SPECIAL ISSUE: PERSPECTIVE

IJCP THE INTERNATIONAL JOURNAL OF CLINICAL PRACTICE

The population attributable fraction and confounding: buyer beware

Linked Comment: www.youtube.com/IJCPeditorial

Linked Comment: Ghaemi & Thommi. *Int J Clin Pract* 2010; 64: 1009–14.

Linked Comment: Lexchin. *Int J Clin Pract* 2010; 64: 1015–8.

Linked Comment: Citrome. *Int J Clin Pract* 2010; 64: 997–8.

In 1997, my colleagues and I estimated the fraction of new cases of diabetes in the United States population attributable to a 10-year weight gain of ≥ 5 kg (1). To estimate this population attributable fraction (PAF), we used a formula that multiplied just two quantities: (i) diabetes 'relative risk' (RR) – the probability of developing diabetes in those who gained ≥ 5 kg divided by the probability in those who gained < 5 kg and (ii) the proportion that gained ≥ 5 kg. We estimated that 27% of new cases of diabetes in the United States were attributable to gaining ≥ 5 kg.

Unfortunately, we got the wrong answer. The correct answer is 21%; we over-estimated the PAF by nearly 30%. What did we do wrong? We followed a well-established, but incorrect, tradition of putting *adjusted* RRs in the crude (unadjusted) PAF formula. The crude formula is only appropriate when the impact of exposure (i.e. ≥ 5 kg gain) on the development of disease (i.e. diabetes) is *not confounded* by other factors, i.e. the crude RR accurately estimates the exposure–disease relationship. We ignored the fact that our RR was adjusted for confounders, i.e. factors correlated with weight gain and associated with developing diabetes *independent of* weight gain.

We adjusted for age, gender, race, education, smoking status, cholesterol, blood pressure, antihypertensive medication, body mass index and alcohol consumption. Our error attributed too many cases of diabetes to weight gain, when some of these cases were attributable to factors we had adjusted for. We could have used an alternative PAF formula that fully adjusts for confounding factors by using *proportion of cases* exposed to ≥ 5 kg gain, instead of proportion of total sample exposed (cases + non-cases). Ironically, we reported all information needed to estimate correctly the PAF in the study itself (1, tables 2 and 4) (see Appendix).

Our error in putting adjusted RRs in the crude formula has been referred to as 'Probably the most common error…' associated with PAF (2, p. 16). One review estimated that at least one in four published studies make this mistake (3). A simulation study putting adjusted RR in the crude PAF formula yielded results that '…were severely biased in most situations' (4, p. 2087). In a recent analysis, this error was shown to over-estimate Unites States mortality attributable to obesity by > 100,000 deaths (5).

Although this error has been discussed in technical literature, it is not easily accessible to clinicians. In this commentary, I try to provide a relatively non-technical understanding of the issue. Probably, if epidemiologists continue to make this error, journal readers will at least be alert to the error and its implications.

Our error in putting adjusted RRs in the crude formula has been referred to as 'Probably the most common error…' associated with PAF

Concepts

The definition of PAF and its fundamental formula are deceptively simple: '… the fraction of all disease cases in a population that are attributable to (caused by) the risk factor under study' (6, p. 131):

$$\mathrm{PAF} = (A - E)/(A + B)$$

where A is the expected number of cases among those exposed to the factor, B is the expected number of cases among those not exposed to the factor, E is the expected number of exposed cases that are not caused by the factor, $A - E$ is the excepted number of cases caused by the factor (6, pp. 131–132).

A, E and B represent unknown population values that can only be estimated from a sample. E and B warrant further discussion.

The language now becomes a bit awkward because we are discussing things that can only occur in the imagination rather than in fact. Among persons exposed to a risk factor, E represents the number of cases of disease that would *hypothetically* occur if they had never been exposed. Diseases often have more than one cause, and removing a risk factor

doi: 10.1111/j.1742-1241.2010.02443.x

might not prevent all cases of the disease; a person with hypertension who smoked might still get a heart attack even if they had hypothetically been normotensive but still smoked. Therefore, $A - E$ represents the *counterfactual* estimate of the number of cases of disease that would occur if we could compare people to themselves, changing only their exposure.

As we never observe the counterfactual quantity $A - E$, we use B – cases that occurred among those that were unexposed – as a surrogate for E. If the sub-populations that give rise to E and B are of the same size and do not differ in prevalence of other risk factors for the disease, $A - B$ provides an unbiased (i.e. unconfounded) estimate of number of cases of disease attributable to the risk factor under study. Although the common computing formulae below account for differences in the size of the E and B sub-populations, the crude PAF formula is only appropriate when the sub-populations do not differ in prevalence of other risk factors for the disease.

