A heuristic approach to the formulas for population attributable fraction

J A Hanley

Abstract

***Background*—As the definitional formula for population attributable fraction is not usually directly usable in applications, separate estimation formulas are required. However, most epidemiology textbooks limit their coverage to Levin's formula, based on the (dichotomous) distribution of the exposure of interest in the population. Few present or explain Miettinen's formula, based on the distribution of the exposure in the cases; and even fewer present the corresponding formulas for situations with more than two levels of exposure. Thus, many health researchers and public health practitioners are unaware of, or are not confident in their use of, these formulas, particularly when they involve several exposure levels, or confounding factors.**

***Methods/Results*—A heuristic approach, coupled with pictorial representations, is offered to help understand and interconnect the structures behind the Levin and Miettinen formulas. The pictorial representation shows how to deal correctly with several exposure levels, and why a commonly used approach is incorrect. Correct and incorrect approaches are also presented for situations where estimates must be aggregated over strata of a confounding factor.**

(*J Epidemiol Community Health* 2001;**55**:508–514)

Department of Epidemiology and Biostatistics, McGill University, 1020 Pine Avenue West, Montreal, Quebec, Canada H3A 1A2

Correspondence to:
Dr Hanley
(James.Hanley@McGill.CA)

Accepted for publication
14 January 2001

The population attributable fraction (AF_p) is defined[1] (page 295) as "the fraction of all cases (exposed and unexposed) that would not have occurred if exposure had not occurred." This fraction can be estimated using two equivalent formulas, based on the distribution of exposure in the population,[2] or in the cases.[3] Most textbooks deal only with the exposure in the population and do not consider confounding factors. Many of them focus on deriving the formula algebraically or minimising the number of computational steps, thereby providing limited insight into the structure of the formula. Only a few texts elaborate on the "case-based" formula. As a result, although it is increasingly used to derive estimates of AF_ps from complex data,[4–8] the case-based formula is less widely known and less well understood.

Likewise, despite the long existence[3 9] of the corresponding AF_p formulas for more than two levels of an exposure of interest, and despite the fact that three advanced textbooks[10–12] do present and even illustrate them, many authors seem to be unaware of them. This author has recently encountered three pre-publication examples where, with multiple exposure levels, the AF_p was calculated incorrectly. Table 1 shows a published example[13] of this same error.

Stratification and—increasingly—regression models are used to provide confounder adjusted rate ratio (RR) estimates as inputs to the calculation of AF_ps. As textbooks do not discuss such situations, and understanding of first principles is limited, the AF_p is often miscalculated in such instances too.[14] Indeed, in addition to mishandling a trichotomous exposure, the above cited report[13] also fails to correctly incorporate the adjusted RR into the AF_p calculation.

The primary aim of this article is to promote understanding of the AF_p formulas for a polytomous exposure. To do so, the article begins with the more familiar all or none exposure. A numerical example and a diagram allow the Levin and the Miettinen formulas to be understood directly from first principles, without algebra. This heuristic approach provides a foundation from which to extend the AF_p formulas correctly to polytomous exposure data, and to data stratified on a confounding variable.

Population (or population time) at risk and cases will be denoted by the letters P and C respectively. The fractions of the population (or population time) in the various exposure categories are denoted by "population fractions" (PFs), while the distribution of exposure in the cases is denoted by "case fractions" (CFs).[3] The terms overall and population attributable fraction are used interchangeably. Given the difficulties[15] of interpreting it as a true "aetiological" fraction, particularly when a long time span and competing risks can substantially change denominators, the AF_p is simply regarded as an "excess" fraction.[15]

*Table 1 Example of incorrect and correct calculation of population attributable fraction for trichotomous "Exposure"**

			Calculation of population attributable fraction (%)	
Exposure level	*% of population*	*RR*	*Incorrect†*	*Correct‡*
Low	50	1.0	—	—
Moderate	30	1.4	10.7	9.5
High	20	1.7	12.3	11.1
All	100		23.0	20.6

*Adapted from table 4 of reference[13] by rounding the reported exposure percentages and RRs. †$(0.4 \times 0.3) / (1 + 0.4 \times 0.3)$ and $(0.7 \times 0.2) / (1 + 0.7 \times 0.2)$. ‡$(0.4 \times 0.3) / (1 + 0.4 \times 0.3 + 0.7 \times 0.2)$ and $(0.7 \times 0.2) / (1 + 0.4 \times 0.3 + 0.7 \times 0.2)$.

All or none exposure

The exposed and unexposed categories are denoted by 1 and 0 and the ratio of the event rates in these two categories as RR: 1.

