8.0 @23 n1 8.0 @33 p0 6.4 @40 p1 6.4 @47 rr 6.4 @54 m 2.0 @57 alpha 5.3
@63 beta 5.3
        @70 total_n 7.0 @80 k 5.3 @87 power 5.3 @93 ex_est_m 8.3 @102 ex_est_2_5 8.3 @111 ex_est_md 8.3 @120 ex_est_97_5
8.3
        @129 vari_m 8.3 @138 gamma 5.3;
  run;
  proc datasets;delete d16-d17;run;quit;

  %mend summary;

/* output file -------------------------------------------------------- */
  filename fname1 'C:¥sp_each_we.out';
  filename fname2 'C:¥sp_power_we.out';

/* Note: there will be an error message on the log window at the first
 running of this SAS code since the output file is appended to an
 existing file which has the same file name. If the file is not prepared,
 a new file is created. You can disregard the error message.           */

/* input simulation condition parameters ----------------------------- */
  data d1;
    input type k m p0 rr alpha beta nite gamma;
    cards;
       1 4    1 0.001 3 0.05 0.2 1000 0.3
       2 4    1 0.001 3 0.05 0.2 1000 1
       3 4    1 0.001 3 0.05 0.2 1000 2.5
```

```
    4 1    1 0.001 3 0.05 0.2 1000 0.3
    5 1    1 0.001 3 0.05 0.2 1000 1
    6 1    1 0.001 3 0.05 0.2 1000 2.5
    7 0.25 1 0.001 3 0.05 0.2 1000 0.3
    8 0.25 1 0.001 3 0.05 0.2 1000 1
    9 0.25 1 0.001 3 0.05 0.2 1000 2.5
  ;
run;

%new_sim_we(1);
%summary;
```