

Assessing Environmental Modifiers of Disease Risk Associated with Rare Mutations

Alice S. Whittemore

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, Calif., USA

Key Words

Ascertainment bias · Case-case design · Odds-ratio · Prospective designs · Retrospective designs · Survival bias

Abstract

Objective: As disease-predisposing mutations are increasingly identified, there is growing need to assess the effects of lifestyle and environmental factors on disease risks in mutation carriers. Such assessment is difficult when the mutations are rare and evaluating them in large population samples is costly. **Methods:** This paper describes four study designs for evaluating the effects of environmental exposures in carriers of rare disease-predisposing mutations. **Results:** The strengths and weaknesses of the designs are assessed, and strategies for analyzing the data obtained from such designs are considered. **Conclusions:** When exposure effects in noncarriers are well-established and exposure is independent of carrier status in the population of disease-free controls, the case-only design provides a feasible and efficient method for inferring effects in carriers. When exposure effects in noncarriers are not well established, the most feasible design options are those that compare exposures in carrier cases to either untyped controls or to carrier controls. These two designs have complementary strengths and weaknesses; thus inferences are stronger when measures of association estimated using the two designs are consistent.

Copyright © 2007 S. Karger AG, Basel

Introduction

Many study design options are available for evaluating alterations in disease risk associated with exposures to modifiable lifestyle and environmental factors in general populations. However evaluating these alterations in carriers of rare disease-predisposing mutations poses several challenges. A major obstacle is the difficulty of ascertaining representative samples of mutation carriers when genetic testing is costly and the yield in carriers is low. This difficulty is compounded when the disease of interest also is rare. Typically measures of exposure-disease association are estimated using data collected on carriers identified through ongoing genetic testing programs [e.g. 1, 2]. An advantage of this approach is that it allows relatively large numbers of diseased and disease-free carriers to be identified rapidly. However, genetic testing is targeted at individuals with a strong family history of disease, so that the selection of carriers is not random with respect to disease status.

Evaluation of potential risk modifiers in mutation carriers is of considerable clinical importance, as these carriers struggle with options for preventive strategies. Moreover the central question carriers face is likely to be: 'How will my risk change if I am exposed to the factor, compared to what it would be if I were to remain unexposed?' Note that this question is focused on risks in carriers, and risks in noncarriers are not directly relevant. Thus the questions of interest concern the effects of ex-

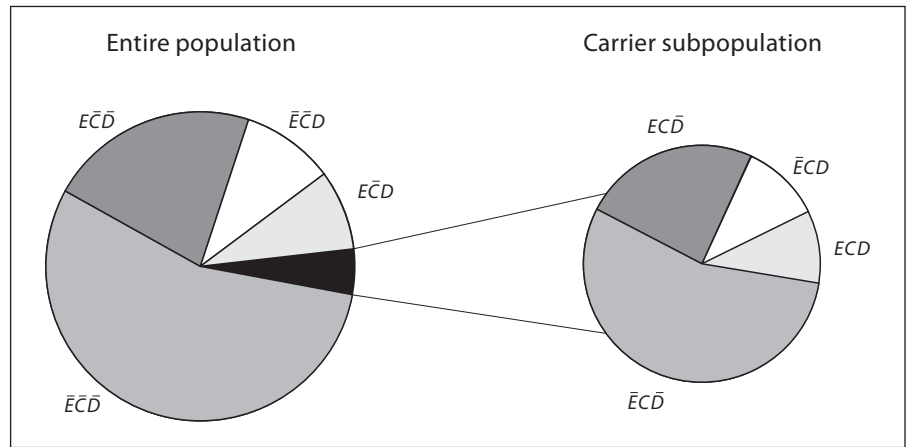


Fig. 1. Classification of the population of interest into eight subpopulations corresponding to status with respect to exposure (E = exposed, \bar{E} = unexposed), carrier status (C = carrier, \bar{C} = noncarrier), and disease (D = diseased, \bar{D} = disease-free).

posure compared to nonexposure in carriers, rather than whether the disease-exposure association differs in carriers and noncarriers.

An example is provided by efforts to evaluate the effects of oral contraceptive use on the risks of breast and ovarian cancer in carriers of pathogenic mutations of the genes BRCA1 and BRCA2. The relation between oral contraceptive use and risk of these cancers among women without BRCA mutations has been studied extensively [e.g. 3, 4]. Consequently, we now know that, while long-term oral contraceptive users are at reduced ovarian cancer risk compared to nonusers, their breast cancer risk is less clear, with some evidence suggesting risk elevation if use occurs before age 20 years or before first pregnancy. However since we do not understand the pathogenesis of BRCA-related breast and ovarian cancers, we cannot extrapolate measures of association from the general population to BRCA mutation carriers. There is particularly strong need to know the effects of oral contraceptive use on breast and ovarian cancer risks in these carriers, because their cancer risks are elevated [5], and it is important for them to avoid exposures that would elevate them even further.

This paper describes and evaluates four study designs that could be used to estimate measures of exposure effects in carriers. The strengths and weaknesses of several designs are discussed, as are strategies for analyzing the data obtained from such designs.