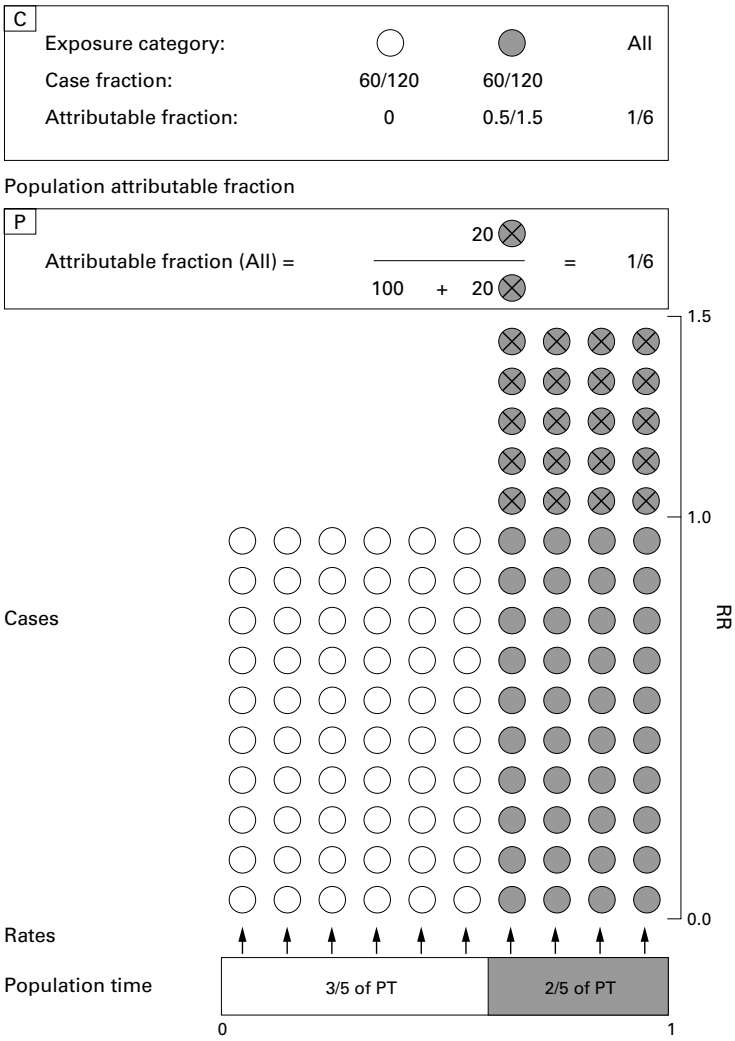

Figure 1 Population (that is, "overall") attributable fraction (AF_p) when exposure is all(1) or none(0), based on distribution of exposure in population (P) and in cases (C).

The PF_0 = 3/5th of the population time (PT) in the unexposed category is shown in white at the bottom of the diagram, and the PF_1 = 2/5th of the population time (PT) in the exposed category is shown shaded. Relative rates in these categories are RR=1 and 1.5. The corresponding numbers of cases arising from the two exposure categories are shown as white and shaded circles. The number of "excess" cases is depicted as shaded circles marked with an "X" (for "excess").

In the classic AF_p structure (summarised in inset "P", with numbers of cases scaled up by 100), the total number of cases is divided into the lower (square) array denoting the number of expected cases (circles without an "X", irrespective of exposure category) and the upper rectangular array denoting the number of excess cases. As the square array of "expected" cases has a base of width 1, and a height of RR = 1, it has an area of 1, representing one expected case. With these relative horizontal and vertical scales, the rectangle of excess cases has a base of PF_1 = 2/5 and a height of RR—1 = 1.5—1 = 0.5, so that its area (the "number of excess cases per 1 expected case") is (RR—1) × PF_1 = 0.5 × 2/5 = 0.2. This 0.2 is 1/6th of the overall total of 1.2 cases.

In the case-based AF_p structure (summarised in inset "C"), the total number of cases is divided first by exposure category. None of the unexposed cases (white circles) are "excess" cases. Of the exposed cases (shaded circles, constituting 1/2 of all cases), only a fraction represent excess cases. As 1 of every 1.5 exposed cases is "expected", some 0.5 of the 1.5, or 1/3rd, are excess. Thus, 1/3 of 1/2 = 1/6 of all cases are excess cases.

FORMULAS

Classic (Levin) structure for AF_p, based on distribution of exposure in population

Denote by PF_1 the proportion (or fraction) of the total population time in the exposed category, and by PF_0 the proportion in the unexposed category. The most popular[1 11 16–18] formula for AF_p is Levin's original version. Levin began by defining the AF_p: its denominator is the rate (or number of cases) in the overall population, and its numerator is the difference between this and the one that would prevail if all of the person time were in the unexposed category. From this, he algebraically derived the estimating formula

$$AF_p = \frac{\{RR-1\} \times PF_1}{1+\{RR-1\}PF_1} \qquad [1P]$$

Attributable fractions for specific exposure categories.

