Commentary

Use and Misuse of Population Attributable Fractions

Beverly Rockhill, PhD, Beth Newman, PhD, and Clarice Weinberg, PhD

Introduction

How much of the disease burden in a population could be eliminated if the effects of certain causal factors were eliminated from the population? To address this question, epidemiologists calculate the population attributable fraction. As noted in a recent editorial in the Journal, population attributable fraction estimates can help guide policymakers in planning public health interventions.[1] Despite numerous articles on population attributable fraction estimation,[2-7] errors in computation and interpretation persist. In addition, in certain settings, the value of a population attributable fraction estimate may be questionable. This commentary considers computational and conceptual issues relevant to population attributable fraction estimation that are infrequently discussed elsewhere, with illustrations from the breast cancer literature.

Background

In 1953, Levin[8] first proposed the concept of population attributable fraction. Since then, the phrases "population attributable risk," "population attributable risk proportion," "excess fraction," and "etiologic fraction" have been used interchangeably to refer to the proportion of disease risk in a population that can be attributed to the causal effects of a risk factor or set of factors. Greenland and Robins[4] distinguish between excess fraction (what epidemiologists usually estimate when they compute "population attributable risk" or "population attributable fraction") and etiologic fraction, which is not estimable without strong biologic assumptions. Our use of the term "population attributable fraction" corresponds to Greenland and Robins' (population) excess fraction.

The population attributable fraction is most commonly defined as the proportional reduction in average disease risk over a specified time interval that would be achieved by eliminating the exposure(s) of interest from the population while distributions of other risk factors in the population remain unchanged. This also can be interpreted as the proportion of disease cases over a specified time that would be prevented following elimination of the exposures, assuming the exposures are causal.

While population attributable fractions usually are estimated for single risk factors, they also can be estimated for groups of factors considered simultaneously. In this situation, a population attributable fraction estimates the proportional amount by which disease risk would be reduced if all of the factors were to be simultaneously eliminated from the population. The exposed group consists of those exposed to at least one of the factors. A population attributable fraction for a set of risk factors considered simultaneously is sometimes termed a summary population attributable fraction.

The preceding definitions show that the word "risk" in attributable risk is technically incorrect; it is more correct to speak of proportion or fraction of risk. For this reason, although the term "population attributable risk" is most commonly used, terms such as "population attributable risk propor-

Beverly Rockhill is with Channing Laboratory and Harvard School of Public Health, Boston, Mass. Beth Newman is with the Department of Epidemiology, University of North Carolina, Chapel Hill. Clarice Weinberg is with the National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Requests for reprints should be sent to Beverly Rockhill, PhD, Department of Epidemiology, CB #7400, University of North Carolina, Chapel Hill, NC 27599-7400.

tion" and "population attributable fraction" are more accurate.

Basic Computational Issues

The expression corresponding to the preceding definition of population attributable fraction can be written as

$$\frac{P(D) - \Sigma_C P(D|\bar{C}, \bar{E})\, P(C)}{P(D)}$$

where $P(D)$ is the average probability of disease in the population (containing both exposed and unexposed individuals) over a specified time interval and $\Sigma_C P\ (D|\bar{C}, \bar{E})$ $P\ (C)$ represents the marginal conditional probability of disease given no exposure, averaged over strata of other risk factors or confounders (C). Several formulas more commonly seen than the preceding one are used to estimate population attributable fractions. Some of these formulas are valid only under the assumption of no confounding of the exposure–disease association. Table 1 presents the most commonly seen computational formulas and discusses the limitations, if any, on the use of each formula. Several authors have provided derivations and detailed discussions of the various formulas.[5–7,9]

The Distributive Property of the Population Attributable Fraction

A property of the population attributable fraction that is not always appreciated by epidemiologists is what Wacholder et al. term the distributive property.[9,10] A population attributable fraction can be quantitatively partitioned, or distributed, into exposure-category-specific attributable fractions, which then sum to the population attributable fraction. A category-specific attributable fraction is the fraction of total disease risk in the population that would be eliminated if persons in only that specific exposure category were to be shifted to the unexposed group. When several risk factors are being considered simultaneously, the exposure categories arise from a complete cross classification of the risk factors under consideration.

