On the Use of Familial Aggregation in Population-Based Case Probands for Calculating Penetrance

Colin B. Begg

Background: Estimating the lifetime risk associated with (i.e., the penetrance of) genetic abnormalities that predispose individuals to cancer is important for genetic counseling. (Penetrance may be estimated from the degree of familial aggregation of cancer, that is, the extent to which cancers cluster in families.) Early penetrance studies of BRCA1 and BRCA2 mutations used high-risk families with multiple cases of breast cancer, a study design that led to very high penetrance estimates. However, such studies were subject to potential ascertainment biases. To offset such biases, recent studies have used data from family members of probands ascertained from population-based incident cases of cancer. The use of case probands is, however, also subject to bias because all risk factors are over-represented in case patients. To draw attention to this problem, literature on the penetrance of breast cancer in BRCA1 and BRCA2 carriers is reviewed. Methods: A theory is presented to show that the use of case probands is itself biased, leading to inflated penetrance estimates. The strategy is unbiased only if all carriers share an identical risk. Any unexplained heterogeneity of risk caused by unknown genetic or shared environmental factors within families leads to an inflated estimate of penetrance. Results: Eight published studies using population-based methods are reviewed. All but one of the family-based studies used case probands. Conclusions: Penetrance estimates from case proband studies must be inflated if other factors influence breast cancer risk in addition to the specific genetic abnormality. Thus, women with such genetic abnormalities and a strong family history of breast cancer are likely to possess a much higher risk for breast cancer than women with such abnormalities but without a strong family history. Methodologic techniques to improve the prediction of cancer risk are needed. [J Natl Cancer Inst 2002;94:1221–6]

In recent years, the discovery of genes associated with the risk for common cancers has led to an intense research effort to evaluate the epidemiologic characteristics of germline abnormalities in these genes. Founder mutations in BRCA1 and BRCA2 genes that are associated with breast cancer have received especially intense scrutiny because of the high incidence of this disease and because the population prevalences of these mutations are non-negligible—especially among Ashkenazi Jews, approximately 2.5% of whom carry a founder mutation (1–3). Of particular interest is the penetrance, defined as the lifetime risk for developing the disease among carriers, frequently reported as the probability of cancer by the age of 70 years. This statistic is especially important in the genetic counseling of individuals who are identified as carriers by genetic testing. Such individuals are faced with daunting options for cancer prevention, such as the prophylactic surgical removal of both breasts (4). For such women, the accuracy of statistical estimates of penetrance is profoundly important.

Almost all studies to date have used the tools of genetic epidemiology to estimate the penetrance of BRCA1 and BRCA2 mutations. Penetrance is derived from the degree of familial aggregation of cancer—that is, the extent to which cancers cluster within families. These data have been used to determine age-specific risk. Early studies involved families with multiple occurrences of breast cancer that had been used to establish the linkage with BRCA1 and BRCA2 in the first place. The Breast Cancer Linkage Consortium has assembled many such families, and these data have led to penetrance estimates ranging from 71% to 85% by 70 years of age (5–8). However, ascertainment of these families is subject to ill-defined selection effects leading to serious doubts about the validity of the estimates (9,10).

Concerns about this issue have led other investigators to pursue the ascertainment of carrier families that are unselected on the basis of family history of cancer. An influential study was conducted in the Washington, DC, metropolitan area, in which the investigators advertised for volunteers who were self-identified as Jewish. A total of 5318 volunteers provided blood for genetic analysis, and 120 carriers were identified who had one (or more) of the three breast cancer-associated mutations common in Ashkenazi Jews (185delAG and 5382insC in BRCA1 and 6174delT in BRCA2). Examination of the occurrence of breast cancer in the relatives of these carriers led to a combined penetrance estimate of 56%, a result substantially lower than the earlier estimates from cancer families (3). Since then, several studies have been published in which data from population-based case-control studies or consecutive series of hospital-based case patients with incident cancers have been used to identify carriers unselected on the basis of family history of cancer. These studies have exclusively used probands who have been diagnosed with cancer and have led to penetrance estimates that are considerably lower than the earlier estimates from families with multiple cases of breast cancer.