The case-based version uses as one of its inputs the "attributable fraction in the exposed", namely

$$\frac{RR-1}{RR}$$

This is a specific AF, as it restricts attention to exposed cases. To emphasise this specificity, we label it AF_1. The under-appreciated fact that the "attributable fraction in the *un*exposed" is 0 becomes important later, and so we label the AF specific to cases in that category as $AF_0 = 0$.

(Miettinen) structure for AF_p, based on distribution of exposure in cases

The case-based version uses as its other input the number of exposed cases, expressed as a fraction of the overall number of cases. Denote this fraction as the "case fraction",[3] CF_1. Then the case-based formula for AF_p is

$$AF_p = \frac{RR-1}{RR} \times \frac{\text{number of exposed cases}}{\text{overall number of cases}},$$

or in the notation used here,

$$AF_p = AF_1 \times CF_1. \qquad [1C]$$

NUMERICAL EXAMPLE

Suppose, as is depicted in figure 1, that PF_1 = 2/5th of the population time (PT) is in the exposed category. Although the AF_p involves relative rather than absolute rates, suppose—for concreteness—that the event rates in the exposed and unexposed categories are 1.5 and 1.0 cases per 10^4 PT units, so that the RR = 1.5, and the rate difference = 0.5 cases per 10^4 PT units. Suppose further that the total population time is 10^6 PT units.

Substitution of PF_1 = 2/5 and RR—1 = 0.5 into formula [1P] yields

$$AF_p = \frac{0.5 \times (2/5)}{1+0.5 \times (2/5)} = \frac{0.2}{1+0.2} = \frac{1}{6}.$$

THE NUMBERS AND STRUCTURES BEHIND THE FORMULAS

Conceptually, the Levin formula directly divides the total number of cases into "expected" cases—those that would occur even if all of the PT were in the unexposed category—and "excess" cases. With a total of 10^6 units of PT, there are $10^6 \times (1.0 \times 10^{-4})$ = 100 "expected" cases. Some 2/5th of the overall 10^6 PT units are in the exposed category, where the excess rate is 0.5 per 10^4 PT units. The product of this PT and the excess rate in this category is 20 "excess" cases. These 20 represent 1/6th of the overall total of 100 + 20 = 120 cases. Note that the 20 can also be represented as the "observed−expected" number, while the 120