A category-specific attributable fraction is estimated as

$$pd_i\left(\frac{RR_i - 1}{RR_i}\right)$$

where RR_i is the (adjusted) relative risk for the ith exposure category (relative to the unexposed stratum) and pd_i represents the proportion of total cases in the population arising from the ith exposure category. The category-specific attributable fraction for the unexposed group ($i = 0$) is 0, since the RR_i is 1.0 by definition. The sum of the category-specific attributable fractions is thus

$$\sum_i pd_i\left(\frac{RR_i - 1}{RR_i}\right),$$

which can be simplified to

$$1 - \sum_i \frac{pd_i}{RR_i}$$

(formula 5 in Table 1).

Important implications of the distributive property have been previously noted.[7,10,11] The population attributable fraction will increase with an increasingly inclusive definition of exposure, provided that each group added to the "exposed" segment has a relative risk greater than 1.0 (in comparison with the remaining unexposed group). However, there may be a loss of precision with a broad exposure definition, since the standard error of the population attributable fraction increases as the proportion of exposed cases and controls increases above 0.50.[12,13]

Errors in Computation

Perhaps as a result of the proliferation of computational formulas for population attributable fractions, errors in estimation are common. Probably the most common error is the use of adjusted relative risks in formula 3 (see Table 1).[14–19] The magnitude of bias resulting from this error will depend on the degree of confounding.

Another type of computational error that is likely to involve more substantial bias is illustrated in an article on poverty and mortality in the United States.[20] The authors inappropriately used formula 3 (see Table 1) to estimate a "weighted" population attributable fraction across strata of a third (nonexposure) variable; that is, they misapplied the stratum reference in formula 3. As a result, the published estimates of the fraction of all-cause US mortality attributable to poverty were overestimated by approximately a factor of three.[21]

Conceptual Issues

Summing Population Attributable Fraction Estimates

Some epidemiologists inappropriately sum single risk factor population attributable fraction estimates in an attempt to derive the total fraction of disease risk attributable to all of the factors. This strategy is rarely appropriate and will almost always yield a value larger than the correctly calculated summary population attributable fraction for all of the factors considered simultaneously. Walter[12] discusses the limited conditions under which individual population attributable fractions may be validly summed.

One corollary of the preceding discussion is that it is possible, albeit counterintuitive, that a set of individual population attributable fractions will sum to more than 1.0. An implication of the "multicausal" model under which most epidemiologists work is that a given case of disease can be prevented by eliminating any one of the necessary causal factors present. For a specific disease, therefore, population attributable fractions computed separately for different risk factors are not constrained to sum to 1.0 or less. This issue has sometimes led to inappropriate analyses. At least two papers have attempted to attribute cancer risk to a variety of life-style and environmental factors considered singly, and the authors forced the population attributable fractions for the single factors (including a catchall factor of "unknown cause") to sum to 1.0.[22,23]

Interpretation and Communication

Perhaps the most important aspect of population attributable fraction estimation is correct interpretation and communication. Consider an extensively cited paper devoted to population attributable fraction estimation for "established" breast cancer risk factors.[19] Seidman et al. estimated population attributable fractions of 0.21 in the 30 to 54-year age group and 0.29 in the 55 to 84-year age group for 10 breast cancer risk factors. Misinterpretations of the population attributable fractions presented in this paper have been common, both in the scientific and lay literatures.

The most frequent error involves equating the population attributable fraction with the proportion of cases having any risk factors: "Although various risk factors have been identified as causes of breast cancer, the fact remains that in 75% of all breast cancer no identifiable risk factor can be found."[24] This error was made in an article advising clinicians on patient education: "Only 21 per cent of the cancers occurring in women from 30 to 54 years of age and 29 per cent in the women over 50 could be attributed to one or more risk factors, meaning that the majority of cancers occur in women with no risk fac-