Case probands are used in these population-based studies for pragmatic reasons. [Throughout this article, the term proband will be used to mean the individual used to identify the family for study. Thus, the proband need not have been affected with the disease, in contrast to conventional usage of this term (11): a case proband is a proband who has been diagnosed with breast cancer, and a control proband is a proband who is disease free.] Because mutations in the BRCA1 and BRCA2 genes occur in a small percentage of the population at risk for breast cancer, genotyping of population-based control subjects will identify carriers very infrequently. By contrast, carriers are more frequent among case patients with incident breast cancer. Thus, the strategy of using case patients to identify probands with and without mutations and then using the relatives of these probands to calculate penetrance is appealing.

Correspondence to: Colin B. Begg, Ph.D., Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021 (e-mail: beggc@mskcc.org). See “Notes” following “References.”

© Oxford University Press
However, little attention has been paid to the methodologic implications of using case probands rather than control probands in studies of penetrance, and advice to epidemiologists on this issue has been vague. Wacholder et al. (12), although stating the importance of the fact that the probands must be representative of the population at risk, nonetheless allow for the use of a “series of patients with tissue available for genotyping …” for convenience, acknowledging that this approach may lead to the overestimation of penetrance because of “overrepresentation of high-risk families.” Hopper et al. (13), in an exposition of principles of genetic epidemiology study designs, advocate collecting family history data from population-based control subjects but do not articulate the role of control subjects in the estimation of penetrance.

Unfortunately, case probands are fundamentally different from control probands, regardless of carrier status, because of size-biased sampling. That is, all risk factors for breast cancer are over-represented in incident cases of breast cancer, and so a sample of case probands who are mutation carriers will also have a higher frequency of all other risk factors than control probands who are mutation carriers, unless the mutation is the sole determining cause of cancer in all mutation carriers. The purpose of this article is to draw attention to this problem and to review the literature on the penetrance of breast cancer in BRCA1 and BRCA2 carriers so that the potential for bias in published estimates can be evaluated.

METHODS

Once a set of probands has been identified, family-based methods for estimating penetrance involve identifying close relatives at risk, establishing the probability that each relative is a carrier of a mutation in BRCA1 or BRCA2 by use of Mendelian inheritance, and then accumulating the breast cancer events and the age-specific periods of follow-up to estimate age-specific risks, after adjusting for the probabilities that the contributing family members are carriers. One popular approach is the “kin–cohort design” (12). In this approach, penetrance is estimated actuarially with a cohort of individuals who are first-degree relatives of genotyped probands. Comparisons of the risks in carriers with the risks in noncarriers can be expressed very simply in terms of the observed risk estimates in relatives of carriers (“carrier kin”) versus relatives of noncarriers (“non-carrier kin”). Indeed, for studies of a low-prevalence mutation limited to first-degree relatives, the difference between the risk for carriers and the risk for noncarriers is approximately double the difference between the risk for carrier kin and the risk for noncarrier kin. A closely related approach, termed the genotyped proband design, has been developed by Gail et al. (14). The same study design is used, but the statistical analysis is adapted to use the relative frequencies at which a mutation is observed in the probands.

A fundamental assumption in both of these methods is that the risks are presumed to be no different for family members of a carrier who is a case patient than for family members of a carrier who is a control subject. Risk is thus assumed to depend solely on the presence of the mutation, an assumption also used by Parmigiani et al. (15) in their algorithm to predict mutational status on the basis of a family history of breast and ovarian cancer. Indeed, in their primary analysis of data from the Washington Ashkenazi Study, Wacholder et al. (12) combined the data from their probands regardless of whether they were patients with (or survivors of) breast and/or ovarian cancer (27 mutation carriers) or subjects at risk of breast cancer—that is, those who did not have the disease (62 mutation carriers). Subsequently, investigators studying the penetrance of BRCA1 or BRCA2 mutations have abandoned control probands entirely and used only data from case probands (16–21).