When the crude RR is greater than the adjusted RR the incorrect PAF formula will under-estimate the true PAF. When the crude RR is less than the adjusted RR the incorrect PAF formula will over-estimate the true PAF

Common PAF computing formulae

When there are no factors that confound the relationship between the risk factor and the disease, PAF is estimated from a population sample using the computing formula: (7, formula 5, p. 57]

$$\text{Crude PAF} = P_T(RR_C - 1)/(1 + P_T(RR_C - 1))$$

where P_T is the prevalence of the risk factor in the total sample, RR_C is the crude RR.

My colleagues and I used a version of this crude PAF formula appropriate for multiple categories of exposure (8, formula 4, p. 906), but we mistakenly used the RR adjusted for confounders mentioned earlier, rather than the crude RR.

When the relationship between the risk factor and the disease is confounded, the correct PAF computing formula is: (7, formula 7, p. 58)

$$\text{Adjusted PAF} = P_D((RR_A - 1)/RR_A)$$

where P_D is the prevalence of exposure *among cases of disease*, RR_A is the RR adjusted for confounding factors.

My colleagues and I should have used the version of this formula appropriate for multiple exposure categories (8, formula 9, p. 907).

Directions of bias

Bias that occurs when adjusted RRs are used in the crude PAF formula can be mathematically defined, but its form is somewhat complex (6, p. 133). One can, however, predict the direction of bias for a single dichotomous exposure and single dichotomous confounder. When the crude RR is greater than the adjusted RR the incorrect PAF formula will under-estimate the true PAF. When the crude RR is less than the adjusted RR the incorrect PAF formula will over-estimate the true PAF. Table 1 shows two hypothetical scenarios in which the confounder is associated with *increased* incidence of disease (When the confounder is associated with *decreased* incidence of disease, the directions of bias will be reversed).

In the first scenario, the confounder (age) is positively correlated with exposure (obesity) – of 250 old, 200 are obese; of 750 young, only 100 are obese

Table 1 Hypothetical examples of bias in estimates of population attributable fraction (PAF) when there is confounding and adjusted relative risk (RR) used in the crude formula for PAF (inspired by ref. 17, figure 3)*

Confounder	Exposure	N	No. cases of disease	Probability of developing disease	Crude RR (RR_C)	Adjusted RR (RR_A)	Proportion of total N exposed (P_T)	Proportion of cases exposed (P_D)	Incorrect PAF for obesity†	Correct PAF for obesity‡	Bias
Scenario with age as confounder											
Young	Obese	100	9	0.09							
	Not obese	650	39	0.06							−10 ppts
Old	Obese	200	120	0.600	5.10	1.50	0.30	0.69	13%	23%	
	Not obese	50	20	0.400							−43%
Total		1000	188	0.188							
Scenario with smoking as confounder											
Non-smoker	Obese	450	90	0.200							
	Not obese	300	40	0.133							+5 ppts
Smoker	Obese	50	47	0.940	0.83	1.50	0.50	0.45	20%	15%	
	Not obese	200	125	0.625							+33%
Total		1000	302	0.302							

*In both scenarios, the adjusted RR (RR_A) is constant across the confounder strata indicating no effect modification. †Incorrect PAF formula = P_T ($RR_A - 1$)/(1 + P_T ($RR_A - 1$)). ‡Correct PAF formula = P_D (($RR_A - 1$)/RR_A).

– and age is also positively correlated with disease – of 250 old, 140 cases occurred; of 750 young, only 48 cases occurred. If we do not adjust for this confounding by age, we mistakenly conclude that the obese are 5.1 times more likely (crude RR) to develop disease than the non-obese; they are actually only 1.5 times more likely (adjusted RR) to develop disease. Therefore, to correctly estimate the PAF for obesity, the adjusted RR should be used in the adjusted PAF formula to estimate that 23% of cases of disease are attributable to obesity. However, if we put the *adjusted* RR in the *crude* PAF formula, we incorrectly estimate that 13% of cases are attributable to obesity. We did not *fully account* for the fact that older people – in whom most cases of disease occur – are also more likely to be obese. Too few cases of disease were attributed to obesity by mistakenly attributing them to older age.

In the second scenario, the confounder (smoking) is negatively correlated with exposure (obesity) – of 250 smokers, only 50 are obese; of 750 non-smokers, 450 are obese – and smoking is also positively correlated with disease – of 250 smokers, 172 cases occurred; of 750 non-smokers, only 130 cases occurred. If we do not adjust for this confounding by smoking, we mistakenly conclude that the obese are only 0.83 times as likely (crude RR) to develop disease as the non-obese. They are actually 1.5 times *more* likely (adjusted RR) to develop disease. To correctly estimate the PAF for obesity, the adjusted RR should be used in the adjusted PAF formula to estimate that 15% of cases of disease are attributable to obesity. However, if we put the *adjusted* RR in the *crude* PAF formula, we incorrectly estimate that 20% of cases are attributable to obesity. We did not fully account for the fact that smokers – in whom most cases of disease occur – are less likely to be obese than non-smokers. Too many cases of disease were attributed to obesity when they were actually attributable to smoking.

