American Journal of Epidemiology
Published by the Johns Hopkins Bloomberg School of Public Health 2007.

Vol. 166, No. 4
DOI: 10.1093/aje/kwm097
Advance Access publication June 7, 2007

Original Contribution

Using Risk-based Sampling to Enrich Cohorts for Endpoints, Genes, and Exposures

Clarice R. Weinberg[1], David L. Shore[2], David M. Umbach[1], and Dale P. Sandler[3]

[1] Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC.
[2] Westat, Durham, NC.
[3] Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Received for publication August 17, 2006; accepted for publication February 21, 2007.

Targeting the first-degree relatives of people with a particular complex disease can offer a powerful approach to building a risk-based cohort for prospective studies of etiologic factors. Such a cohort provides both a sizable increase in the rate of accrual of newly incident cases, enriching for risk factors that are known or even unknown, and a high level of motivation among participants. A nationwide study of breast cancer in the United States and Puerto Rico, the Sister Study, made up of women who are each the sister of a woman with breast cancer, exemplifies this approach. In this paper, the authors provide power calculations to aid in the design of such studies and quantify their benefits for detecting both genetic variants related to risk and interactive effects of genetic and environmental factors. While the risk-based cohort can have markedly increased prevalences of rare causative alleles, most of the power advantages for this design is due to the increased rate of accrual of newly incident cases rather than the increase in any one individual allele.

cohort studies; disease susceptibility; genes; prospective studies; risk; sampling studies

Most diseases are complex because both genetic and environmental factors contribute to their etiology, and many are difficult to study prospectively because they are uncommon. Retrospective case-control studies offer important advantages for rare diseases, but enrolling individuals only after disease has occurred can distort findings and prevent the collection of information that could shed light on complex etiology. Thus, strategies for enhancing the power of prospective studies are needed. Risk-based cohorts can be constructed either by enrolling first-degree relatives of probands who have experienced a particular adverse event, such as breast cancer, or by enrolling people at risk of recurrence of an adverse event. The latter idea was applied (1) in a study of folate supplementation in pregnant women who had previously delivered a child with a neural tube defect. To improve power to detect genetic contributions to risk, Antoniou and Easton (2) also suggested a case-control design in which cases who have a family history of the disease are selected.

We here consider a risk-based sampling design in which siblings of affected individuals are recruited and followed prospectively. Such a design provides well-motivated volunteers and, to the extent that risk follows a familial pattern, an accrual rate for new cases that is elevated compared with a random cohort. For example, for breast cancer, the elevation in incidence among sisters is about twofold (3). For certain other conditions, such as autism (4), the relative increase could be more substantial. A risk-based cohort will have increased prevalence of both risk-related genetic variants and exposures that tend to co-occur within families, enhancing power to identify such factors.

In two-stage case-control sampling (5–8), one can oversample people with particular risk profiles by using prespecified outcome-and-covariate-dependent recruitment probabilities. Risk-based sampling is similar in that susceptible people are overrepresented but different in that the prevalence of both known and yet-unknown genetic and environmental risk factors is increased.

Correspondence to Dr. Clarice R. Weinberg, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, NC 27709 (e-mail: weinberg@niehs.nih.gov).

Another, less intuitive consequence of risk-based sampling is a tilting of measurable exposure relative risks away from the null for any exposure that interacts supermultiplicatively with an uncommon genetic variant. This augmentation and the corresponding increase in statistical power happen even when the genetic cofactor is unknown. While this exaggeration of the measurable exposure effect in the presence of supermultiplicative interaction happens in both the source population and a risk-based cohort, it is enhanced in the latter because the genetic cofactor is more prevalent.

This paper explores effects of risk-based sampling on power through these several mechanisms. Our results should aid investigators in determining required sample sizes under a range of relative risk and gene-by-environment interaction scenarios. We illustrate benefits of risk-based sampling by using as an example the Sister Study (http://www.sisterstudy.org), which is recruiting a cohort of sisters of women with breast cancer.

Conceptually, risk for the sibling of a proband case is elevated in part because both share the same parents and hence will carry many of the same genetic variants. The offspring of a proband case would also carry many of the same genetic variants and be at increased risk, so a ''son or daughter study'' would offer similar advantages, although it could be subject to time trends in risk-related exposures. As a third variant, sampling based on the previous occurrence of events such as cancer or pregnancy complications can be beneficial but needs separate consideration because the same person is studied at a later time point. For cancer, in particular, treatment of the original disease can influence recurrence. Here, we limit our attention to designs in which enrichment is achieved by enrolling siblings of affected individuals. We return briefly to other designs in the Discussion section of the paper.