Penetrance estimates are meaningful only for individuals who are at risk, i.e., women who do not have breast cancer. Thus, the population at risk is the population of female carriers who do not (yet) have breast cancer. However, the following analysis shows that the distribution of risks in the population at risk is fundamentally different (i.e., lower risks in general) than the distribution of risks in incident cases. Let \( r \) denote the risk of a randomly selected carrier who is at risk in the chosen population, and consider the possibility that there is heterogeneity in the risks for different carriers. For simplicity, consider risk to be a single measure for each individual, even though risk actually changes as the individual ages. Let \( p(r) \) be the probability density of the risks in carriers, with the mean risk denoted by \( \mu \).

Then, \( \mu \) is the penetrance that is being estimated. In an ideal, though impractical, study design, one would identify a large number of carriers in the population and follow them over time to observe incident cancers and to estimate this (average) penetrance directly.

Consider now the distribution of risks in carriers identified from an incident series of cases. The risk distribution of these carriers would not be \( p(r) \) but, instead, would be size-biased, in the terminology of Patil and Rao (22). That is, individuals with higher risks are more likely to be sampled in proportion to the risks themselves. Specifically, if the probability density of risks among these carriers is \( q(r) \), then

\[
q(r) = \frac{r p(r)}{\int r p(r) dr} \tag{1}
\]

Let us denote the mean risk of this distribution by \( \mu_c \), where the subscript \( c \) denotes the use of case probands. It can be shown that

\[
\frac{\mu_c}{\mu} = 1 + \frac{\nu^2}{\mu^2} \tag{2}
\]

where \( \nu^2 \) is the variance of the risks in the population. Thus, the mean risk in carriers identified from case probands is greater than the mean risk in carriers identified from control probands, and equation 2 shows that the square of the coefficient of variation determines the degree by which the mean risk (penetrance) is inflated. That is, the greater the variation in risk among carriers in the population, the greater the bias.

It is important to clarify that the risk distributions correspond to the inherent risks in the probands themselves. Because penetrance estimates are based on cancer incidence in relatives of probands, size bias occurs because of the impact of the additional risk factors that aggregate within families, e.g., genetic variations other than the gene under investigation (e.g., BRCA1 or BRCA2) or shared environmental risk factors. Furthermore, when we apply logic similar to that of Wacholder et al. (12), the upward bias in first-degree relatives of case probands who are carriers will be approximately half that of the probands themselves, as defined by equation 2.
The size-biased sampling paradigm characterized by equation 1 is fundamental to the case-control methodology that has formed the basis of cancer epidemiology research for approximately 50 years, since the work of Cornfield (23). That is, all cancer risk factors (e.g., smoking history, reproductive history, or others) are size-biased in a series of population-based incident case subjects in exactly the manner of equation 1. For example, let us consider any binary exposure with a prevalence of $p$ and relative risk of $\psi$. The exposed and nonexposed case subjects will be sampled in proportion to the product of the (relative) risks and the population prevalences, i.e., in the proportions $p\psi$ and $(1 - p)$, respectively. Thus, the prevalence of the exposure among case subjects is $q = p\psi/(1 - p + p\psi)$. The odds ratio, calculated from the cross-product of the case-control status and the exposure status, is thus $q(1 - p)/p(1 - q) = \psi$. This paradigm allows us to estimate the relative risk associated with any candidate risk factor but does not permit us to directly estimate absolute risks without the use of extraneous data on incidence to anchor the estimation (e.g., from cancer incidence registries). In other words, in identifying an incident case subject, we identify an individual who has been subject to a selection mechanism that affects all factors that influence risk. Thus, incident case subjects can be used for estimating absolute risks only if appropriate adjustments are made for the influences of all risk factors, both known and unknown. This concept has related implications for the interpretation of data on the incidence rates of second primary cancers, an issue that has been examined in detail for melanoma (24).

