

COMMENTARY

Apportioning causes, targeting populations and predicting risks: Population attributable fractions

Nilanjan Chatterjee¹ & Patricia Hartge^{2,3}

¹Biostatistics Branch; ²Office of Director, Epidemiology and Biostatistics Program; ³Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, D.H.H.S, USA

In this issue of the journal, Gustavsson and his co-authors describe a simple method of computing the population attributable fractions (*PAF*) associated with two causes and present a correct and simple way to estimate the variance. The *PAF*, which goes by various names, including etiologic fraction and population attributable risk, is the portion of the total burden of a disease in a population that should be ascribed to certain cause(s) of the disease. A moderately arcane parameter at first blush, the *PAF* deserves attention because it measures disease burden. It helps to describe the effect of different interventions on various target populations. It is often useful for building risk prediction models such as the Gail model [1]. Improvements in estimation, interpretation, and use of the *PAF* make an important contribution to epidemiology.

Gustavsson concentrates on simple bifurcation of the population into exposed and unexposed, but often we need to go further. To gauge the impact on risk of two exposures that occur together in the population, whether they interact biologically or merely coexist, we should compute a *PAF* for each as well as an overall *PAF*. The individual *PAF* should be computed by adjusting for the other risk factors: it measures the fraction of cases that would be eliminated by removing the single exposure from the population if the distribution of the other exposures were unchanged. A brief review of the mathematics is useful for subsequent calculations.

Based on the above definition of adjusted *PAF*, the *PAF* for an exposure *E* adjusted for a polytomous factor *C* with *K* levels (strata) can be defined as [2, 3]

$$PAF(E) = 1 - \sum_{k=1}^K P(C_k)P(D|\bar{E}, C_k)/P(D), \quad (1)$$

where $P(C_k)$ is the fraction of the whole population that is in the *k*th stratum and $P(D|\bar{E}, C_k)$ is the probability of the disease among unexposed subjects in the *k*th stratum. An alternative formula that can be useful for computing adjusted *PAF* from case-control data (both matched and unmatched) is given by [4]

$$PAF(E) = \sum_k \Pr(C_k|D)PAF_k(E), \quad (2)$$

where $P(C_k|D)$ is the fraction of all cases that are in the *k*th stratum and $PAF_k(E)$ is the *PAF* due to *E* in the *k*th stratum. In (2), $PAF_k(E)$ can be estimated from case-control studies using the standard formula

$$PAF_k(E) = \Pr(E|D, C_k) \frac{RR_k(E) - 1}{RR_k(E)}, \quad (3)$$

where $RR_k(E)$ is relative risk associated with *E* in stratum C_k . Neither of formulae (1) or (2) requires that the relative risk associated with *E* is constant across strata. Both of the formulae are applicable for adjustment of confounders as well as effect modifiers.

Gustavsson and co-authors use a hypothetical example involving two exposures *A* and *B* in their Table 1. Based on the formula (1) given above, we can compute the *PAFs* due to the exposures *A* and *B* individually (adjusted for each other) as 0.24 and 0.44, respectively. Based on this data, the authors report that the overall *PAF* due to *A* and *B* is 0.56. These calculations give several insights. First, the sum of *PAF* for the two individual risk factors exceeds the overall *PAF* due to the two risk factors together. Thus, in this example, the burden of the disease due to the two risk factors together cannot be partitioned into terms of the burdens of the disease due to the individual risk factors. Second, while a total of 56% of the cases could be eliminated by removing both *A* and *B*, as much as 44% of cases can be eliminated by removing *B* alone. Thus, in terms of determining public health policy for reducing burden of the disease from the population, it seems removal of *A* and *B* (assuming they are modifiable) both would yield only a modest benefit over removing *B* alone.

Gustavsson's real-world example considers lung cancer incidence attributable to exposure to asbestos and combustion products in Stockholm, Sweden (Table 2). Based on formula (2) and (3), we estimate the individual *PAF* for asbestos and combustion product to be 4.4 and 2.4%. The authors estimate the overall *PAF* due to asbestos and combustion product together to be about 6.9%. Unlike the hypothetical example, here, the total *PAF* due to the two risk factors together roughly equals the sum of *PAF* for the individual risk factors. That is, the total burden of the disease due to asbestos and combustion product together can be partitioned in terms of the burden of the disease due to the individual risk factors.

There are critical links between the parameters of *PAF*, relative risks and risk differences. In general, partitioning of the total *PAF* in terms of the *PAF* for the individual risk factors is possible if, but only if, the effects of the two risk factors on the disease risk are additive, that is, the excess risk due to exposure to both of two factors must equal to sum of the excess risks due to exposure to the individual factors [5]. In terms of relative risks, this condition for two exposures *A* and *B* can be stated as

$$RR(AB) = RR(\overline{AB}) + RR(A\overline{B}) - 1, \quad (4)$$

where $RR(\overline{AB})$, $RR(A\overline{B})$ and $RR(AB)$ denote the relative risks corresponding to the exposure groups \overline{AB} , $A\overline{B}$ and AB , respectively, in reference to the common baseline exposure \overline{AB} . In the example of lung cancer, it is easy to see from Table 2 that condition (4) is satisfied approximately for exposures to asbestos and combustion product ($2.25 \approx 1.62 + 1.67 - 1$).

