Different, Not Biased

To the Editor:

In their paper “Bias in using family history as a risk factor in case-control studies of disease,” Khoury and Flanders have shown that the odds ratio (OR) for a positive family history, calculated from viewing the data from a case-control design perspective (that is, exposure = disease in one or more relatives), is a function of the number of relatives of cases and controls. They compared this measure with the relatives’ risk ratio (RRR), a measure of familial aggregation calculated from viewing the data as a cohort study of relatives (that is, exposure = case/control status of proband), and showed that the two estimates differed under varying conditions.

They have, however, interpreted the difference between OR and RRR as representing “bias”—note the first word of their title. In particular, they claim that the OR is an overestimate. In my opinion, this is an incorrect use of the term bias. (Epidemiologists often use the word “bias” to reflect design issues, such as differential recall, but that does not appear to be the concept being alluded to by Khoury and Flanders.)

Last2 defines bias as “any effect at any stage of investigation or inference tending to produce results that depart systematically from the true values.” In statistical parlance, an estimator, $\hat{\theta}$, is biased if its expected (average) value differs from that of the population parameter it is purportedly estimating, $\theta$. In the situation under discussion, there are two estimators directed at different entities; $\hat{\theta}_1 = OR$ is estimating the risk to an individual associated with the existence of one or more affected relatives ($\theta_1$), whereas $\hat{\theta}_2 = RRR$ is estimating the risk to an individual associated with just one relative being affected ($\theta_2$). OR is only “biased” if it is thought to be an estimator of $\theta_2$, which it clearly is not (unless there is always exactly one relative in the exposure set). Part of the problem probably lies with the authors not differentiating between estimators and parameters—an unfortunate and common practice in epidemiology. The issue of any bias per se of these two estimators was not addressed in the paper. It is well known that the sample odds ratio is an unbiased estimator of the population risk ratio, rate ratio, or odds ratio, depending on the method used to sample controls.

Pearson’s and Kendall’s correlation coefficients are both measures of association, but we do not talk of one of them as being a biased estimator because it gives different values from the other. Similarly, OR and RRR are both measures of familial aggregation for a disease, and there is no compulsion that the estimators should be equivalent, especially when one looks closely at their definitions.

The RRR may well be preferred to OR for a number of reasons. (Note, however, that the data may not be in a form from which RRR can be calculated, and that, as its standard error estimates are typically based on independence within exposure sets, they will be underestimates by an amount depending on the strength of familial aggregation.) RRR attempts to make more use of the information in the data, and the parameter it estimates has an unambiguous meaning, whereas OR is difficult to interpret if it is calculated from a sample of (case and control) families of varying sizes. It is not justified, however, to consider that an estimator based on using family history as a risk factor in case-control studies is biased, in either the statistical or epidemiologic sense of the word.

John L. Hopper
Australian National Health and Medical Research Council Twin Registry, Department of Public Health and Community Medicine, University of Melbourne, 200 Berkeley Street, Carlton, Victoria 3053, Australia (address for correspondence)

References


The Authors’ Reply:

While Dr. Hopper’s point is well taken, the issue is mostly semantic. In many epidemiologic studies, positive family history odds ratio (OR) is used to estimate relatives’ risk ratio (RRR). We have shown in our paper that OR is indeed a biased estimator of RRR. Family history OR, however, may be a consistent estimator of a parameter of familial aggregation, but this parameter is a complex function of the number of case and control relatives, their age distribution, and age at onset of disease. The whole point of our article is to illustrate how using family history OR in case-control studies can lead to varying and often overinflated measurement of familial aggregation, even when there is no case-control difference with respect to family size and age distribution. Most importantly, because family history is not a personal attribute or characteristic such as smoking or alcohol use, odds ratios derived from using family history cannot be compared across different studies and populations. We are glad that Dr. Hopper agrees with us that RRR is preferable to whatever parameter positive family history OR may be estimating. We also agree with Hopper that most case-control data may not be in a form from which RRR can be estimated. The latter point is more of a reason to caution epidemiologists against using positive family history as an easy and convenient way to measure familial aggregation.