CALCULATING EFFECTS OF RISK-BASED SAMPLING

Calculations for a sibling cohort start with the parents, even though the parents are not studied. The family's genotype for a diallelic locus can be represented by the number of copies of the variant allele carried by the mother, the father, and the affected proband, denoted by M, F, and P, respectively. Note that it is inconsequential which allele is considered the ''variant,'' although geneticists usually take ''variant'' to be the less common of the two. Table 1 shows the 15 possible MFP triads. To simplify calculations, we assume Hardy-Weinberg equilibrium in the source population. Thus, if p is the prevalence of the variant allele, the proportion of individuals carrying two copies is p^2 and the proportion carrying one is $2p(1-p)$. We also assume that the parental genotypes are symmetric, so that, for example, the combination M = 2 and F = 1 is as likely as M = 1 and F = 2. Finally, we assume that the parents' genotypes are not related to the number of surviving children they now have.

Because their offspring have been identified as having developed the disease under study, the genotype distribution among parents is enriched for risk-related genetic variants. In column 6 of table 1, R_1 and R_2 denote the relative risks for offspring with one or with two copies of the variant allele, respectively, relative to those with no copies.

TABLE 1. Quantifying the effect of sampling siblings of affected individuals*

M	F	P	Population Pr[MF]	Population Pr[offspring | MF]	Pr[MFP triad]	Exposure factor	Pr[Sibling has 0 copies, given the parents]	Pr[Sibling has 1 copy, given the parents]	Pr[Sibling has 2 copies, given the parents]
2	2	2	p^4	1	R_2p^4B	$1-r+rTU_2$	0	0	1
2	1	2	$2p^3(1-p)$	1/2	$R_2p^3(1-p)$B	$1-r+rTU_2$	0	1/2	1/2
2	1	1	$2p^3(1-p)$	1/2	$R_1p^3(1-p)$B	$1-r+rTU_1$	0	1/2	1/2
1	2	2	$2p^3(1-p)$	1/2	$R_2p^3(1-p)$B	$1-r+rTU_2$	0	1/2	1/2
1	2	1	$2p^3(1-p)$	1/2	$R_1p^3(1-p)$B	$1-r+rTU_1$	0	1/2	1/2
1	1	0	$4p^2(1-p)^2$	1/4	$p^2(1-p)^2$B	$1-r+rT$	1/4	1/2	1/4
1	1	1	$4p^2(1-p)^2$	1/2	$2\,R_1p^2(1-p)^2$B	$1-r+rTU_1$	1/4	1/2	1/4
1	1	2	$4p^2(1-p)^2$	1/4	$R_2p^2(1-p)^2$B	$1-r+rTU_2$	1/4	1/2	1/4
1	0	0	$2p(1-p)^3$	1/2	$p(1-p)^3$B	$1-r+rT$	1/2	1/2	0
1	0	1	$2p(1-p)^3$	1/2	$R_1p(1-p)^3$B	$1-r+rTU_1$	1/2	1/2	0
0	1	0	$2p(1-p)^3$	1/2	$p(1-p)^3$B	$1-r+rT$	1/2	1/2	0
0	1	1	$2p(1-p)^3$	1/2	$R_1p(1-p)^3$B	$1-r+rTU_1$	1/2	1/2	0
2	0	1	$p^2(1-p)^2$	1	$R_1p^2(1-p)^2$B	$1-r+rTU_1$	0	1	0
0	2	1	$p^2(1-p)^2$	1	$R_1p^2(1-p)^2$B	$1-r+rTU_1$	0	1	0
0	0	0	$(1-p)^4$	1	$(1-p)^4$B	$1-r+rT$	1	0	0

* For a diallelic gene, genetic enrichment when siblings are studied can be calculated from factors given here. Simplifying assumptions are that the gene is in Hardy-Weinberg equilibrium and that exposure and gene occur independently in the population. Probabilities are based on a multiplicative model; R_1 and R_2 are the relative risks for those with one and two copies, respectively. Allele prevalence is p, exposure prevalence is r, the main effect of the exposure is T, and the two interaction parameters are U_1 and U_2. B is simply a normalizing constant, ensuring that probabilities sum to 1.0. M, mother; F, father; P, affected proband.