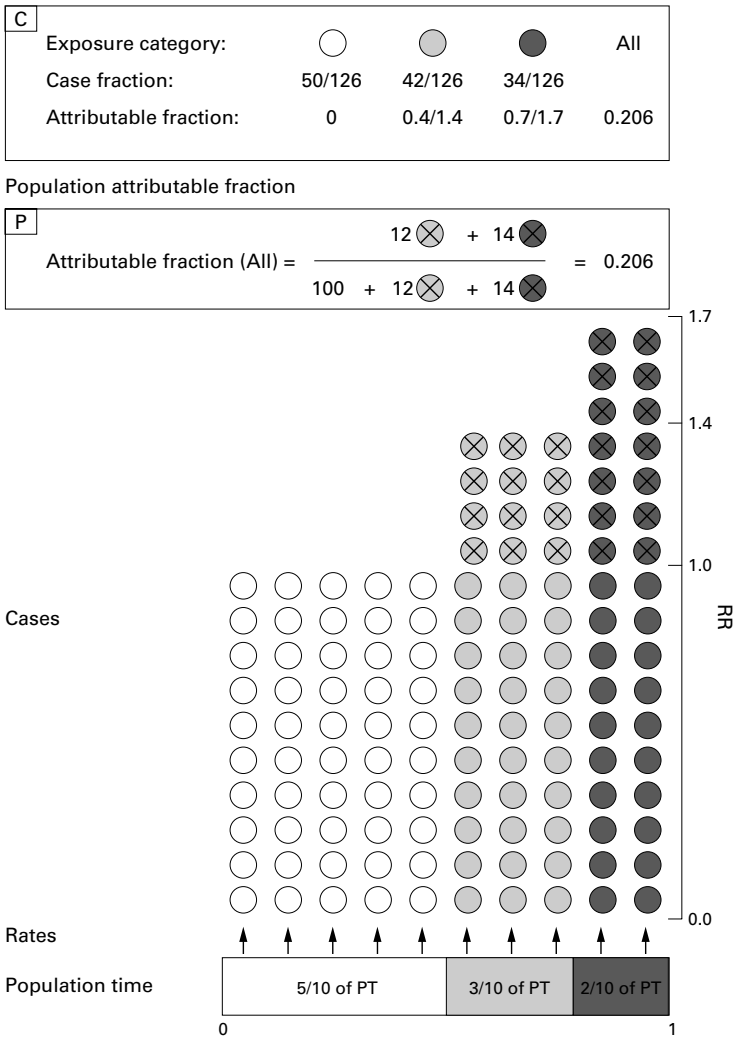

Figure 2 Population (that is, "overall") attributable fraction (AF_p) when exposure is trichotomous, based on distribution of exposure in population (P) and in cases (C). Data are from table 1.

The PF_0 = 5/10th, PF_1 = 3/10th, and PF_2 = 2/10th of the population time (PT) in the low, moderate, and high exposure categories are shown at the bottom of the diagram using increasing levels of shading. Relative rates in these categories are RR_0 = 1, RR_1 = 1.4 and RR_2 = 1.7. The numbers of cases arising from the three categories are shown as correspondingly shaded circles. The number of "excess" cases in each category is depicted as circles marked with an "X" (for "excess").

In the classic AF_p structure (summarised in inset "P", with numbers of cases scaled up by 100), the total number of cases is divided into the lower (square) array denoting the number of expected cases (circles without an "X", irrespective of exposure category) and the two (upper) rectangular arrays denoting the numbers of excess cases.

In the case-based AF_p structure (summarised in inset "C"), the total number of cases is divided first by exposure category—CF_0 = 50/126, CF_1 = 42/126 and CF_2 = 34/126. None of those occurring in the lowest risk category (white circles) are "excess" cases, while 4/14th and 7/17th of the fractions in the higher risk categories are. Thus, 4/14th of 42/126 + 7/17th of 34/126 = 20.6% of all cases are excess cases.

represent the "observed" number, in keeping with the structure in Miettinen's 1985 text (page 254–256),[3] and Levin's[2] original conceptualisation.

The case-based formula begins with the same total of 120 cases and immediately "rules out" all unexposed cases, as, by definition, none of them are "excess" cases. Based on the amount of unexposed PT, they number 60, or 3/5th of the 100 "expected" cases discussed above. This leaves 60 exposed cases (1/2 of the overall total); if RR > 1 and finite, then only a fraction of these 60 exposed cases (or of the 1/2) are excess cases (that is, the maximum possible AF_p is 1/2). What fraction of these 60 represent excess cases? The RR=1.5 implies that of every 1.5 exposed cases, 1 is "expected", while 0.5 of the 1.5, or 1/3rd, are excess. Thus, of the subtotal of exposed 60 cases, 20 are "excess" cases. As the subtotal of 60 exposed cases constitute 1/2 of all cases, and as only 1/3rd of the 60 are excess, then 1/3rd of 1/2, that is, 1/6 of all cases are excess cases. This 1/6th is thus a "fraction of a fraction". The one fraction (1/2) is simply what fraction of all cases are exposed cases—which we have denoted by CF_1. The other, RR—1 as a fraction of RR, that is, 1/3, is the AF specific to the exposed category, namely AF_1.

THESE NUMBERS AND FORMULAS REPRESENTED PICTORIALLY

Figure 1 begins "at the base" with the denominators—the PT distribution—that generated the cases. Some 3/5th (= PF_0) of the PT are in the unexposed category (empty area) and 2/5th (= PF_1) are in the exposed category (shaded area). The cases that arise from these are represented by empty or shaded circles respectively; "excess" cases are marked with an "X" (for "excess"), while "expected" cases are not.

The number of excess cases as a fraction of all cases can be seen from two different views. In the first (Levin), the total number of cases is directly subdivided into two arrays—the (bottom) square array of expected cases, and the (upper) rectangular array of excess cases. With its height (representing the rate in category 0) arbitrarily scaled to 1, and with its entire base of 1, the square array of "expected" cases has an area of 1, representing one expected case. The width of the rectangle of excess cases is PF_1 = 2/5 and its height is RR—1 = 1.5—1 = 0.5, so that its area (the number of "excess cases per one expected case") is (RR—1) × PF_1 = 0.5 × 2/5 = 0.2, yielding AF_p = 0.2/(1 + 0.2) = 0.2/1.2 = 1/6.