The Basic Model

Suppose we wish to evaluate a measure of association between a dichotomous nongenetic exposure (such as oral contraceptive use) and the risk of disease development in a given time period, in a population containing

some individuals who carry high-risk genotypes for the gene of interest. Let x_e denote an indicator for exposure prior to disease occurrence, assuming the value 1 if an individual is exposed and 0 otherwise, let x_c be an indicator for carrier status of a high-risk genotype for the disease-susceptibility gene, and let d be an indicator for disease occurrence in a given, fixed time period. In this simple situation the population can be divided into eight subgroups, consisting of individuals classified according to their joint statuses with respect to exposure, carriage of a high-risk genotype, and disease (fig. 1). Let p_{ij} denote the probability that a randomly sampled individual has exposure status $x_e = i$ and carrier status $x_c = j$, $i, j = 0, 1$, with $\sum_{i=0}^1 \sum_{j=0}^1 p_{ij} = 1$. We assume the following model for disease risk in terms of exposure to the risk factor and carrier status:

$$\text{logit}[\Pr(d=1|x_e, x_c)] \equiv \log \frac{\Pr(d=1|x_e, x_c)}{\Pr(d=0|x_e, x_c)} = \mu + \beta_e x_e + \beta_c x_c + \gamma x_e x_c \quad (1)$$

In equation (1) the quantities μ , β_e , β_c , γ are fixed parameters that characterize the population. The parameter $\mu \text{logit}[\Pr(d=1|0,0)]$ represents the log odds of disease in noncarriers who are unexposed to the environmental factor. The parameter

$$\beta_e = \text{logit}[\Pr(d=1|1,0)] - \text{logit}[\Pr(d=1|0,0)]$$

represents the log odds-ratio relating disease to exposure in noncarriers. Similarly, the parameter

$$\beta_c = \text{logit}[\Pr(d=1|0,1)] - \text{logit}[\Pr(d=1|0,0)]$$

represents the log odds-ratio relating disease to carrier status in unexposed individuals. The parameter γ measures the difference between the log odds-ratios associating disease with exposure in carriers and noncarriers.

From (1) we can write the probability that a randomly sampled individual has exposure, carrier and disease status i, j, d as

$$\Pr(i, j, d) = P_{ij} \frac{e^{(\mu + \beta_e + \beta_c + \gamma)j + \gamma ij} d}{1 + e^{\mu + \beta_e + \beta_c + \gamma ij}} \quad (2)$$

As discussed in the previous section, our primary focus concerns the effect of exposure on disease risk in carriers. This effect is measured by the log odds-ratio relating disease to exposure in carriers, which is the difference between the logit of disease risk in exposed and unexposed carriers. According to model (1), this difference is

$$\begin{aligned} \log OR_{DE|x_c=1} &= \text{logit}[\Pr(d=1 | 1,1)] - \text{logit}[\Pr(d=1 | 0,1)] \\ &= \mu + \beta_e + \beta_c + \gamma - (\mu + \beta_c) \\ &= \beta_e + \gamma \equiv \eta. \end{aligned} \quad (3)$$

In what follows interest will focus on the odds-ratio e^η .

So far we have ignored the temporal aspects of exposure and disease development. For some designs it may be appropriate to express the basic model (1) not as the logit of disease risk, but rather as the log incidence rate $\lambda(t)$ for disease at age t years:

$$\log \lambda(t) = \log \lambda_0(t) + \beta_e x_e + \beta_c x_c + \gamma x_e x_c \quad (4)$$

Here $\lambda_0(t)$ is the incidence rate in unexposed noncarriers, and x_e is taken to be an indicator for exposure during the period from birth to age t . In this setting the parameters β_e, β_c, γ and η have interpretations as appropriate log hazard ratios.

Our aim is to sample individuals from the population and assess their status with respect to disease, exposure and carriage of a high-risk genotype, in order to infer the parameter of interest.

Potential Biases

The designs used to assess exposure effects in carriers differ with respect to which of the eight population subgroups are ascertained, and which characteristics are assessed in ascertained individuals. All of them involve the sampling of carriers from the population. For pathogenic variants of low frequency, such sampling is complicated by the difficulty of ascertaining a representative sample of carriers. Testing a large sample of asymptomatic individuals for carriage typically is costly, and the carrier yield is apt to be too small for reliable inferences. We call this dilemma the *carrier ascertainment* problem. To deal with it, investigators often ascertain carriers by sampling and testing diseased individuals (who form a subpopulation enriched for carriers), or the relatives of diseased individuals who have been found to carry the variants. This strategy raises several issues.