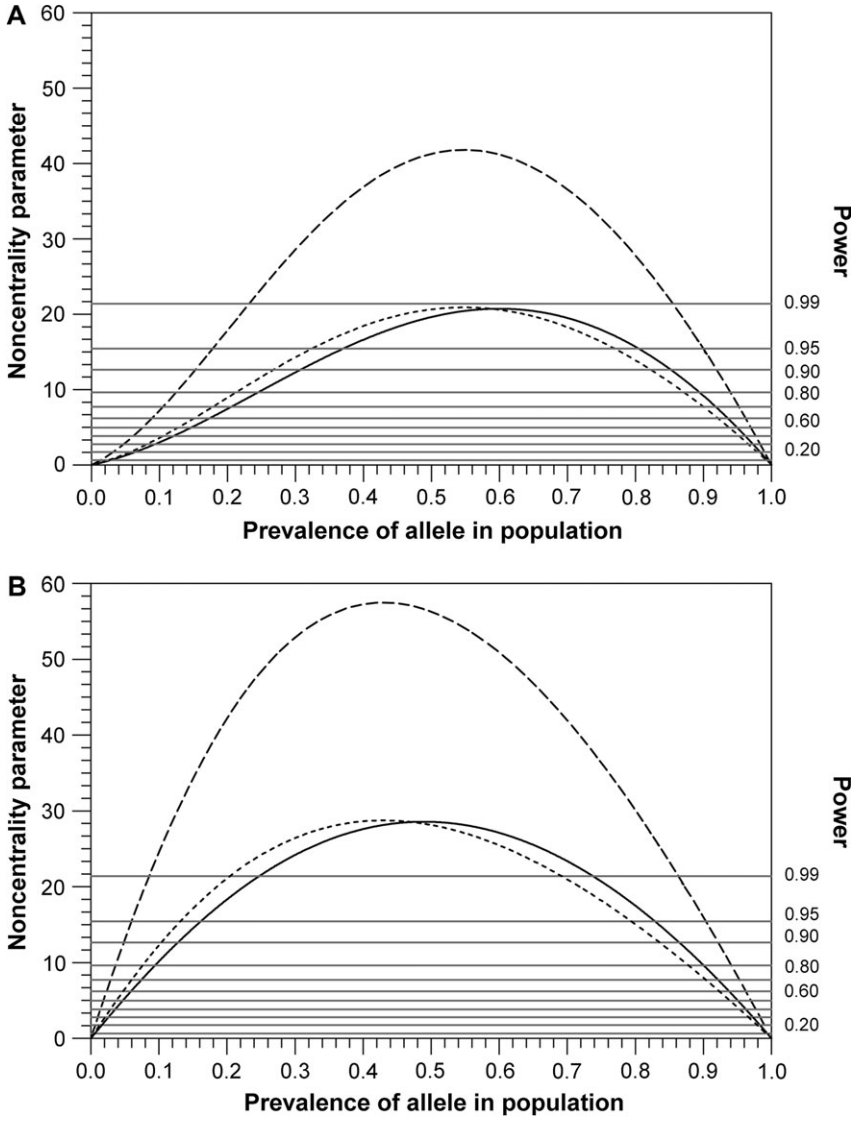

FIGURE 1. Noncentrality parameter (left *y*-axis) and associated power (right *y*-axis) as a function of allele prevalence for a 2-degrees-of-freedom chi-square likelihood ratio test for genetic main effects. Calculations assume a nested case-control study with 400 cases and 800 controls within a risk-based cohort (short-dash line) or a randomly sampled cohort (solid line). An example curve corresponding to doubled accrual, based on 800 cases and 1,600 controls, within the risk-based cohort, is also shown (long-dash line). The genetic risk scenarios are A) $R_1 = 1.2$, $R_2 = 2.0$, $T = U_1 = U_2 = 1.0$; and B) $R_1 = 1.5$, $R_2 = 2.5$, $T = U_1 = U_2 = 1.0$. Parameters R_1 and R_2 denote the relative risks for offspring with one and with two copies of the variant allele, respectively, relative to those with no copies. The parameter T is the relative risk associated with exposure. U_1 and U_2 are the multiplicative interaction parameters for exposure acting jointly with one and with two inherited copies of the allele, respectively.

If the genetic variant interacts with an exposure, another multiplier will be needed (9), as shown for a dichotomous exposure, E, with population prevalence r in column 7 of table 1. To simplify calculations, E is assumed to occur independently of genotype in the source population. (The analysis relies on only the much weaker assumption that E in offspring is independent of allele transmission.) The parameter T is the relative risk associated with E. U_1 and U_2 are the multiplicative interaction parameters for E acting jointly with one and two inherited copies of the allele, respectively. If the interaction parameters are both 1, then exposure has no effect on the distribution of parental genotypes: the multiplicative exposure effect is simply absorbed into the normalizing constant, B. The final three columns of the table give the Mendelian probabilities for the unaffected sibling, conditional on the genotypes of the parents.

One can use the elements of table 1 to calculate the genotype distribution for the siblings of probands, that is, for the risk-based cohort, under any postulated scenario characterized by the frequency of the variant allele, the genotype relative risks, the exposure prevalence, the exposure relative risk, and the multiplicative gene-by-exposure interaction