Muin J. Khoury
Birth Defects and Genetic Diseases Branch, Centers for Disease Control and Prevention, MS F45, 1600 Clifton Road, Atlanta, GA 30333 (address for correspondence)

W. Dana Flanders
Rollins School of Public Health, Emory University, Atlanta, GA

Parental Age and Breast Cancer Mortality

To the Editor:

Holmberg et al1 found a higher relative hazard for death from breast cancer of 1.30 (95% confidence interval = 0.85–1.98)
in daughters born to mothers age 45 years or older, compared with mothers under the age of 20 years. Their finding was in accord with at least five other studies and, according to the authors, substantiates the claims for intrauterine influences of endogenous pregnancy estrogens on later breast cancer risk. They were embarrassed, however, by the absence of a clear linear trend for higher risk with increasing age of the mother at birth.

Close inspection into the maternal age categories after full adjustment for known confounders (that is, age at menarche, age at first pregnancy, and parity), however, indicates that the higher risks were present not only at the end, but at both extremes of reproductive age (that is, adolescence <20 years of age) and premenopause (≥45 years of age). Thompson and Janerich also found young maternal age at birth to be a risk factor. A U-shaped relation with maternal age is well established for menstrual and hormonal disturbances and, thus, for optimum maturation of the oocyte. A higher risk for breast cancer for twins also shifts our attention from intrauterine to intratubal or even intrafollicular influences. As in the case of clear cell adenocarcinoma, overripeness ovopathy should be considered as an alternate causal trigger for carcinogenesis, as pointed out by Witschi: "The persistence of an embryonic appearance of the cells, designated as asplasia or progressive failure of cells to differentiate, is the most constant effect produced by overripeness (of the egg); if combined with considerable growth, it leads to the formation of neoplasms."

Some anthropometric and reproductive characteristics, such as early menarche, late menopause, and menstrual disturbances, are related to breast cancer and have been shown to be connected with overripeness ovopathy. This factor, in concert with other components of the causal pathway, can be put to the test by examining the hazard for breast cancer in the younger maternal age categories and the other high-risk conceptions (for example, "spring" and "autumn" conceptions, short and long interpregnancy intervals, endocrinologic diseases, and so on).  

Piet Hein Jongbloet
Carla van Gils
Huub Straatman

Department of Epidemiology, University of Nijmegen, P.O. Box 9101, 6500 HB Nijmegen, the Netherlands (address correspondence to: Piet Hein Jongbloet)

References

Software for Optimal Matching in Observational Studies

To the Editor:

Individual matching can be done using either greedy or optimal algorithms. The latter method has the advantage of finding the set of case-control matches that are closest among all possible pairings. This method was recently illustrated in a report by Cologne and Shibata, in which they encouraged the implementation of optimal matching routines in major statistical computer packages. Here, we describe a SAS macro, %match, written to implement the optimal matching algorithm of Rosenbaum. The macro can also perform matching within risk sets as defined within nested case-control studies, and it can be used in matched cohort studies that match unexposed subjects to exposed subjects.

To determine which control is "best" for a particular case, it is necessary to provide some notation and a definition of distance between cases and potential controls. Let

\[
X^i = \{x^i_1, x^i_2, \ldots, x^i_p\}
\]

and

\[
X^0 = \{x^0_1, x^0_2, \ldots, x^0_p\}
\]

be the vector of \(p\) matching variables for the \(N\) cases and \(M (\geq N)\) potential controls, respectively. Let \(D_{ij}\) be "distance" between the \(i\)th case and the \(j\)th potential control. Choices for \(D_{ij}\) include the Euclidean distance (using standardized \(X^i\)'s), Mahalanobis distance, the sum of the absolute (or squared) difference in the ranks of the \(X^i\)'s, or the absolute (or squared) difference in propensity or "balancing" scores.  

For this macro, we define the distance as:

\[
D_{ij} = \sum_{k=1}^{p} \left| x^i_k - x^0_k \right| \cdot W_k,
\]

where \(W_k\) is an arbitrary nonnegative weight associated with matching variable \(k\).

This definition allows considerable flexibility. Instead of the actual \(X^i\)'s, standardized \(X^i\)'s, a propensity score, or the ranks of the \(X^i\)'s could be used. In addition, the weights allow the user considerable flexibility to modify the distance definition.

The optimal matching algorithm is designed to minimize the total \(D_{ij}\) over the set of all possible matchings. "Greedy" algorithms, on the other hand, involve proceeding sequentially through the list of cases, selecting the best available control at each step. Our experience and that of Rosenbaum suggest that optimal matching produces matched sets that are 5-10% "closer" than those defined with the greedy algorithm.

The SAS macro %match was written to provide an efficient and reproducible method of matching cases.