In the other view (Miettinen), the total number of cases is first subdivided into two arrays (and thus "case-fractions") on the basis of exposure. Only the exposed (shaded) cases (a fraction CF_1 of all cases) are "eligible" to be excess cases. These exposed cases are then further subdivided into subarrays of excess and expected cases, yielding the attributable fraction AF_1 specific to the exposed cases. An attraction of this "fraction of a fraction" structure of the overall AF_p is that it does not explicitly involve the 3/5 : 2/5 exposure distribution in the source PT—a distribution that is not always easy to estimate—but rather the 60:60 split of the cases themselves.

For those who prefer algebra to pictures, an algebraic derivation of the case-based formula is given in the appendix.

POPULATION ATTRIBUTABLE FRACTION AS A WEIGHTED AVERAGE

Unfortunately, immediately "eliminating" the unexposed cases, and focusing on the exposed ones, distracts from the fact that the case-based AF_p can also be viewed as a weighted average of the two category specific attributable fractions AF_0(= 0) and AF_1. Naturally, as the focus is on

all cases, the weights are given by the relative numbers of cases in exposure categories 0 and 1—that is, by the proportions CF_0 and CF_1. Thus the weighted average of the two category specific fractions AF_0 and AF_1 across both categories of cases is $0 \times CF_0 + AF_1 \times CF_1 = AF_1 \times CF_1$. This representation of AF_p as a weighted average[3] is key to understanding the case-based formula for the polytomous exposure situation considered next.

Polytomous exposure

Figure 2 depicts the data, and illustrates the correct AF_p calculations, for the "three exposure levels" example given in table 1.

CLASSIC STRUCTURE, BASED ON DISTRIBUTION OF EXPOSURE IN POPULATION

Again, the expected cases are shown in the square array of unmarked circles. Now, there are two sets of excess cases, denoted by the lightly shaded and more heavily shaded rectangular arrays of cases marked with an "X". The scaled heights, $\{RR_1—1\}$ and $\{RR_2—1\}$, of these rectangles, multiplied by their widths, PF_1 and PF_2 , yield excess "areas" of $\{RR_1—1\} \times PF_1$ and $\{RR_2—1\} \times PF_2$ respectively. These products represent the number of excess cases for every one "expected". This "expected" versus "excess" partition of the cases leads immediately to the formula

$$AF_p = \frac{PF_1\{RR_1 - 1\} + PF_2\{RR_2 - 1\}}{1 + PF_1\{RR_1 - 1\} + PF_2\{RR_2 - 1\}}. \qquad [2P]$$

For every one expected case, there are $0.12 + 0.14 = 0.26$ "excess" cases, yielding an AF_p of $0.26/1.26 = 20.6\%$ (last column of table 1). Note that applying formula 1P [all or none exposure] twice[13] (second last column of table 1), overestimates the overall fraction of excess cases.

STRUCTURE BASED ON DISTRIBUTION OF EXPOSURE IN CASES

In this view, the AF_p is a sum of 2 "fractions of fractions", that is,

$$AF_p = CF_1 \times \frac{RR_1 - 1}{RR_1} + CF_2 \times \frac{RR_2 - 1}{RR_2} = CF_1 \times AF_1 + CF_2 \times AF_2 \qquad [2C]$$

as originally given in reference 3. As $AF_0=0$, the AF_p can also be seen as a weighted average of the category specific AFs over all 3 levels 0, 1 and 2

$$AF_p = CF_0 \times AF_0 + CF_1 \times AF_1 + CF_2 \times AF_2 \qquad [2C']$$

For the data in table 1 and figure 2, the appropriate calculation is

$$(50/126) \times 0 + (42/126) \times (0.4/1.4) + (34/126) \times (0.7/1.7),$$

yielding the "CF weighed" average, AF_p = 20.6%.

The appendix explains a version that is useful when there are several strata or covariate patterns.

EFFECT OF COLLAPSING CATEGORIES OF HIGHER RISK INTO ONE

Several authors have noted[11 19 20] or shown[7] that the AF_p involving an exposure with levels 0, 1, ..., k is the same as if one first combined categories 1 to k and used the formula for the "all or none" situation. This is easy to see from figure 2, where the RR for the "moderate or high" category, relative to low, is 1.52, and $PF_{moderate/high} = 0.5$. Thus, for every one expected case, there are $0.5 \times 0.52 = 0.26$ excess cases, yielding $AF_p = 0.26/1.26 = 20.6\%$.

The "distributive"[21] property of the AF_p is useful in multiple regression if, instead of aggregating exposure categories, one subdivides them, to the point that each case defines its own exposure category. Details are given in the appendix.