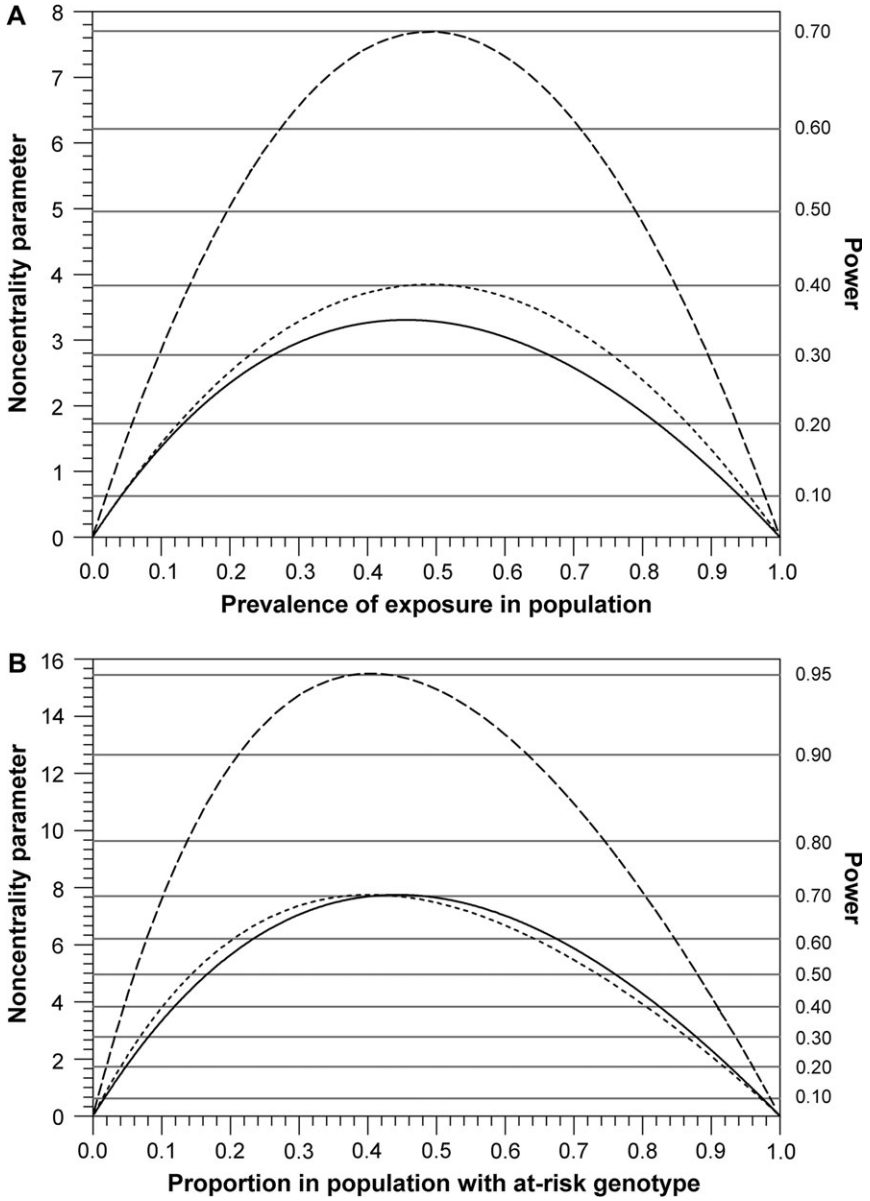

FIGURE 2. Noncentrality parameter (left *y*-axis) and associated power (right *y*-axis) for a 2-degrees-of-freedom chi-square likelihood ratio test for interaction. The risk scenario is for a pure interaction: $R_1 = R_2 = T = 1.0$, $U_1 = U_2 = 2.0$. Calculations assume a nested case-control study with 400 cases and 800 controls within a risk-based cohort (short-dash line) or a randomly sampled cohort (solid line). An example curve corresponding to doubled accrual within the risk-based cohort, based on 800 cases and 1,600 controls, is also shown (long-dash line). Noncentrality is plotted as a function of A) exposure prevalence, with allele prevalence fixed at 0.05; and B) proportion of the population that carries the allele, with exposure prevalence fixed at 0.4. Parameters R_1 and R_2 denote the relative risks for offspring with one and with two copies of the variant allele, respectively, relative to those with no copies. The parameter T is the relative risk associated with exposure. U_1 and U_2 are the multiplicative interaction parameters for exposure acting jointly with one and with two inherited copies of the allele, respectively.

parameters. This distribution, together with relative accrual rates, enables calculation of power.

CALCULATING POWER

We first assume no exposure effects or interactions. We presume that after a specified length of follow-up, a nested case-control analysis is carried out. We calculate the power of the corresponding 2-degrees-of-freedom likelihood ratio chi-square test of the null hypothesis that both genotype relative risks R_1 and R_2 are equal to 1.

The power is derived as follows (10). First, calculate the expected counts in the 2×3 table (case-control status by number of alleles) for a case-control sample nested within

the risk-based cohort, using table 1. The probability that a control carries zero, or one, or two copies of the variant is given by summing products of the triad probabilities (column 6) and the respective sibling probabilities (columns 8, 9, or 10). This calculation implicitly uses a rare-disease assumption, by assuming that the genotype distribution in controls is that of participants in general. The three case genotype probabilities then depend on the assumed relative risks applied to those control genotype probabilities. The expected counts are the specified totals for cases or controls multiplied by the corresponding probabilities. One calculates the change in deviance (the negative of twice the log maximized likelihood) based on applying logistic regression to the expected counts in the 2×3 table. This number is the *noncentrality parameter* for a 2-degrees-of-freedom chi-square distribution. This parameter characterizes the right-skewing of the noncentral chi-square distribution under the specified alternative to the null and determines the power. Widely available statistical software (e.g., SAS (Statistical Analysis System; SAS Institute, Inc., Cary, North Carolina), GAUSS Mathematical and Statistical System (Aptech Systems, Inc., Black Diamond, Washington)) provides power for a given noncentrality parameter by computing the probability in the upper tail of the noncentral chi-square distribution. For comparison, we also calculate the noncentrality parameter for the same case-control study nested instead within a randomly sampled cohort.