**RESULTS**

Eight studies of breast cancer penetrance of BRCA1 and BRCA2 mutations in case patients unscreened on the basis of family history of cancer have been published, and the results are summarized in Table 1. For most of these studies, the penetrance was reported by age 70 years, except for studies by the Anglian Group (16) and Risch et al. (17), who reported penetrance by age 80 years. The design of these studies varied in a number of important ways. Most studies used case probands exclusively, with the exception of Struwing et al. (3), who used both case and control probands. The studies by Antoniou et al. (18) and Risch et al. (17) used women diagnosed with ovarian cancer as case probands. Some studies were limited to young probands [Hopper et al. (19) and the Anglian Group (16)]. In Struwing et al. (3), Thorlacius et al. (20), and Warner et al. (21), only known founder mutations were analyzed, and in other studies, more extensive sequencing was performed. Some studies analyzed only first-degree relatives, and others analyzed the entire available pedigrees (16,18). The sample sizes were variable, and the studies frequently had only a few carrier probands, leading to substantial statistical variation in the penetrance estimates, as indicated by the wide confidence intervals. The last study reported in Table 1 [Satagopan et al. (25)] is radically different in design from the others, in that family data were not used to estimate penetrance. In this study, the age-specific odds ratios for breast cancer associated with BRCA1 and BRCA2 mutations were estimated separately, as in a case-control study, and the penetrance was calculated based on the assumption that overall incidence rates in the Jewish population and in the whole U.S. population are the same. Thus, this study was not subject to the bias induced by size-biased sampling.

Given the variability in study design, the considerable variation in estimated penetrance is not surprising. Despite the variability, the results suggest the following conclusions: First, the penetrance of BRCA2 mutations appears to be lower than the penetrance of BRCA1 mutations, on the basis of the relatively low estimates for the penetrance of BRCA2 mutations in the studies by Thorlacius et al. (20), Warner et al. (21), Risch et al. (17), and Satagopan et al. (25), although the relatively small study conducted in England [Anglian Group (16)] found the opposite result. Second, the impact of size-biased sampling in most of the studies should be reflected in lower estimates in the

<table>
<thead>
<tr>
<th>Study</th>
<th>Gene(s)</th>
<th>Mutation carriers</th>
<th>Penetration (95% CI), %*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Struwing et al. (3)</td>
<td>BRCA1/BRCA2</td>
<td>27 case probands†</td>
<td>56 (40 to 73) at age 70 y</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62 control subjects</td>
<td></td>
</tr>
<tr>
<td>Thorlacius et al. (20)</td>
<td>BRCA2§</td>
<td>69 case patients (male or female)</td>
<td>37 (22 to 54) at age 70 y</td>
</tr>
<tr>
<td>Warner et al. (21)</td>
<td>BRCA1/BRCA2</td>
<td>48 case patients</td>
<td>60 (—) at age 70 y (BRCA1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28 (—) at age 70 y (BRCA2)</td>
<td></td>
</tr>
<tr>
<td>Hopper et al. (19)</td>
<td>BRCA1/BRCA2</td>
<td>18 case patients‖</td>
<td>40 (15 to 65) at age 70 y</td>
</tr>
<tr>
<td>Anglian group (16)</td>
<td>BRCA1/BRCA2</td>
<td>24 case patients‖</td>
<td>48 (7 to 82) at age 80 y (BRCA1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>74 (7 to 94) at age 80 y (BRCA2)</td>
<td></td>
</tr>
<tr>
<td>Antoniou et al. (18)</td>
<td>BRCA1</td>
<td>12 case patients#</td>
<td>45 (22 to 76) at age 70 y**</td>
</tr>
<tr>
<td>Risch et al. (17)</td>
<td>BRCA1/BRCA2</td>
<td>60 case patients#</td>
<td>68 (—) at age 80 y (BRCA1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>“No excess risk” (BRCA2)</td>
<td></td>
</tr>
<tr>
<td>Satagopan et al. (25)</td>
<td>BRCA1/BRCA2</td>
<td>79 case patients††</td>
<td>46 (31 to 80) at age 70 y (BRCA1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>62 control subjects</td>
<td>26 (14 to 50) at age 70 y (BRCA2)</td>
</tr>
</tbody>
</table>

*CI = confidence interval; — = CIs were not reported in these studies.
†Genotyped for mutations common in Ashkenazi Jews: 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2.
‡Case probands include survivors of breast or ovarian cancer.
§Genotyped for Icelandic founder mutation: 999del5.
¶Case patients with breast cancer diagnosed at younger than 40 years of age.
#Patients with breast cancer diagnosed at younger than 55 years of age.
**Model-dependent “best estimate” that does not use kin-cohort methodology.
††Hospital-based patients with incident cases of cancer. The same control subjects were used here and by Struwing et al. (3). Family data were not used to estimate penetrance.
study by Satagopan et al. (25), which was not subject to this bias and, to a lesser extent, in the study by Struwing et al. (3) in which many of the probands were control probands. The penetrance estimates of 46% for BRCA1 and 26% for BRCA2 from Satagopan et al. (25) are, indeed, generally lower than those of the other studies, although the combined estimate of 56% in Struwing et al. (3) is surprisingly high.