Stratified data

To correctly understand how to aggregate stratum specific AF_ps, first write

$$AF_p \text{ in the aggregate} = \frac{\text{excess number of cases in the aggregate}}{\text{number of cases in the aggregate}}$$

Then, with Σ denoting summation over the strata that form the aggregate, dis-aggregate the numbers of cases so that

$$AF_p \text{ in the aggregate} = \frac{\Sigma \text{ excess number of cases}}{\Sigma \text{ number of cases}}$$

Finally, rewrite this as

$$AF_p \text{in the aggregate} = \frac{\Sigma \text{ number of cases } \times \frac{\text{excess number of cases}}{\text{number of cases}}}{\Sigma \text{ number of cases}} = \frac{\Sigma \text{ number of cases } \times AF}{\Sigma \text{ number of cases}},$$

that is, as a weighted average of stratum specific AF_ps, with the numbers of cases in the each stratum as weights.

Figure 3 illustrates the correct calculation. Whether one arrives at the stratum specific AF_ps "by P or by C", one must average them using the stratum specific numbers (or proportions) of cases as weights. The figure also illustrates the commonly used, but incorrect practice of coupling adjusted (RR-1)s with the marginal distribution of exposure in the overall source.

Discussion

The primary aim of this article is to promote understanding of the AF_p formulas for a polytomous exposure, and for stratified data. To this end, you must begin with the simpler and more familiar, but not fully understood, representations for an all or none exposure. The article also shows how the two seemingly

very different representations can both be derived—without algebraic manipulations—directly from the same diagram. A third aim is to increase awareness of the case-based formulas.

There are a number of possible explanations for the limited awareness and understanding of the case-based formulas. Some textbooks focus on the overall AF before (or without ever) dealing with the specific AFs that are aggregated to create the overall AF. Also, the case-based formula is given in fewer textbooks, usually without a complete derivation. The usually cited source[3] does not explicitly derive it; instead it cites another source,[22] which in turn cites an unpublished source. It was acknowledged (page 331)[3] that the basis for the formula "may not be immediately obvious" and a cryptic explanation was offered. The "fraction of a fraction" formula is derived 11 years later (equation A.2.17, page 256)[10] but in a seemingly different context, and using a purely algebraic manoeuvre that does not reveal the logic behind it (see appendix). The lengthy way in which the formula continues to be algebraically derived in subsequent articles and textbooks[4 5 7 11 23] suggests that the simplicity and "immediate obviousness" of its structure have not been fully or widely understood.

The most important practical benefit of the case-based version is AF_p estimation from stratified, or individually matched, case-control studies[22] where the classic formula is inappropriate.[19 24 25] Variations on this version (see appendix) are also increasingly used to derive—and quantify the sampling variability of—estimators of AF_p from stratified data or multiple logistic regression.[4–8]

The case-based structure also has conceptual benefits. Firstly, it emphasises that AFs refer to cases, and that the observed numbers of cases are the denominators of these AFs. This is in contrast with most epidemiological calculations, where numbers of cases serve as

Confounder stratum	*Exposure level*	*% of population*	*RR*	*Numbers of cases★ (excess cases)*	*Population attributable fraction*
—	—	40	1	40	
—	+	10	1.5†	15(5)	
+	—	10	2†	20	
+	+	40	3†	120(40)	
Total		100		195(45)	45/195 = 23%
Incorrect calculation					
	—	50	1	*60*	
	+	50	1.5‡	*135*§	$\frac{0.5 \times (1.5 - 1)}{1 + 0.5 \times (1.5 - 1)}$ = 20%
Correct calculation					
—			1.5‡	55(5)	$\frac{5}{55}$ = 9.1%
+			1.5‡	140(40)	$\frac{40}{140}$ = 28.6%
Total					$\frac{55 \times \frac{5}{55} + 140 \times \frac{40}{140}}{55 + 140}$ = 23%

★ Numbers of cases are proportional to (% of population) × RR.
† Relative to rate in the (— —) cell; 3/2 = 1.5/1, so no modification of RR.
‡ Adjusted (common) RR.
§ AFp can be calculated correctly using RR = 1.5 in formula [1C]: (135/195) × (0.5/1.5).

Figure 3 Incorrect and correct calcuation of population attributable fraction in presence of confounding factor (hypothetical data).

the numerators of statistics. As exemplified in figure 3, this difference has important implications for how to correctly aggregate stratum specific AF_ps—no matter which version of the formula (classic or case-based) is used to calculate the stratum specific AF_ps. Failure to appreciate this focus on cases may explain why authors, such as those of reference 13, incorrectly couple adjusted (RR-1)s with the marginal distribution of exposure in the source via formulas [1P] and [2P]. This is a common mistake.[14] It is of note, and testimony to the naturalness of the case-based representation, that in the example in figure 3, the weighted average of the stratum specific AF_ps (the AF_ps having been derived from adjusted RRs), using the stratum specific numbers (or proportions) of cases as weights, yields the correct AF_p for the aggregate.