In short, we can gain important insight if we compute the *PAF* for to the individual risk factors and compare them with the overall *PAF* for the combination of factors together. For categorical exposure variables, both the overall *PAF* and the adjusted *PAF* for the individual risk factors can be computed empirically without any model assumption on the joint relative risk parameters. For dealing with continuous exposures or exposures with large number of ordered categories, on the other hand, model based methods may be needed for obtaining stable estimate of these different types of *PAF* [6, 7].

We suspect *PAF* will gain new uses as epidemiologists turn increasing attention to the interplay of genetic (*G*) and environmental (*E*) factors in the etiology of complex diseases. Obviously, the overall *PAF* due to *G* and *E*, defined as the fraction of cases that would be eliminated if both the exposure *G* and *E* are removed from the population is not directly relevant because the genetic exposure is not modifiable. Still, appropriate *PAF* can serve public health purposes by revealing the possible impact of an intervention on *E*. For population based intervention that impacts the population as a whole, such as banning a carcinogenic agent, the relevant measure of impact of the ban is the *PAF* for *E* adjusted for *G*; this estimates the fraction of cases that will be eliminated if *E* is removed from the population without affecting *G*. Furthermore, if *G* and *E* are roughly independently distributed in the population, adjustment for *G* is not necessary for computation of the *PAF* due to *E*, irrespective of whether *G* is an effect modifier of *E* or not. This result, although generally not known, can be derived from formula (1) by noting that under *G*-*E* independence assumption $\Pr(G) = \Pr(G|\overline{E})$, that is, the probability of exposure to *G* in the whole population is same as that in the subpopulation of subjects who are not exposed to *E*.

For intervention aimed at an individual, such as counseling for smoking cessation, however, *PAF* is not directly relevant; the correct measure of impact is given by the excess absolute risk for an individual subject due to the exposure *E*. In this context, an important question is whether genetic information can be used to target subjects for which elimination of *E* would be most beneficial. The net benefits of prevention due to elimination of the exposure *E* for subgroups of subjects with and without the genetic exposure *G* can be measured by the risk differences $RD_G(E) = \Pr(D|E, G) - \Pr(D|\overline{E}, G)$ and $RD_{\overline{G}}(E) = \Pr(D|E, \overline{G}) - \Pr(D|\overline{E}, \overline{G})$, respectively [8].

We also suspect the *PAF* will see expanded use because risk prediction models such as Gail model [1] are often built on it. Occasionally, risk prediction models can be derived from cohort studies in which the exposure-specific absolute risks of the disease can be estimate directly. For case-control studies, the baseline disease probabilities not being known, the absolute risks of the disease cannot be computed directly. An overall *PAF* due to all risk factors together can be computed from case-control studies that enroll a fully representative case group. If the probability of the disease ($\Pr(D)$) in the underlying population can be estimated from other sources of data, the baseline disease probability $\Pr(D|\overline{E})$, defined as the probability of the disease who are not exposed to any of the risk factors, can be estimated using the relationship

$$(1 - PAF) \Pr(D) = \Pr(D|\overline{E}).$$

Using this baseline, the prediction model yields absolute risk estimates for various combinations of exposures.

We share the view of Gustavsson and co-authors that measures of *PAF* involving multiple exposures deserve attention as they will play important roles in epidemiology and public health.

References

1. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989; 81(24): 1879-1886.
2. Whittemore AS. Statistical methods for estimating attributable risk from retrospective data. *Stat Med* 1982; 1(3): 229-243.
3. Whittemore AS. Estimating attributable risk from case-control studies. *Am J Epidemiol* 1983; 117(1): 76-85.
4. Walter SD. The estimation and interpretation of attributable risk in health research. *Biometrics* 1976; 32(4): 829-849.
5. Walter SD. Effects of interaction, confounding and observational error on attributable risk estimation. *Am J Epidemiol* 1983; 117(5): 598-604.

6. 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(5): 904-914.
7. Benichou J. A review of adjusted estimators of attributable risk. *Stat Meth Med Res* 2001; 10(3): 195-216.
8. Wacholder S, Weinberg CR. Selecting subpopulations for intervention. *J Chronic Dis* 1986; 39(7): 513-519.

Address for correspondence: N. Chatterjee, 6120 Executive Blvd, EPS 8038, Rockville MD 20852, USA
E-mail: chattern@mail.nih.gov