The first issue is selection bias. If disease risk is heterogeneous within the population of carriers (due to variants with differing pathogenicity or to risk-modifying genetic or nongenetic factors that are correlated within families), carriers ascertained in this way are selected to have higher risks than those without a diseased relative [6]. In this case estimates of their lifetime disease risks would be biased upward. It has been proposed that even if the risks estimated using this ascertainment scheme are too large, measures of association between exposure and disease may well be invariant across subgroups of carriers at different risk. However this argument is inconsistent with the logic underlying the goal of evaluating exposure effects in carriers. This goal implicitly includes the possibility that exposure influences risk differently in carriers and noncarriers (else why go to the trouble of evaluating exposure effects in carriers?). Thus exposure may also influence risk differently in subgroups of carriers with different disease risks. In the face of possible within-carrier heterogeneity in both risk and exposure effects, one might argue that effects in carriers selected for high disease risk have the most clinical relevance, since these individuals carry the brunt of the disease burden. Still, the possibility of selection bias should be kept in mind when counseling carriers who present clinically without disease and with little or no family history of disease (see Liu et al. [7] for analysis of the impact of risk heterogeneity on estimates of absolute risk among carriers as a whole).

Despite the arguments in the preceding paragraph, measures of exposure-disease association often are assumed to be invariant across subgroups of the carrier population. Specifically, much of our current knowledge derives from designs that rely on the *Carrier Homogeneity* assumption, which states that the odds-ratio e^η relating disease to exposure does not vary within the subpopulation of carriers. Suppose, for example, that 'exposure' is educational level, and that the educational level of BRCA carriers ascertained in a family cancer clinic differs from that of unascertained carriers. The Carrier Homogeneity assumption would be violated only if this difference varied with disease status, because only then would the exposure-disease odds-ratio differ between the two subgroups of carriers (those available and unavailable to be ascertained). For the same reason, the assumption would be violated in a case-control study of gene carriers at increased risk of Parkinson's disease if cigarette smoking cases were less likely to participate than nonsmoking cases, but control participation rates were independent of smoking status. In this case, the difference between exposure (smoking) prevalence in ascer-

Table 1. Assumptions invoked to infer disease-exposure association in carriers of rare pathogenic genetic variants

Assumption	Specifications	Related References
Carrier homogeneity	Disease-exposure odds ratio does not vary within the subpopulation of carriers	Kelsey et al., pp 14–15 [27] Risch and Narod [28]
Outcome independence	Outcome variable (disease or exposure) occurs independently among related carriers	Khoury et al., p 188 [29] Slager and Schaid [24] Whittemore and Halpern [20]
Exposure-survival independence	Exposure is unrelated to risk of death from the disease of interest	Friedman, p 99 [30]
Control independence	Exposure and carrier status are unrelated in subpopulation of disease-free individuals	Piegorsch et al. [31] Yang and Khoury [11] Botto and Khoury [12]

tained and non-ascertained carriers varies by disease status, being negative for carrier cases and zero for carrier controls.

Note that in both of the previous two examples, the disease risks in ascertained carriers may differ from those in unascertained carriers (selection bias) without producing bias in estimated carrier odds-ratios e^n . Thus if the Carrier Homogeneity assumption holds, then the carriers ascertained for a study need not represent all carriers in either exposure prevalence or disease risk. Instead, one need only ensure that measures of exposure-disease association in the subpopulation of carriers available for ascertainment be representative of that pertinent to all carriers.

A second issue raised by the strategy of ascertaining carriers as relatives of diseased cases who are themselves carriers is the possibility of correlated outcomes among study subjects who are closely related. If the assessed outcome is disease status, then possible correlation in disease risk among related carriers should be accommodated in analysis of risk among exposed and unexposed. If the assessed outcome is exposure, then possible exposure correlations within families should be accommodated in analysis of exposure differences among case and control carriers. Analyses that fail to acknowledge such correlations rely on the *Outcome Independence* assumption, which states that the outcome variable (which would be disease occurrence for prospective studies and exposure for retrospective studies) occurs independently among related carriers included in the study. Violation of this assumption and failure to accommodate it in the analysis typically does not bias estimates of the association parameters of interest, but it can bias measures of uncertainty such as variances and confidence limits. As dis-

cussed in the Data Analysis section, robust variance estimates are available through the use of analytic methods that allow the extent of pairwise correlation between relatives to vary with the closeness (kinship coefficients) of the relative pairs.

A third issue is one of survival bias. Most designs ascertain only living case carriers, because of the difficulty of obtaining accurate exposure histories from relatives of the deceased. Living case carriers often report their prediagnostic exposures several years after their disease diagnoses, and these individuals are called *prevalent* (rather than *incident*) cases. If the exposure is associated with altered survival, then this selection strategy may result in biased measures of exposure-disease association. Designs using prevalent cases for inferences about the etiological effects of prediagnostic exposures rely on the *Exposure-Survival Independence* assumption, which states that the exposure of interest is unrelated to the risk of death from the disease of interest. Table 1 lists the assumptions that may be needed for valid inferences.

Study Designs

The following description of design options begins with an idealized design for assessing the effects of exposure on disease risk in carriers. Because of the infeasibility of this design, other, more practical ones are commonly used. Common alternative designs are described and their strengths and weaknesses are summarized in table 2.