Secondly, although originally derived for empirical estimates from case-control studies, the versatility of the case-based representation can be used in a broader context—for example, to structure the very AF_p parameter itself.[3 8] (section 2, page 866). For these practical and conceptual reasons, the case-based representation needs to be better understood, and not presented in textbooks and articles simply as an algebraic fact.

Greater awareness and understanding of the formulas for polytomous exposure should also decrease computational errors. Even in the absence of confounding, the repeated application of formula [1P], once for each exposure category separately[13] yields an overestimate of AF_p. As is made obvious by figure 2, a single application of formula [2P] yields the correct estimate.

Although its purpose was "multivariate" from the outset, the paper by Eide and Gefeller[26] uses a graphical depiction similar to that presented here. It is helpful to think all of the covariate patterns shown in figure 1 of the Eide and Gefeller article as different levels of a single composite factor, in the spirit of the single polytomous factor in figure 2 of the present article.

The heuristic approach also gives insights into more realistic, and more complex, scenarios than are discussed in introductory textbooks. Indeed, it was questions from a colleague, in a study involving four levels of risk, and the consequences of switching, not to the lowest risk category, but to lower risk categories, that prompted the author to produce diagrams similar to figure 2. Readers are referred elsewhere[10] (appendix 2.3, page 254–6)[12 27 28] for more on this topic.

Technical details on estimating AF_p from regression models can be found in papers by Benichou[7] and Greenland and Dresher (page 1763).[8] Benichou warns that his method for calculating the precision of the estimates is "complex". The methods used by Greenland and Dresher are more tractable, but the matrix notation and associated calculations may still require the help of a statistician. The portion of the article by Oja *et al*[29] that deals with conventional logistic regression modelling, and in particular the hand workable example in the appendix, is a useful point of departure before tackling either of these papers. If you wish to avoid matrix calculations, then bootstrap confidence intervals are an attractive alternative.[30]

Appendix

FORMAL ALGEBRAIC DERIVATION OF CASE-BASED FORMULA 1C

Even for those who prefer algebra to pictures, the majority of the published derivations of formula 1C are much more tedious than they need be. The simplest algebraic derivation is found in Miettinen's text (page 256).[10] It uses the same "fraction of a fraction" logic used to determine that the percentage of eligible subjects who respond to a survey is the percentage of eligible subjects contacted × the percentage of contacted subjects who respond.

$$\begin{aligned} AF_p &= \frac{\text{no. of excess cases}}{\text{total no. of cases}} \\ &= \frac{\text{no. of excess cases}}{\text{no. of eposed cases}} \\ &\qquad \times \frac{\text{no. of exposed cases}}{\text{total no. of cases}} \\ &= AF_1 \quad \times \quad CF_1 \end{aligned}$$

A USEFUL RE-EXPRESSION OF THE CASE-BASED FORMULAS

The version of the "case-based" structure that has become popular as a point of departure for multivariate applications in the past 15 years is, for the three exposure levels example (equation 12, page 327).[3]

$$AF_p = 1 - \left(CF_0 \times \frac{1}{RR_0} + CF_1 \times \frac{1}{RR_1} + CF_2 \times \frac{1}{RR_2} \right), \qquad [2C^*]$$

where $RR_0 = 1$. One can algebraically derive formula 2C⋆ from formula 2C, by rewriting each specific AF in terms of the corresponding RR, and simplifying terms. However, it is more instructive to view the process as taking the complement of the "expected" fraction. In figure 2, the overall "expected" fraction is the sum of three fractions: Of the (50) cases in exposure category 0, the fraction of "expected" cases is 1; of the (42 and 34) cases in categories 1 and 2, the corresponding fractions are $1/RR_1 = 10/14$th, and $1/RR_2 = 10/17$th. Thus, the fraction of the overall cases that are "expected" is the weighted average of the fractions 1, 10/14, and 10/17, with weights given by the case fractions $CF_0 = 50/126$, $CF_1 = 42/126$, and $CF_2 = 34/126$.

Thus, the complement of AF_p

$$= (50/126) \times 1 + (42/126) \times 10/14 + (34/126) \times 10/17$$

$$= 100/126,$$

leading immediately, by subtraction, to formula 2C⋆. Note, however, that unlike formula 2C, this "complement" method requires summation over all levels of the exposure.