We focus on noncentrality parameters because, for a given case-control ratio and risk scenario, one can use the noncentrality parameter calculated for one sample size to easily calculate the noncentrality parameter (and power) for a different sample size: multiply the calculated noncentrality parameter by the ratio of the desired sample size divided by the sample size used in the calculation. For simplicity, we assume twice as many controls as cases, but the reader can modify this ratio by redoing the calculation.

Initially, we assume the same number of cases in the risk-based cohort as in a randomly sampled cohort, although of course there would typically be a much higher number in the risk-based cohort because increased incidence leads to increased accrual. Our artificial equalization allows us to look separately at the gain in power due to increasing the prevalence of risk-related alleles in the risk-based cohort as distinct from the additional improvement that results from the obvious increase in achievable sample size.

Power to detect an effect of genotype

For two risk scenarios, we computed the noncentrality parameters and powers for an α-level 0.05 test from a study of 400 cases and 800 controls within either a risk-based cohort or a randomly sampled cohort. In figure 1A, the relative risks for one and two copies of the gene are assumed to be 1.2 and 2.0, respectively. The lower of the two dashed curves accounts for only genetic enrichment and not for increased accrual in a risk-based cohort; the higher dashed curve shows the parameters that also reflect the doubled accrual expected when the relative risk associated with first-degree family history is 2, as for breast cancer. For this curve, the calculations assume 800 cases and 1,600 controls. For other applications, the relative risk associated with having an affected first-degree relative could be higher or lower than 2, and values (for a study that also has a 2:1 ratio of controls to cases) can be found by multiplication. Power is shown on the right-hand *y*-axis. Figure 1B shows results with slightly larger relative risks.

The risk-based cohort always has a higher allele prevalence when the relative risks exceed 1.0; consequently, as the population allele prevalence becomes large enough, if we do not account for increased accrual, the randomly sampled cohort slightly outperforms the risk-based cohort. Since the prevalence of a risk-enhancing allele in the risk-based cohort always exceeds that in the corresponding population, the risk-based cohort reaches the point of diminishing power advantage from increasing prevalence sooner (thus, the curves cross). However, when we account for increased accrual, the benefit remains substantial.

Power to detect genotype-by-exposure interaction

The above calculations assumed no effect of exposure. We now consider a pure interaction scenario, where E doubles risk but only for those who carry one or two copies of the variant allele, and genotype does not affect risk in the absence of exposure. Figure 2A shows results based on comparing a saturated model for joint effects with a model with just the multiplicative main effects, across the range of exposure prevalences when allele prevalence is 0.05. The noncentrality parameter is plotted as a function of the proportion of the population carrying the allele. When the exposure is fairly common, the power to detect this interaction is modestly improved by the risk-based sampling through the increased allele prevalence alone. The increase in case accrual, however, increases power further, and one can generally achieve reasonable power to detect an interaction of this magnitude provided the allele prevalence is not close to 0 or to 1. Figure 2B shows the same model comparison as a function of the proportion of the population carrying the allele when exposure prevalence is 0.4.

Figures 3A and 3B show powers for detecting interaction when there is also a main effect of genotype, with relative risks of 1.5 and 3.0 for one copy and two copies, respectively. Figure 3A shows that for not very common risk alleles, both enrichment and increased accrual confer benefit. Interestingly, for more common alleles, the increase in allele prevalence conferred by risk-based sampling can actually limit the increase in power by making an already common allele too common in the cohort. However, this loss is more than made up for by increased case accrual. Figure 3B shows the power results for tests based on an additive null, testing against the same alternative. Note that the specified alternative is supermultiplicative and that the power based on the additive null is superior.

Our calculations assumed that the exposures of the proband and of the enrolled sibling are uncorrelated. If they were positively correlated and exposure prevalence were low, then the power enhancement would be increased to the extent that the exposure prevalence would be higher in the risk-based cohort (11).