Design 1: Comparison of Disease Risk in Exposed and Unexposed Carriers

An ideal design would obtain representative samples of exposed and unexposed carriers from the population,

Table 2. Study designs for evaluating disease-exposure association in carriers of rare pathogenic genetic variants

Study design	Individuals ascertained	Characteristics assessed	Examples	Strengths	Weaknesses
1. Comparison of disease risk in exposed and unexposed carriers	exposed and unexposed carriers	disease status; time to disease	not available	allows direct assessment of exposure-disease association in carriers; robustness against failure of Exposure-Survival Independence and Control Independence assumptions	infeasible (see text) Vulnerable to failure of Carrier Homogeneity and Disease Independence assumptions
2. Exposure comparison in carrier and noncarrier cases (case-only design)	carrier and noncarrier cases	exposure status	OC use and breast cancer risk in carriers of BRCA mutations [Ursin et al., 1997]	straightforward implementation; robustness to failure of Carrier Homogeneity, Exposure Independence and Exposure-Survival Independence assumptions	allows inference for interaction parameter (subject to Control Independence assumption) but not for exposure-disease association in carriers
3. Exposure comparison in carrier cases and untyped controls	carrier cases and noncarrier controls	exposure status	OC use in relation to risk of breast cancer [Milne et al., 2005] and ovarian cancer [McGuire et al., 2004]	same as those of Design 2	vulnerable to failure of Control Independence assumption
4. Exposure comparison in carrier cases and carrier controls	carrier cases and carrier controls	exposure status	OC use in relation to risk of ovarian cancer [Narod et al., 1998; Whittemore et al., 2004] and breast cancer [Haile et al., 2006]	allows direct assessment of exposure-disease association in carriers; robustness against failure of Control Independence assumption	vulnerable to failure of Carrier Homogeneity, Outcome Independence, and Exposure-Survival Independence assumptions

follow them for disease occurrence, and evaluate associations by comparing disease risk in the two groups (upper panel of fig. 2). This design has the advantage of allowing direct assessment of the parameter η that measures disease-exposure association in carriers. In the terminology of Fleiss et al. [8], the design might be cross-sectional (ascertained because of carrier status but not because of status for disease or exposure) or prospective (ascertained because of carrier and exposure status but not because of disease status). As noted above, the carrier ascertainment problem complicates the task of ascertaining a random sample of the carriers in the population. In addition, for diseases affecting specific organs, carriers often opt to reduce or effectively eliminate their risks by undergoing prophylactic organ removal (e.g. prophylactic mastectomy or oophorectomy). This choice leads to the following catch-22. On the one hand, currently available preventive strategies (such as prophylactic surgery) are invasive and information on the efficacy of less drastic alternatives is needed. On the other hand, such information cannot be obtained (at least using this ideal design), because of lack of power due to the fact that many carriers have removed the organ at risk.

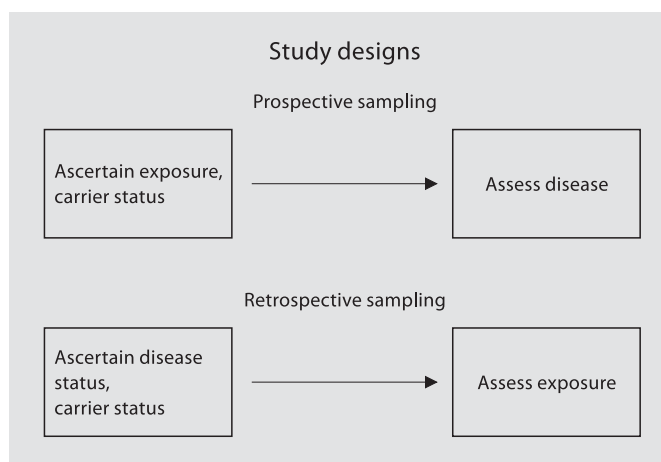


Fig. 2. Study designs for ascertaining the effects of exposure on disease risk in carriers.

Because of the infeasibility of this design, epidemiologists typically use one of the following retrospective designs (lower panel of fig. 2). These designs rely on the invariance of the exposure-disease odds-ratio interpreta-

tion with respect to disease risk and exposure prevalence, and its estimability from data collected retrospectively [9]. As applied to carriers, all three designs rely implicitly on one or more of the assumptions in table 1.

Design 2: Exposure Comparison in Carrier and Noncarrier Cases (the Case-Only Design)

In this design a representative sample of diseased cases is ascertained and tested for carrier status. The outcome variable is exposure, which is assessed in all sampled cases. This ‘case-only’ or ‘case-case’ design has been used to study the effects of oral contraceptive use on breast cancer risk in BRCA mutation carriers [10], and it has been discussed widely in the literature [e.g. 11, 12]. Its validity does not require the Carrier Homogeneity assumption (since the carrier cases are ascertained as a representative sample of all diseased carriers in the population) and, if incident cases are ascertained, it does not require the Exposure-Survival assumption. The design allows estimation of the exposure-carrier odds-ratio in cases, which is denoted by $OR_{EC|d=1}$. As shown in the Appendix, this odds-ratio can be written in terms of the interaction parameter γ and the exposure-carrier odds-ratio $OR_{EC|d=0}$ in controls:

$$OR_{EC|d=1} = e^\gamma OR_{EC|d=0}. \quad (5)$$

Equation (5) shows two important properties of the case-only design. First, it allows estimation of the interaction parameter γ but not the odds-ratio e^γ associating risk to exposure in carriers, which is of primary interest. As can be seen from equation (3), this property need not be a limitation if the parameter β_e measuring disease-exposure association in noncarriers is known. For example, it may be known from extensive prior research that there is no association between disease and exposure in noncarriers (i.e., $\beta_e = 0$). Then equation (3) indicates that $\gamma = \eta$ and this design is directly informative for η . Alternatively, as with ovarian cancer and long-term oral contraceptive use, the odds-ratio e^{β_e} relating exposure to disease may be known (say $e^{\beta_e} = 0.5$). Then an estimate for γ that differs significantly from $\log(.5)$ would be also be informative for η . However if β_e is not known the design cannot inform η , the parameter of interest.