TABLE 1—Commonly Seen Formulas for Attributable Fraction Estimation

1.	$\frac{IP_t - IP_0}{IP_t}$	Empirical approximation of $\frac{P(D) - \sum_C P(D \mid \bar{C},\bar{E})\, P(C)}{P(D)}$ IP_t = cumulative proportion of total population developing disease over specified interval; IP_0 = cumulative proportion of unexposed persons who develop disease over interval. Valid only when no confounding of exposure(s) of interest exists. If disease is rare over time interval, ratio of average incidence rates I_0/I_t approximates ratio of cumulative incidence proportions, and thus formula can be written as $(I_t - I_0)/I_t$. Both formulations found in many widely used epidemiology textbooks.
2.	$\frac{p_e(RR-1)}{p_e(RR-1)+1}$	Transformation of formula 1. Not valid when there is confounding of exposure–disease association. p_e = proportion of source population exposed to the factor of interest. RR may be ratio of two cumulative incidence proportions (risk ratio), two (average) incidence rates (rate ratio), or an approximation of one of these ratios. Found in many widely used epidemiology texts, but often with no warning about invalidness when confounding exists.
3.	$\frac{\sum_{i=0}^{k}(p_i)(RR_i-1)}{1+\sum_{i=0}^{k}(p_i)(RR_i-1)} = 1 - \frac{1}{\sum_{i=0}^{k} p_i\,(RR_i)}$	Extension of formula 2 for use with multicategory exposures. Not valid when confounding exists. Subscript i refers to the ith exposure level. p_i = proportion of source population in ith exposure level, RR_i = relative risk comparing ith exposure level with unexposed group ($i = 0$). Derived by Walter[12]; given in Kleinbaum et al.[29] but not in other widely used epidemiology texts.
4.	$pd\left(\frac{RR-1}{RR}\right)$	Alternative expression. Produces internally valid estimate when confounding exists and when, as a result, adjusted relative risks must be used.[9] pd = proportion of cases exposed to risk factor. In Kleinbaum et al.[29] and Schlesselman.[30]
5.	$\sum_{i=0}^{k} pd_i\left(\frac{RR_i-1}{RR_i}\right) = 1 - \sum_{i=0}^{k}\frac{pd_i}{RR_i}$	Extension of formula 4 for use with multicategory exposures. Produces internally valid estimate when confounding exists and when, as a result, adjusted relative risks must be used. pd_i = proportion of cases falling into ith exposure level; RR_i = relative risk comparing ith exposure level with unexposed group ($i = 0$). See Bruzzi et al.[5] and Miettinen[9] for discussion and derivations; in Kleinbaum et al.[29] and Schlesselman.[30]

tors.”[25(p608)] Such statements reflect misunderstanding about the meaning of the population attributable fraction. The proportion of patients exposed to the considered risk factor(s) is different from the population attributable fraction. In the Seidman et al. study, the proportions of breast cancer patients who had at least one of the considered factors were 0.76 in the 30 to 54-year age stratum and 0.82 in the 55 to 84-year age stratum.

Seidman et al. may have contributed to misinterpretations with the wording of their conclusion: “Given our current understanding of breast cancer risk factors, we are unable to identify. . .the ‘causes’ of more than about one-quarter of all cases.”[19] An average population attributable fraction estimate of 0.25 across the two age strata means that 25% of the population risk of breast cancer would be eliminated if all 10 risk factors were to be eliminated from the population or, equivalently, that 25% of cases would be prevented following the risk factor eliminations. As just discussed, it does not mean that 25% of women who develop breast cancer will have one or more of the 10 risk factors; nor does it mean that epidemiologists can identify the cause(s) of breast cancer for a quarter of individuals with the disease. The population attributable fraction does not address probability of causation for a specific case of disease, nor does its estimation enable epidemiologists to discriminate between those cases caused by, and those not caused by, the risk factors under consideration.

A more recent report on population attributable fractions and breast cancer risk factors has similarly been misinterpreted. Bruzzi et al.[5] considered four established factors and estimated a population attributable fraction of 0.55. Referring to this estimate, a recent article included the following misstatement: “Another report estimates that 55 percent of breast cancers have one or more risk factors.” [26(p5)] In fact, as a result of the broad risk factor definitions used, 99% of the breast cancer cases in the Bruzzi et al. analysis involved one or more risk factors!