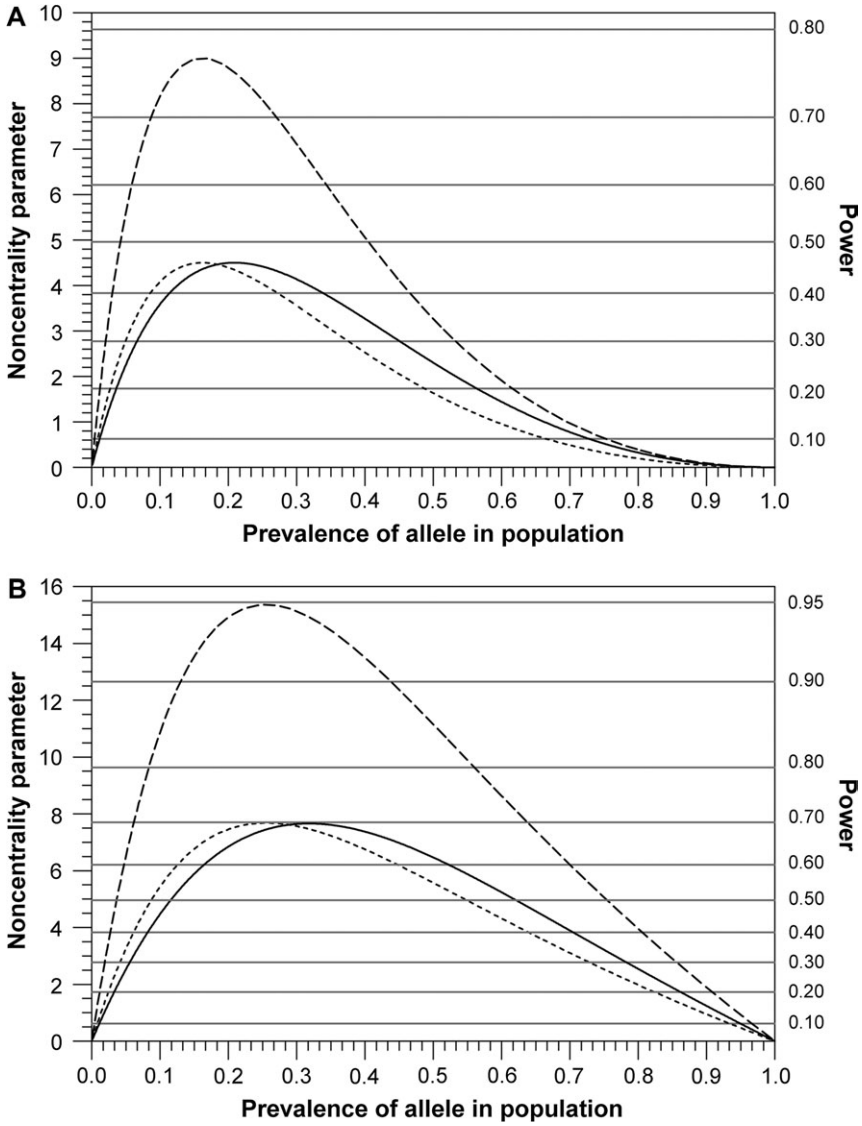

FIGURE 3. Noncentrality parameter (left *y*-axis) and associated power (right *y*-axis) as a function of allele prevalence for a 2-degrees-of-freedom chi-square likelihood ratio test for interaction effects. The interaction risk scenario: $R_1 = 1.5$, $R_2 = 3.0$, $T = 1$, $U_1 = U_2 = 2.0$, with exposure prevalence fixed at 0.15. Calculations assume a nested case-control study with 400 cases and 800 controls within a risk-based cohort (short-dash line) or a randomly sampled cohort (solid line). An example curve corresponding to doubled accrual within the risk-based cohort, based on 800 cases and 1,600 controls, is also shown (long-dash line). Noncentrality is based on A) a multiplicative null model for no interaction, and B) an additive null model for no interaction. Parameters R_1 and R_2 denote the relative risks for offspring with one and with two copies of the variant allele, respectively, relative to those with no copies. The parameter T is the relative risk associated with exposure. U_1 and U_2 are the multiplicative interaction parameters for exposure acting jointly with one and with two inherited copies of the allele, respectively.

Power to detect an exposure that interacts with an unstudied gene

Suppose that the exposure relative risk for those with no copies of the variant allele is 1.5 and the interaction with one or two copies of the variant allele is also twofold ($U_1 = U_2 = 2$), so that those with both have a threefold increase in risk. The gene itself has no effect in the absence of exposure ($R_1 = R_2 = 1$). Suppose, however, that we did not know about this gene and have not studied it. Its variant allele has a frequency of 0.2. The relative risk for the exposure is a weighted average of that among carriers and that among noncarriers, being about 2.04 in the hypothetical randomly sampled cohort and slightly larger, 2.16 at most (across exposure prevalences), in the risk-based cohort. Figure 4 shows the results across all possible exposure prevalences. The large power benefit for the risk-based cohort is almost entirely due to increased accrual. The effect of increasing