The second design property evident from equation (5) is that unbiased estimation of γ requires the assumption that $OR_{EC|d=0} = 1$, i.e., that there is no association between exposure and carrier status among the controls. This assumption is typically unverifiable because of the carrier ascertainment problem. I shall call it the *Control Independence* assumption (row 4 of table 1). It states that

exposure and carrier status are independent in the subpopulation of disease-free individuals. For example, this assumption would be violated if carrier controls were more likely than noncarrier controls to use oral contraceptives.

Design 3: Exposure Comparison in Carrier Cases and Untyped Controls

This design is an offshoot of the classical case-control study design. It is illustrated by several studies evaluating the effects of oral contraceptive use on carriers’ risks of breast or ovarian cancer [13–15]. In these studies, representative samples of incident cases and disease-free controls are ascertained without regard to carrier status, and only the cases are tested for carriage of a pathogenic variant. The controls are not tested because of the ascertainment problem. The prevalence of exposure in carrier cases is compared to that in the full set of all controls.

Advantages and disadvantages of this design are similar to those of the case-only design. Advantages include its robustness to departures from the Carrier Homogeneity assumption (since the carrier cases are a representative sample of all diseased carriers) and from the Exposure-Survival Independence assumption (since only cases with incident disease are included). A weakness is its reliance on the Control Independence assumption, which is needed to insure that exposure prevalence in the comparison group of untested controls is similar to that in carriers without the disease. The importance of this assumption can be verified formally by noting that this design allows inference on the exposure odds in carrier cases divided by the exposure odds in untyped controls:

$$\frac{\Pr(x_e = 1 | x_c = 1, d = 1)}{\Pr(x_e = 0 | x_c = 1, d = 1)} \div \frac{\Pr(x_e = 1 | \cdot, d = 0)}{\Pr(x_e = 0 | \cdot, d = 0)}. \quad (6)$$

Here $\Pr(x_e = 1 | \cdot, d = 0)$ denotes exposure prevalence in the untyped controls, which can be written

$$\Pr(x_e = 1 | \cdot, d = 0) = \frac{\Pr(x_e = 1 | x_c = 1, d = 0)q + \Pr(x_e = 1 | x_c = 0, d = 0)(1 - q)}{\Pr(x_e = 1 | x_c = 0, d = 0)(1 - q)}, \quad (7)$$

where q represents carrier prevalence among controls. The Control Independence assumption states that exposure and carrier status are independent among controls, i.e., that

$$\Pr(x_e = 1 | x_c = 1, d = 0) = \Pr(x_e = 1 | x_c = 0, d = 0).$$

Using this assumption in equation (7) gives

$$\Pr(x_e = 1 | \cdot, d = 0) = \Pr(x_e = 1 | x_c = 1, d = 0).$$

Thus (6) can be expressed as

$$\frac{\Pr(x_e = 1 | x_c = 1, d = 1)}{\Pr(x_e = 0 | x_c = 1, d = 1)} \div \frac{\Pr(x_e = 1 | x_c = 1, d = 0)}{\Pr(x_e = 0 | x_c = 1, d = 0)} = OR_{DE|x_c=1} = e^\eta.$$

For this design there also is need to insure that the controls are matched to the subgroup of carrier cases with respect to potential confounding variables, such as age. For example, the null findings of Modan et al. [13] concerning oral contraceptive use and ovarian cancer in BRCA mutation carriers may have occurred because the oral contraceptive use of carrier cases was compared to that of controls who were older than they were. This difference occurred because controls were age-matched to all cases, whereas the carrier cases were younger than the noncarrier cases, and the analytic comparisons were stratified only in broad ten-year age categories. Thus the comparison controls were born earlier than the carrier cases, and had less opportunity for long-term exposure to oral contraceptives, which did not become widespread until after 1960.

Design 4: Exposure Comparison in Carrier Cases and Carrier Controls

This design is illustrated by several studies of oral contraceptive use in relation to risk of breast or ovarian cancer in carriers of BRCA mutations [e.g. 2, 16–18]. It shares with Design 1 the advantage of allowing direct assessment of the log odds-ratio η that measures disease-exposure association in carriers. It also is robust against departures from the Control Independence assumption, because noncarriers are not ascertained. Offsetting these strengths are some important limitations, including vulnerability to departures from the Carrier Homogeneity, the Exposure Independence and the Exposure-Survival Independence assumptions. A key assumption is that while ascertainment may depend on a carrier's disease status, it should not depend on his or her exposure.