EFFECT OF SUBDIVIDING (INDIVIDUALISING) CATEGORIES OF HIGHER RISK

Imagine that, instead of aggregating exposure categories, you continue to subdivide them, to the point that each case defines its own exposure category (this would happen if the exposure takes on values on a continuum, or is a multivariate "x" vector in a multiple regression. Then by the "distributive" property[21] of the AF_p,

$$AF_p = \frac{1}{\text{no. of cases}} \sum \frac{RR_i - 1}{RR_i}$$

$$= 1 - \frac{1}{\text{no. of cases}} \sum \frac{1}{RR_i},$$

where RR_i is the (unconfounded) RR for the covariate pattern of the i-th case, and where the summation is over all of the individual cases. This structure is useful in complex designs[30] and when constructing AF_p from a logistic regression in which each case has a unique covariate pattern.

The author is grateful to Drs Robert Allard, Jean-François Boivin, Michael Kramer and Olli Miettinen for comments and advice on the various versions of this manuscript.

Funding: this work was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada.
Conflicts of interest: none.

1 Rothman KJ, Greenland S. *Modern epidemiology*. 2nd ed. Philadelphia : Lippincott-Raven, 1998.
2 Levin ML. The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum* 1953;**9**:531–41.
3 Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol* 1974;**99**:325–32.
4 Bruzzi P, Green SB, Byar DP, *et al*. Estimating the population attributable risk for multiple factors using case control data. *Am J Epidemiol* 1985;**122**:904–14.
5 Kuritz SJ, Landis JR. Attributable risk ratio estimation from matched-pairs case-control data. *Am J Epidemiol* 1987;**125**: 324–8.
6 Benichou J, Gail MH. Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models. *Biometrics* 1990;**46**:991–1003.
7 Benichou J. Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Stat Med* 1991;**10**:1753–73.
8 Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* 1993;**49**:865–72.
9 Walter S. The estimation and interpretation of attributable fraction in health research. *Biometrics* 1976;**32**:829–49.
10 Miettinen, OS. *Theoretical epidemiology: principles of occurrence research in medicine*. New York: Wiley; 1985:254–6.
11 Kleinbaum DG, Kupper LL, Morgenstern H. *Epidemiologic research: principles and quantitative methods*. Belmont (CA): Lifetime Learning Publications, 1982:section 9.1, 160–4.
12 Schlesselman JJ. *Case control studies: design, conduct, analysis*. New York: Oxford University Press, 1982:section 7.9, 220–6.
13 Moss ME, Lanphear BP, Auinger P. Association of dental caries and blood lead levels. *JAMA* 1999;**281**:2294–8
14 Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health* 1998; **88**:15–19.
15 Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol* 1988;**128**:1185–97.
16 Kramer MS. *Clinical epidemiology and biostatistics: a primer for clinical investigators and decision-makers*. Berlin: Springer-Verlag, 1988.
17 MacMahon B, Trichopoulos D. *Epidemiology: principles and methods*. Boston: Little, Brown, 1996.
18 Hennekens CH. *Epidemiology in medicine*. 1st ed. Boston: Little, Brown, 1987.
19 Walter SD. Calculation of attributable risks from epidemiological data. *Int J Epidemiol* 1978;7:175–82.
20 Breslow NE, Day NE. *Statistical methods in cancer research*. Vol I. Analysis of case-control studies. Lyon: International Agency for Research on Cancer Scientific Publications, no 32,1980.
21 Wacholder S, Benichou J, Heineman EF, *et al*. Attributable risk: advantages of a broad definition of exposure. *Am J Epidemiol* 1994;**140**:303–9.
22 Panayoutou PP, Kaskarelis DB, Miettinen OS, *et al*. Induced abortion and ectopic pregnancy. *Am J Obstet Gynecol* 1972; **114**:507–10.
23 Armitage P, Berry G. *Statistical methods in medical research*. 3rd ed. London: Blackwell, 1994:520.
24 Cole P, MacMahon B. Attributable risk percent in case-control studies. *Br J Prev Soc Med* 1971;**25**:242–4.
25 Whittemore AS. Estimating attributable risk from case-control studies. *Am J Epidemiol* 1983;**117**:76–85.
26 Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. *J Clin Epidemiol* 1995;**48**:645–55.
27 Morgenstern H, Bursic ES. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J Community Health* 1982;7:292–309.
28 Drescher K, Becher H. Estimating the generalized impact fraction from case-control data. *Biometrics* 1997;**53**:1170–6.
29 Oja H, Alho O-P, Laara E. Model-based estimation of the excess fraction (attributable fraction): day care and middle ear infection. *Stat Med* 1996;**15**:1519–34.
30 Rockhill B, Weinberg C, Newman B. Population attributable fraction estimation for established breast cancer risk factors: considering the issues of high prevalence and unmodifiability. *Am J Epidemiol* 1998;**147**:826–33.