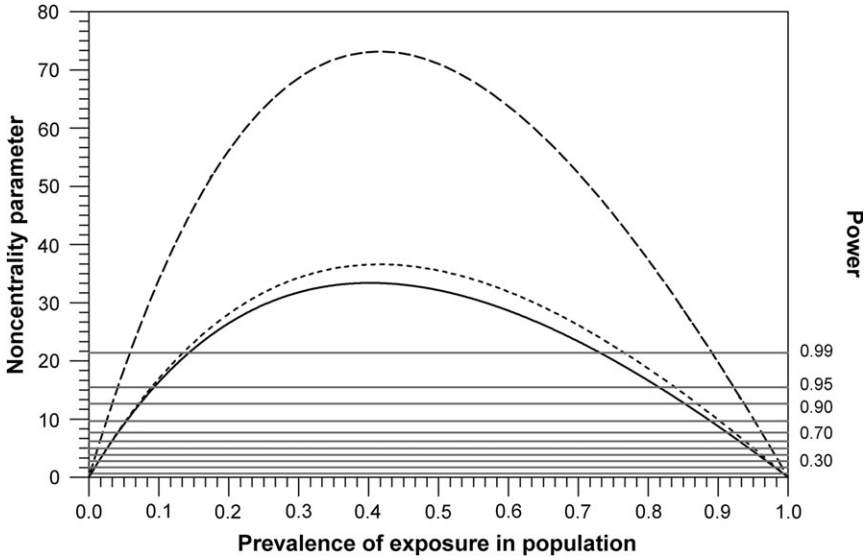

FIGURE 4. Noncentrality parameter (left *y*-axis) and associated power (right *y*-axis) as a function of exposure prevalence for a test of exposure main effects, showing the influence of unstudied genetic effects. Risk scenario: $R_1 = R_2 = 1.0$, $T = 1.5$, $U_1 = U_2 = 2.0$, with allele prevalence fixed at 0.2. Calculations assume a nested case-control study with 400 cases and 800 controls within a risk-based cohort (short-dash line) or a randomly sampled cohort (solid line). An example curve corresponding to doubled accrual within the risk-based cohort, based on 800 cases and 1,600 controls, is also shown (long-dash line). Parameters R_1 and R_2 denote the relative risks for offspring with one and with two copies of the variant allele, respectively, relative to those with no copies. The parameter T is the relative risk associated with exposure. U_1 and U_2 are the multiplicative interaction parameters for exposure acting jointly with one and with two inherited copies of the allele, respectively.

the prevalence of the at-risk genotype is a linear increase in the noncentrality parameter, with again little effect of enrichment per se (data not shown).

EXAMPLE: THE SISTER STUDY

Risk-based sampling is being used for the Sister Study, a prospective study of environmental and genetic contributors to risk of breast cancer (http://www.sisterstudy.org). Its prospective design addresses concerns about interpreting causal associations with environmental exposures in a retrospective study. For example, if assessing effects of pesticides or vitamin D on breast cancer risk, one would worry that breast cancer or its treatment would alter measured levels of specific metabolites or would modify behaviors, leading to bias with a retrospective design. However, a prospective study of the general population would need to be very large or long term to yield the number of cases needed to carry out studies of gene-environment interactions for relatively rare genes or exposures.

Taking advantage of the fact that sisters of women with breast cancer have, on average, a doubled risk of developing breast cancer, the Sister Study is recruiting 50,000 women aged 35–74 years, each of whom had a sister with breast cancer. The index sister is not enrolled or genotyped. Not only will twice as many sisters develop breast cancer during the course of follow-up as would be seen in a random sample (with 300 cases per year expected among the 50,000 sisters) but also more of them will have genetic factors related to breast cancer risk. To the extent that sisters also share lifestyles and early life exposures (a feature our calculations have not exploited), more of them may also have exposures related to increased breast cancer risk. The study recruits sisters, rather than mothers or daughters, because they are likely to be in the same age group as the index sister who developed breast cancer and thus in the age group of greatest breast cancer risk. Furthermore, the similarity in ages for siblings will increase the likelihood of sharing common exposures and lifestyle. An added advantage of this design is that many sisters of women with breast cancer are highly motivated because of their family history.

DISCUSSION

Risk-based sampling can greatly improve the efficiency of cohort studies for assessing the contribution of uncommon genetic and environmental factors to risk of diseases that are not extremely rare and that have an etiology that is in part genetic. When appropriately targeted, such strategies will likely also improve compliance and data quality. Our power calculations document that the benefit of risk-based sampling is attributable to increasing the prevalence of uncommon risk alleles as well as to the expected accrual of additional cases, the latter being the more important contributor. Nonetheless, the increased accrual is likely the result of the combined impact of increases in the prevalence of many rare risk alleles and exposures.

Although we have documented substantial power gains, these gains are most modest for causative genetic variants

that are very common (refer to the figures) or, equivalently, for protective genetic variants that are uncommon. However, causative variants may often have a prevalence in the range of 10 percent or less, where achieving adequate power will be the investigator's greatest challenge, especially for detection of gene-by-environment interactions.

Closely related designs have been used to study the risk of events associated with an elevated recurrence risk, such as pregnancy complications. For conditions expressed early in life, maternally mediated genetic effects may be important (12) because the mother's genotype can act on her phenotype during pregnancy and thereby influence her phenotype and that of her offspring. Such maternal influences may even be important for the development of later diseases, such as schizophrenia or breast cancer (13). In the presence of such a mechanism, use of these designs will again enrich the parental genotype distribution for causative genes compared with what would be seen with random sampling, potentially reflecting both maternal and offspring genotype effects.

To the extent that maternally mediated effects are biologically plausible, if the mothers of cohort (or case/control) participants are not themselves genotyped, then any estimated effect of the offspring genotype must be interpreted with caution. This is classic confounding: the maternal genotype can act as a cause of both the offspring genotype and the condition. So, particularly when studying a disease with onset early in life, one should consider genotyping both the offspring and the mothers (14). The implications of risk-based sampling for such a study are the subject of ongoing work.