Data Analysis

The analytic plan for a given design should accommodate the way study subjects are ascertained and the characteristics that are assessed in these subjects. Ideally, at least some of the design limitations could be addressed or offset by a judicious choice of analysis. However, it appears that, among the limitations of the four designs discussed in the previous section (i.e., their sensitivity to violation of one or more of the assumptions in table 1), only departures from the Outcome Independence assumption can be addressed in the analysis. This assumption is needed for Designs 1 and 4. The following is a brief discussion of analytic methods for the four designs.

Design 1: Comparison of Disease Risk in Exposed and Unexposed Carriers

This design, if feasible, would allow direct inference regarding measures of exposure-disease association. Departures from the Outcome Independence assumption (in this case the assumed independence of disease occurrence in related carriers) could be handled by assuming either a random effects model or a marginal model for the joint probability of disease among the carriers in a family, given their exposures. However measures of exposure-disease association (odds-ratios, hazard ratios) specified by the two models have different interpretations. The odds-ratio or hazard ratio specified by the random effects model measures exposure-disease association conditional on family membership, while the one specified by the marginal model measures exposure-disease association in a randomly selected carrier, regardless of the family to which he or she belongs (see Diggle et al. [19] for further discussion.) The two measures can actually differ in magnitude when carriers' disease risks are correlated within families [20]. Thus estimates for them are not comparable. When disease occurrence in a fixed time period is modeled via the logistic form (3), family-matched conditional logistic regression produces consistent estimates for the odds-ratio in a random-effects model, and consistent variance estimates. This strategy is problematic, however, if a substantial proportion of the ascertained carriers drop out of the analysis because they lack a matched relative discordant with respect to disease and/or exposure status. An alternative in this case is to assume a marginal logistic model, and use generalized estimating equations and robust variance estimates [21] to infer measures of association. When censored time-to-disease data are assumed to follow the proportional hazards model (4), within family correlation in disease risk can be addressed using random effects frailty models (e.g. [22]), or marginal models [19, 23].

Design 2: Exposure Comparison in Carrier and Noncarrier Cases

We have seen that when the exposure-disease odds-ratio in noncarriers is known, this design can provide useful information for the corresponding odds-ratio in carriers, provided the Control Independence assumption is met. It is well-known that standard logistic regression methods can be used to infer the interaction parameter γ . Potential confounding variables (e.g., age at diagnosis), can be controlled using conditional logistic regression (conditional logistic regression) of age-matched strata, or unconditional logistic regression (uncondition-

al logistic regression) of exposure data from cases and controls frequency-matched on age.

Design 3: Exposure Comparison in Carrier Cases and Untyped Controls

Analytic methods for this design are similar to those used to analyze data from standard case-control studies. Potential confounding variables can be controlled by matching in the design stage or regression modeling in the analysis stage.

Design 4: Exposure Comparison in Carrier Cases and Carrier Controls

Potential pitfalls of this design include sensitivity to departures from three of the four assumptions described in table 1: Carrier Homogeneity, Outcome Independence (here the assumption that the exposures of related carriers are independent) and Exposure-Survival Independence assumptions. As noted above, only sensitivity to the Exposure Independence assumption lends itself to amelioration by analysis. Whittemore and Halpern [20] considered the problem of estimating the exposure-disease odds-ratio e^η in carriers under model (3), using data from related carriers who were ascertained because of their disease phenotypes. The authors describe random effects and marginal models for the joint distribution of exposures among a set of related carriers, conditional on their disease statuses and genetic relationships. As noted above for the analysis of prospective data from Design 1, measures of association defined by the two models have different interpretations and different magnitudes when exposures are correlated among related carriers. Whittemore and Halpern [20] and Slager and Schaid [24] generalize to correlated exposures the results obtained in the absence of correlation by Prentice and Pyke [9], who showed that odds-ratios can be estimated consistently by assuming the data had been collected prospectively as in Design 1. These authors also propose robust variance estimates that are consistent in the presence of correlated exposures among relatives. The variance estimates are computed from a matrix of estimated exposure covariances for pairs of relatives. The exposure covariances can be specific to relative pairs with given kinship coefficients, so that covariances of differing magnitudes in large pedigrees can be handled, provided the data contain adequate numbers of relative pairs with any given degree of relatedness. But in most applications exposure correlations may not be strong among cousins or more distant relatives, so that two types of correlations (e.g. those for 'close' and 'distant' pairs) may be adequate.

In an alternative approach to analyzing data from Design 4, Antoniou et al. [25] treat the data as if they were prospectively ascertained time-to-disease data as in Design 1, and account for the disease-specific ascertainment actually used by differentially weighting the contributions from carrier cases and carrier controls. Specifically, the authors assume a proportional hazards model for censored time to disease among carriers and use weighted Cox regression analysis to estimate the hazard ratio associated with exposure. The idea is similar to that used in large-scale survey studies that oversample specific subpopulations, where weights are introduced to make the resulting estimates consistent with known population values [26]. Here the average diseased risks are not known but must be inferred from average disease risks in the entire carrier population. The method also provides variance estimates that account for the weighting, but not for possible exposure correlation among related carriers in the study.