**DISCUSSION**

The prevalences of risk factors in incident cases reflect the degree to which these factors affect risk. Thus, even among carriers of BRCA mutations, the prevalences of other risk factors will be higher in case patients than in control subjects. This situation will lead to bias if the incidence of cancer in relatives of case probands is used to estimate lifetime risks of cancer, as in the kin–cohort design (12) or the genotyped proband design (15), because the presence of any additional genetic risk factors will elevate risks for the relatives. At present, one can only speculate about the likely degree of bias, even for a well-studied problem, such as the penetrance of BRCA1 and BRCA2 mutations in breast cancer. The extent of bias can be inferred indirectly in various ways. First, one can estimate penetrance without using family data. This strategy was used by Satagopan et al. (25), who derived relative risk estimates conventionally by comparing case patients and control subjects without regard to information about relatives and who obtained lower penetrance estimates than the family-based studies reported in Table 1. Second, one can obtain penetrance estimates in studies using control probands. This approach was used by Struwing et al. (3), whose study consisted mostly of healthy control probands; indeed, the majority of the mutation carriers were control probands. Thus, it is surprising that the penetrance estimate for this study is somewhat high relative to the studies that used case probands. However, the risk of breast or ovarian cancer in noncarriers at age 70 years was observed to be 13%–14% in this study (12). When data from the Surveillance, Epidemiology, and End Results (SEER)1 program are examined, the overall risk of breast cancer at age 70 years should be 8%, and it should be about 9% for breast and ovarian cancer combined, suggesting that the volunteers in the study by Struwing et al. (3) were inadvertently selected from a higher-risk pool. By contrast, in the only other study that contained similar data on the risk in noncarriers, Warner et al. (24) found that observed penetrance at age 70 years for the kin of the noncarrier control probands is approximately 8%, which is equivalent to the SEER rate.

Another indication of the likely degree of bias when using family data in case probands to estimate penetrance comes from the observations that there is far more familial aggregation of breast cancer than can be explained solely by mutations in BRCA1 and BRCA2. Family members of case probands routinely have much higher risks than the family members of control probands, supporting the basic thesis of equation 1—that all factors that influence risk are preferentially sampled in case patients in proportion to their effects on risk. For example, Claus et al. (26), using data from the Cancer and Steroid Hormone Study, identified those relatives of case probands who were very likely to not carry BRCA mutations (91.7% of case probands and 94.9% of control probands) and showed that the familial risk in such women was substantially higher than that in relatives of the control probands. In addition, in one of the most comprehensive studies using linked data from twin registries and cancer registries in Scandinavia, Lichtenstein et al. (27) showed dramatically increased risks for cancer in twins of case patients compared with that in twins of control subjects for every cancer site with sufficient data for the study. For female breast cancer, the relative risk for breast cancer in monozygotic twins was found to be 5.2 (95% confidence interval [CI] = 3.7 to 7.4), and the relative risk for breast cancer in dizygotic twins was 2.8 (95% CI = 2.1 to 3.8), results that may disguise a strong negative association with age at diagnosis (28).