From a public health perspective, estimation of the population attributable fraction is of most use when the factor of interest is clearly causally related to the end point and when there is consensus that the exposure is amenable to intervention. However, many researchers use risk factors that are surrogates for susceptibility attributes that may be unmodifiable (e.g., ethnicity, family history of cancer), as well as factors that are preclinical markers of disease (e.g., history of benign breast biopsy). Some factors included in population attributable fraction estimations are surrogates for more proximate exposures (e.g., poverty, educational level, marital status). Obviously, breast cancer risk will not be reduced by denying women a college education or a breast biopsy or by ensuring that all women marry, assuming that more causally proximate exposures and behaviors remain the same; however, these points are rarely discussed by investigators. The practical and logical limitations of including unmodifiable attributes, potential disease markers, and surrogate factors in population attributable fraction estimation are not always recognized.

Another issue related to interpretation of a population attributable fraction concerns specification of the exposed group. When modifiable risk factors are being considered in order to prioritize public health intervention strategies, the exposure cut point should be chosen so that the "unexposed" level is realistically attainable by those in the exposed category. Otherwise, the population attributable fraction may have theoretic value but will be of little practical public health value. Related to this point is Rose's observation that, for many chronic diseases, susceptibility for any disease is rarely confined to a high-risk minority within the population.[27] More typically, the majority of cases arise from the mass of the population with risk factor values around the population average. For many chronic diseases, population attributable fractions can be made high only by defining risk factors in such a way that almost the entire population is labeled "exposed" or "at elevated risk." The unrealistic implication of such broad exposure definitions is that virtually everyone in the population will need to be "shifted" to the lowest exposure category. This was the implication for the estimate presented by Bruzzi et al.[5] It is also the case with the most recent estimate (0.41) of the summary population attributable fraction for three breast cancer risk factors[28]; the authors estimate that 90% of US women have one or more of these established risk factors for breast cancer, and thus this large proportion of the population will need to be shifted on one or more of the factors if the estimated reduction in breast cancer burden is to be achieved.

A final philosophical point concerns the common practice of equating the population attributable fraction with the proportion of disease cases that are "explained" by the risk factors. For instance, after computing their population attributable fraction of 0.41, Madigan et al. stated that their estimates "suggest that a substantial proportion of breast cancer cases in the United States are explained by well-established risk factors."[28] This use of the word "explain" is somewhat misleading, since many readers probably equate "explain" with "cause." According to the Madigan et al. data, nearly the entire population of women in the United States has at least one of the considered risk factors. Since the vast majority of such "exposed" women will not develop breast cancer, stating that such factors explain a large proportion of breast cancer risk seems euphemistic. As an extreme example, if an age of greater than 15 years is considered a risk factor in a population attributable fraction estimation, virtually all cases of breast cancer can be "explained," in the technical sense of explaining variation in rates between the exposed (those more than 15 years of age) and the unexposed (those 15 years of age or younger); however, to imply that being more than 15 years of age "causes" breast cancer is of no value. Authors who present population attributable fractions should communicate clearly what they mean when they use phrases such as "explained by" and "attributable to," because there is potential for confusion on the part of both scientific and lay readers.

Conclusion

Many public health researchers are interested in evaluating the potential population impacts of identified risk factors. For some of these evaluations, estimation of the population attributable fraction is appropriate and valuable. The assumptions underlying valid population attributable fraction estimation include the following: a causal relationship between the risk factors and disease; the immediate attainment, among those formerly exposed, of the unexposed disease risk following elimination of the exposures; and independence of the considered risk factors from other factors that influence disease risk so that it is possible to conceive of changing the population distributions of the considered factors only. Such assumptions are often not justified. Those who present population attributable fractions have a duty to ensure that estimates are correctly computed and that their limited meaning is correctly communicated, given the interest among researchers, clinicians, and the public in quantitative figures that attempt to summarize the state of etiologic knowledge about a disease. □

Acknowledgments

We thank Beth Gladen, Glinda Cooper, Bob Millikan, Bruce Levin, and the two anonymous reviewers for their thoughtful and very helpful reviews of the manuscript.