Discussion

This article has reviewed the assumptions needed in studies that attempt to assess associations between lifestyle or environmental exposures and disease risk in carriers of rare mutations. The strengths and weaknesses of four study designs have been discussed, with particular emphasis on potential inferential bias due to departures from the assumptions. Analytic methods for the designs have been reviewed.

A natural question in choosing among a set of study design options concerns statistical power. In our context the issue is power to reject the null hypothesis that $\eta = 0$, where η is the log odds-ratio relating exposure to disease risk in mutation carriers. Little has been published concerning the relative power of the four designs discussed here. In the special case that the noncarrier exposure-disease parameter β_e has been determined from previous studies, the case-only design is known to be more powerful for testing hypotheses about η than a standard case-control design in which both exposure and carrier status are assessed in cases and controls. Arguments similar to those underlying this result can be invoked to show that the 'carrier-only' case-control design 4 has more power than one that assesses exposure and carrier status in all cases and controls. Nevertheless relative power considerations seem less important to the choice of a design than potential inferential bias. It is difficult to specify circumstances under which violations of the assumptions in ta-

ble 1 are likely to be extreme enough to cause seriously misleading inferences. Since Designs 3 and 4 have somewhat complementary strengths and weaknesses, inferences are stronger when measures of association using the two designs are consistent.

Acknowledgements

This research was supported by NIH grants CA97359 and CA94069.

Appendix: Odds-Ratio for the Case-Only Design

To verify equation (5) we write the exposure-carrier odds-ratio among cases as

$$OR_{EC|d=1} = \frac{\Pr(1,1,1)\Pr(0,0,1)}{\Pr(1,0,1)\Pr(0,1,1)} \quad (8)$$

where $\Pr(i, j, d)$ denotes the joint probability that an individual in the population has exposure status i , carrier status j , and disease status d . Dividing and multiplying the right side of (8) by the exposure-carrier odds-ratio $OR_{EC|d=0}$ in controls, rearranging terms and using equation (3) gives

$$\begin{aligned} OR_{EC|d=1} &= \left[\frac{\Pr(1,1,1)\Pr(0,0,1)}{\Pr(1,0,1)\Pr(0,1,1)} \div OR_{EC|d=0} \right] OR_{EC|d=0} \\ &= \left[\frac{\Pr(1,1,1)\Pr(0,0,1)}{\Pr(1,0,1)\Pr(0,1,1)} \div \frac{\Pr(1,1,0)\Pr(0,0,0)}{\Pr(1,0,0)\Pr(0,1,0)} \right] OR_{EC|d=0} \\ &= \left[\frac{\Pr(1,1,1)\Pr(0,0,1)}{\Pr(1,1,0)\Pr(0,0,0)} \div \frac{\Pr(1,0,1)\Pr(0,1,1)}{\Pr(1,0,0)\Pr(0,1,0)} \right] OR_{EC|d=0} \\ &= \exp\left[(\mu + \beta_e + \beta_c + \gamma) + \mu - (\mu + \beta_e) - (\mu + \beta_c)\right] OR_{EC|d=0} \\ &= e^\gamma OR_{EC|d=0} \end{aligned}$$

in agreement with (5).

References

- Couch FJ, DeShano ML, Blackwood MA, Calzone K, Stopfer J, Campeau L, Ganguly A, Rebbeck T, Weber BL: BRCA1 mutations in women attending clinics that evaluate the risk of breast cancer. *N Engl J Med* 1997;336:1409–1415.
- Narod SA, Risch H, Moslehi R, Dørum A, Neuhausen S, Olsson H, Provencher D, Radice P, Evans G, Bishop S, Brunet JS, Ponder BAJ, Klijn JGM, for The Hereditary Ovarian Cancer Clinical Study Group: Oral contraceptives and the risk of hereditary ovarian cancer. *Hereditary Ovarian Cancer Clinical Study Group. N Engl J Med* 1998;339:424–428.
- Whittemore AS, Harris R, Itnyre J, Collaborative Ovarian Cancer Group: Characteristics relating to ovarian cancer risk: Collaborative analysis of twelve US case-control studies. II. Invasive epithelial ovarian cancers in white women. *Am J Epidemiol* 1992;136:1184–1203.
- Beral V, Reeves G: Childbearing, oral contraceptive use, and breast cancer. *Lancet* 1993;341:1102.
- Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjäkoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, Easton DF: Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet* 2003;72:1117–1130.
- Begg CB: On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* 2002;94:1221–1226.
- Liu W, Icitovic N, Shaffer ML, Chase GA: The impact of population heterogeneity on risk estimation in genetic counseling. *BMC Med Genet* 2004;5:18.
- Fleiss JL, Levin B, Paik MC: *Statistical Methods for Rates and Proportions*, 3rd ed. Hoboken, NJ, Wiley Interscience, 2003.
- Prentice RL, Pyke R: Logistic disease incidence models and case-control studies. *Biometrika* 1979;66:403–411.
- Ursin G, Henderson BE, Haile RW, Pike MC, Zhou N, Diep A, Bernstein L: Does oral contraceptive use increase the risk of breast cancer in women with BRCA1/BRCA2 mutations more than in other women? *Cancer Res* 1997;57:3678–3681.
- Yang Q, Khoury MJ: Evolving methods in genetic epidemiology. III. Gene-environment interaction in epidemiologic research. *Epidemiol Rev* 1997;19:33–43.
- Botto LD, Khoury MJ: Commentary: Facing the challenge of gene-environment interaction: the two-by-four table and beyond. *Am J Epidemiol* 2001;153:1016–1020.
- Modan B, Hartge P, Hirsh-Yechezkel G, Chetrit A, Lubin F, Beller U, Ben-Baruch G, Fishman A, Menczer J, Struewing JP, Tucker MA, Ebbens SM, Friedman E, Piura B, Wacholder S, for the National Israel Ovarian Cancer Study Group: Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *N Engl J Med* 2001;345:235–240.
- McGuire V, Felberg A, Mills M, Ostrow KL, DiCiccio RA, John EM, West DW, Whittemore AS: Relation of contraceptive and reproductive history to ovarian cancer risk in carriers and noncarriers of BRCA1 gene mutations. *Am J Epidemiol* 2004;160:613–618.
- Milne RL, Knight JA, John EM, Dite GS, Balbuena R, Ziogas A, Andrulis IL, West DW, Li FP, Southey MC, Giles GG, McCredie MRE, Hopper JL, Whittemore AS, for the Breast Cancer Family Registry: Oral contraceptive use and risk of early-onset breast cancer in carriers and noncarriers of BRCA1 and BRCA2 mutations. *Cancer Epidemiol Biomarkers Prev* 2005;14:350–356.