In the studies by Lichtenstein et al. (27), relatives of case patients were compared with relatives of control subjects, without knowledge of individual carrier status or any other risk factors. The contributions of any risk factor to these estimated familial relative risks are greatly attenuated compared with the fundamental relative risks induced by the risk factors (29). Thus, these familial relative risks represent the presence of a substantial genetic component to risk variation. Hopper and Carlin (30) have studied this phenomenon in detail by use of a normally distributed risk factor that is classified into quartiles of risk, showing that a genetic factor that is associated with a relative risk of 20 when the upper and lower quartiles are compared will lead to a familial relative risk of only 3.5 in monozygotic twins and 1.9 in dizygotic twins. In fact, for a known binary risk factor, the contribution to the familial relative risk in monozygotic twins can be calculated by the expression (1 – p + ψp2)/(1 – p + ψ)2, where p is the carrier prevalence and ψ is the relative risk (28). For dizygotic twins or their first-degree relatives, the contribution of a binary risk factor to the relative risk should be approximately half that calculated by this expression. The carrier prevalence of BRCA1 and BRCA2 mutations is believed to be in the range of 0.1%–0.4% in outbred Western populations, such as the Scandinavian population studied by Lichtenstein et al. (27). Peto et al. (31) estimated the prevalence in Britain at just over 0.1%. Even if the relative risk were in the range of 10–20, the contribution of BRCA1 and BRCA2 mutations to the relative risk for monozygotic twins would be, at most, in the range of 1.30–2.24 (or in the range of 1.15–1.62 for dizygotic twins or their first-degree relatives). Thus, the bulk of the genetic variation in the risk of breast cancer remains to be explained, which is a conclusion also reached by other investigators (31,32). The presence of additional unknown genetic variants that substantially influence risk argues against the possibility that the risk of breast cancer shows little or no variation among carriers of BRCA1 and BRCA2 mutations. As demonstrated by equation 2, however, the presence of such variation inevitably leads to bias in the estimation of penetrance when case probands are used.

The theory presented in this article indicates that control probands are an especially precious resource for studies of strong genetic risk factors, because all of the highly penetrant cancer gene mutations identified to date are rare in the population. In the absence of control probands who are mutation carriers, kin–cohort approaches that use case probands are a pragmatic strategy, but methodologic work on statistical techniques for correcting the bias caused by size-biased sampling is needed. Chatterjee and Wacholder (33) have developed an analytic method that recognizes and endeavors to estimate the residual intrafamily correlation that is the consequence of size-biased sampling. However, the method is designed to estimate residual correlation rather than to adjust the estimates of penetrance. Simulation studies by both Gail et al. (34) and Chatterjee and Wacholder (33) show that the presence of heterogeneity leads to a positive
bias in the estimate of penetrance. In both of these investigations, the case patients were not generated in a size-biased manner with respect to the underlying risk, and so the biases observed must be underestimates of the true biases. Furthermore, the authors used case–control sampling fractions ranging from 10% (34) to 50% (33) when, as we have seen, most investigations of the penetrance of major cancer gene mutations use essentially 100% case probands. It may not be possible to devise a reliable statistical method for correcting the bias in studies that use only familial aggregation of cancer in relatives of incident case patients. The more appropriate use of these data may be to estimate the relative risks associated with the mutations, even while recognizing that this strategy is itself susceptible to bias caused by a decrease in the observed relative risks when unmeasured covariates that influence risk are involved (35,36). The use of a conditional likelihood approach may provide a robust analytic strategy in this setting (37). These relative risks can then be used to infer the penetrance of the mutation by using known age-specific incidence rates as a benchmark, such as in the approach of Satagopan et al. (25).

Finally, an important thesis of this article is that substantial heterogeneity exists in the risk for cancer among individuals in the population. The notion that the presence of a germline BRCA1 or BRCA2 mutation in a woman completely defines her risk for breast cancer is probably far from the truth and, in fact, numerous unknown genetic risk modifiers are likely to exist. Thus, a woman who is identified as a carrier and who also has a strong family history of breast cancer is likely to possess a much higher risk for breast cancer than a carrier with no known family history of the disease. These differences should be reflected in the tools used by genetic counselors to predict risk. However, refinement of risk prediction methods such as the Gail model (38) to encompass detailed family history of breast cancer and carrier status (if known) is a major methodologic challenge, in part because of the problems outlined in this article.

References


NOTES

1 Editor's note: SEER is a set of geographically defined, population-based, central cancer registries in the United States, operated by local nonprofit organizations under contract to the National Cancer Institute (NCI). Registry data are submitted electronically without personal identifiers to the NCI on a biannual basis, and the NCI makes the data available to the public for scientific research.

I am grateful to Robert Elston, Mitchell Gail, John Hopper, and Sholom Wacholder for constructive feedback on an early version of this article.

Manuscript received December 11, 2001; revised June 11, 2002; accepted June 25, 2002.