References

1. Northridge ME. Annotation: public health methods—attributable risk as a link between causality and public health action. *Am J Public Health.* 1995;85:1202–1203.
2. Coughlin SS, Benichou J, Weed DL. Attributable risk estimation in case-control studies. *Epidemiol Rev.* 1994;16:51–64.
3. Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Stat Med.* 1991;10:1753–1773.
4. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol.* 1988;128:1185–1197.
5. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol.* 1985;122: 904–914.
6. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. *Stat Med* 1982;1:229–243.
7. Walter SD. Effects of interaction, confounding, and observational error on attributable risk estimation. *Am J Epidemiol.* 1983;117:598–604.
8. Levin ML. The occurrence of lung cancer in man. *Acta Union International Contra Cancrum.* 1953;9:531–541.
9. Miettinen O. Proportion of disease caused or prevented by a given exposure, trait, or intervention. *Am J Epidemiol.* 1974;99:325–332.
10. Wacholder S, Benichou J, Heineman EF, Hartge P, Hoover RN. Attributable risk: advantages of a broad definition of exposure. *Am J Epidemiol.* 1994;140:303–309.
11. Hsieh C-C, Walter SD. The effect of nondifferential exposure misclassification on estimates of the attributable and prevented fraction. *Stat Med.* 1988;7:1073–1085.
12. Walter SD. The estimation and interpretation of attributable fraction in health research. *Biometrics.* 1976;32:829–849.
13. Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of the attributable fraction based on logistic models. *Biometrics.* 1990;46:991–1003.
14. Liao Y, Cooper RS, McGee DL, Mensah GA, Ghali JK. The relative effects of left ventricular hypertrophy, coronary artery disease, and ventricular dysfunction on survival among black adults. *JAMA.* 1995;273:1592–1597.
15. Donders GGG, Desmyter J, De Wet DH, Van Assche FA. The association of gonorrhoea and syphilis with premature birth and low birthweight. *Genitourin Med.* 1993;69:98–101.

16. Hankin JH, Zhao LP, Wilkens LR, Kolonel LN. Attributable fraction of breast, prostate, and lung cancer in Hawaii due to saturated fat. *Cancer Causes Control.* 1992;3:17–23.
17. Mavalankar DV, Gray RH, Trivedi CR. Risk factors for preterm and term low birthweight in Ahmedabad, India. *Int J Epidemiol.* 1992;21:263–271.
18. Stroffolini T, Chiaramonte M, Tiribelli C, et al. Hepatitis C virus infection, HBsAg carrier state and hepatocellular carcinoma: relative risk and population attributable risk from a case-control study in Italy. *J Hepatol.* 1992;16:360–363.
19. Seidman H, Stellman SD, Mushinski MH. A different perspective on breast cancer risk factors: some implications of the nonattributable risk. *CA.* 1982;32:301–312.
20. Hahn RA, Eaker E, Barker ND, Teutsch SM, Sosniak W, Krieger N. Poverty and death in the United States—1973 and 1991. *Epidemiology.* 1995;6:490–497.
21. Rockhill B, Weinberg C. Error in population attributable risk calculation. *Epidemiology.* 1996;7:453.
22. Higginson J, Muir CS. Environmental carcinogenesis: misconceptions and limitations to cancer control. *JNCI.* 1979;63:1291–1298.
23. Wynder EL, Gori GB. Contribution of the environment to cancer incidence: an epidemiologic exercise. *JNCI.* 1977;58:825–832.
24. Freeman HP, Wasfie TJ. Cancer of the breast in poor black women. *Cancer.* 1989;63: 2562–2569.
25. Love SM. Use of risk factors in counseling patients. *Hematol Oncol Clin North Am.* 1989;3:599–610.
26. Garfinkel L. Perspectives on cancer prevention. *CA.* 1995;45:5–7.
27. Rose G. Sick individuals and sick populations. *Int J Epidemiol.* 1985;14:32–38.
28. Madigan MP, Ziegler RG, Benichou J, Byrne C, Hoover RN. Fraction of breast cancer cases in the United States explained by well-established risk factors. *JNCI.* 1995;87: 1681–1685.
29. Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic Research.* Belmont, Calif: Lifetime Learning Publications; 1982:163.
30. Schlesselman JJ. *Case-Control Studies: Design, Conduct, Analysis.* New York, NY: Oxford University Press Inc; 1982.