- 16 Haile RW, Thomas DC, McGuire V, Felberg A, John EM, Milne RL, Hopper JL, Jenkins M, Levine AJ, Daly MB, Buys SS, Senie RT, Andrulis IL, Knight JA, Godwin AK, Southey M, McCredie MR, Giles GG, Andrews L, Tucker K, Miron A, Apicella C, Tesoriero A, Bane A, Pike MC, kConFab Investigators; Ontario Cancer Genetics Network Investigators, Whittemore AS: BRCA1 and BRCA2 mutation carriers, oral contraceptive use and breast cancer risk before age 50. *Cancer Epidemiol Biomarkers Prev* 2006;15:1863–1870.
- 17 McGuire V, John EM, Felberg A, Haile RW, Boyd NF, Thomas DC, Jenkins MA, Milne RL, Daly MB, Ward J, Terry MB, Andrulis IL, Knight JA, Godwin AK, Giles GG, Southey M, West DW, Hopper JL, Whittemore AS: No increased risk of breast cancer associated with alcohol consumption among carriers of BRCA1 and BRCA2 mutations ages <50 years. *Cancer Epidemiol Biomarkers Prev* 2006;15:1565–1567.
- 18 Boyd NF, John EM, McGuire V, Felberg A, Haile R, Milne R, Trench G, Daly M, Buys S, Senie R, Knight J, Giles G, Andrulis I, McCredie M, Godwin A, Terry MB, Miron A, Beck J, Venne V, Ross, E Southey M, Ozelik H, Hopper J, Whittemore AS: Smoking and risk of breast cancer in carriers of mutations in BRCA1 and BRCA2 aged less than 50. *Cancer Epidemiol Biomarkers Prev* (in press).
- 19 Diggle PJ, Liang K-Y, Zeger SL. 1994 *Analysis of Longitudinal Data*. Oxford, Clarendon Press, 1994.
- 20 Whittemore, AS, Halpern J: Logistic regression of family data from retrospective study designs. *Genet Epidemiol* 2003;25:177–189.
- 21 Liang K-Y, Zeger SL: Longitudinal data analysis using generalized linear models. *Biometrika* 1986;73:13–22.
- 22 Nielsen GG, Gill RD, Andersen PK, Sorensen TIA: A counting process approach to maximum likelihood estimation in frailty models. *Scand J Stat* 1992;19:25–43.
- 23 Prentice RL, Hsu L: Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika* 1997;84:349–363.
- 24 Slager SL, Schaid DJ: Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am J Hum Genet* 2001;68:1457–1462.
- 25 Antoniou AC, Goldgar DE, Andrieu N, Chang-Claude J, Brohet R, Rookus MA, Easton DF: A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genet Epidemiol* 2005;29:1–11.
- 26 Korn EL, Graubard BI: Examples of differing weighted and unweighted estimates from a sample survey. *Am Stat* 1995;49:292–295.
- 27 Kelsey JL, Whittemore AS, Evans AS, Thompson WD: *Methods in Observational Epidemiology* 2nd ed. New York, Oxford University Press, 1996.
- 28 Risch HA, Narod SA: Re: On the use of familial aggregation in population-based case probands for calculating penetrance. *J Natl Cancer Inst* 2003;95:73–74.
- 29 Khoury M, Beatty T, Cohn BH: *Fundamentals of Genetic Epidemiology*. New York, Oxford University Press, 1993.
- 30 Friedman GD: *Primer of Epidemiology* 5th ed. New York, McGraw Hill, 2004.
- 31 Piegorsch WW, Weinberg CR, Taylor JA: Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med* 1994;13:153–162.