One concern sometimes raised in the context of risk-based sampling, although with little empirical support, involves the idea that ''bludgeon'' genes may obstruct assessment of environmental effects. Consider the example of *BRCA1* and *BRCA2* genes in the sisters of women with breast cancer. It is estimated that about 0.2 percent of the population and 2 percent of breast cancer cases carry one or more risk alleles at these genes (www.cancer.gov/cancertopics/pdq/genetics/breast-and-ovarian/HealthProfessional/). If so, about 1 percent of sisters of cases are carriers. We estimate that in the Sister Study cohort, approximately 1,500 new cases will be diagnosed in 5 years of follow-up. On the basis of an odds ratio of 10 for mutation carriers, approximately 112 of these new cases will carry a deleterious *BRCA1* or *BRCA2* mutation. Thus, while the increase in prevalence for rare variants is substantial, the majority of cases of disease will not be attributable to this potent cause. Environmental factors should remain detectable. In particular, the study may also be able to identify environmental cofactors for the known breast cancer genes. The same reasoning would apply to any disease where the relative risk associated with having a first-degree relative who is affected is modest.

Two additional concerns are sometimes raised: does risk-based sampling limit generalizability of the findings, and does it limit the investigator's ability to study outcomes other than those selected for? Thus, for example, can findings for risk factors be taken to hold also for individuals who do *not* have a sibling with the condition under study? In addition, can other conditions be validly studied, for example, osteoporosis in the Sister Study cohort?

Even when risk has a familial pattern, however, most diseases are not strongly genetic. There is typically no reason to believe that the siblings of cases are fundamentally biologically different from individuals without an affected sibling. The concordance rate for breast cancer in identical twins is modest (15), suggesting that lifestyle and environmental factors play an important role. Moreover, studies of migrant populations reveal that for many complex diseases such as breast cancer, immigrants adopt the risk of their adopted country within a few generations (16), again suggesting that environmental and behavioral factors play a major role in determining susceptibility. While exposure relative risks that apply to first-degree relatives of cases may be slightly different from those in the general population, marked differences would not be expected. Neither the risk factors themselves nor the directions of association for those factors should be different for siblings of affected individuals.

A risk-based cohort is made up of volunteers, as is true of most cohort studies, and if one does not begin with a known sampling frame and achieve a high participation rate, then the ''worried well'' may be overrepresented. Although volunteer-based cohort studies enjoy internal validity, the measurable relative risks might not coincide precisely with those for the population at large. Nonetheless, risk-based sampling can greatly enhance our ability to design efficient prospective studies of complex conditions to identify both genetic and environmental contributors to risk.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

The authors acknowledge Drs. Beth Gladen and Allen Wilcox for providing many useful suggestions that improved the manuscript.

Conflict of interest: none declared.

REFERENCES

1. Smithells R, Sheppard S, Schorah C, et al. Possible prevention of neural-tube defects by periconceptional vitamin supplementation. Lancet 1980;1:339–40.
2. Antoniou A, Easton D. Polygenic inheritance of breast cancer: implications for design of association studies. Genet Epidemiol 2003;25:190–202.
3. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. Lancet 2001;358:1389–99.
4. Muhle R, Trentacoste S, Rapin I. The genetics of autism. Pediatrics 2004;113:472–86.
5. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982;115:119–28.

6. Breslow N, Cain K. Logistic regression for two-stage case-control data. Biometrika 1988;75:11–20.
7. Weinberg CR, Sandler DP. Randomized recruitment in case-control studies. Am J Epidemiol 1991;134:421–32.
8. Breslow N, Holubkov R. Weighted likelihood, pseudo likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. Stat Med 1997;16:103–16.
9. Umbach D, Weinberg C. The use of case-parent triads to study joint effects of genotype and exposure. Am J Hum Genet 2000;66:251–61.
10. Agresti A. Categorical data analysis. New York, NY: John Wiley & Sons, 1990:558.
11. Hopper JL, Carlin JB. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. Am J Epidemiol 1992;136:1138–47.
12. Wilcox AJ, Weinberg CR, Lie RT. Distinguishing the effects of maternal and offspring genes through studies of ''case-parent triads.'' Am J Epidemiol 1998;148:893–901.
13. Troisi R, Hatch E, Titus-Ernstoff L, et al. Birth weight and breast cancer risk. Br J Cancer 2006;94:1734–7.
14. Starr J, Hsu L, Schwartz S. Assessing maternal genetic associations: a comparison of the log-linear approach to case-parent triad data and a case-control approach. Epidemiology 2005;16:194–303.
15. Lichtenstein P, Niels V, Verkasalo P, et al. Environmental and heritable factors in the causation of cancer. N Engl J Med 2000;343:78–85.
16. Nelson N. Migrant studies aid the search for factors linked to breast cancer risk. J Natl Cancer Inst 2006;98:436–8.