A variation on incorrect adjustment of PAF for confounding is the practice of excluding persons with the confounding factor from the analysis, prior to estimating the PAF. For example, one study excluded smokers before estimating the obesity RR for cancer mortality (9). This RR *for non-smokers* was then used in the PAF formula to estimate the fraction of *all* cancer deaths caused by obesity that occur in the *total* population (*smokers* + non-smokers). This practice inflates the PAF because it misclassifies the many cancer deaths that occur in smokers as deaths because of obesity.

Another way to think about PAF estimation is to remember that adjusting the RR for a confounder stratifies the data by levels of the confounder, each level having its own attributable fraction (see Table 1). The incorrect approach, however, assumes that there is only one attributable fraction in the data. For example, in the scenario where smoking is the confounder, there is a fraction of disease attributable to obesity among smokers and a fraction among non-smokers. The attributable fraction in smokers is 9% and in non-smokers 23%. These two attributable fractions can be added together after weighting them by the *proportion of cases* that are in each of the two smoking strata (0.57 for smokers, 0.43 for non-smokers). The result is the correct PAF of 15%.

Pointers for readers

When reading a study that reports PAF, assess the following:

- Determine if the RR (aka, odds ratio, hazard ratio, rate ratio, risk ratio) was adjusted for any other variables.
- Determine if the crude or adjusted PAF formula was used. Sometimes authors will report that 'a standard formula was used' and give the crude PAF formula. The reader should not be deceived by this practice; there is nothing 'standard' about putting an adjusted RR in the crude PAF formula – it is simply the *wrong* formula.
- Determine if the *proportion of cases* of disease that were exposed to the risk factor is reported. If authors report only the proportion of the total sample exposed, then they probably used the wrong PAF formula.
- Readers should be able to hand-calculate the PAF if they know the RR, the PAF formula used and the proportion of cases exposed to the risk factor.

There is nothing 'standard' about putting an adjusted RR in the crude PAF formula – it is simply the wrong formula

Closing thoughts

When confounding is present, putting the adjusted RR in the crude PAF formula causes bias because this only partially adjusts for confounding (through use of adjusted RR). This practice, however, generally gives a less biased estimate of PAF than if confounding is completely ignored (use of crude RR in crude formula). Perhaps the wrong approach is used because it is thought to be 'better than nothing'.

This commentary has emphasised the use (and misuse) of common computing formulae to estimate PAF. These formulae can be completely avoided, however, through use of statistical models that directly estimate the quantities A, E and B in the

fundamental formula for PAF. For examples see references (10 and 11).

There are other important issues that are beyond the scope of this commentary. PAF should reflect the importance of risk factors that are true *causes* of disease, rather than just correlates of disease; otherwise no insight is gained about the impact of policies to remove risk factors from the population (12). PAF estimates have statistical variability that should be reported with the point estimate (e.g. confidence intervals). I have focused on confounding and PAF, but an equally serious bias arises if the impact of the risk factor on disease is dependent on other factors not accounted for ('effect modifiers'), even if the correct PAF formula is used (13). Reference 14 provides a thorough review of PAF, as well as many other key epidemiological concepts.

Finally, the motivation for this commentary was a letter I published about the incorrect PAF formula used in a recent paper (15). Although the authors acknowledged their error, their revised formula still included the wrong measure of proportion exposed (16).

Acknowledgements

I thank Theodore J. Thompson, M.S. for a careful review of an earlier draft of this study, and Robert Gerzoff, M.S. for independently verifying all calculations. I thank Drs. Katherine F. Flegal, Barry I. Graubard and Mitchell H. Gail for their many insights related to attributable fraction estimation.

Funding

No specific funding sources were used.

Disclosures

None.

D. F. Williamson
Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, USA

Correspondence to:
David F. Williamson, Hubert Department of Global Health, Rollins School of Public Health, Room 740, Emory University, Atlanta, GA 30322, USA
Tel.: 770 488 1054
Fax: 770 488 8550
Email: dfwilli@emory.edu

References

1 Ford ES, Williamson DF, Liu S. Weight change and diabetes incidence: findings from a national cohort of US adults. *Am J Epidemiol* 1997; **146**: 214–22.
2 Rockhill B, Newman B, Weinberg C. Use and misuse of the population attributable fraction. *Am J Pubic Health* 1998; **88**: 15–9.
3 Uter W, Pfahlberg A. The application of methods to quantify attributable risk in medical practice. *Stat Methods Med Res* 2001; **10**: 231–7.
4 Gefeller O. Comparison of adjusted attributable risk estimators. *Stat Med* 1992; **11**: 2083–91.
5 Flegal KF, Graubard BI, Williamson DF, Gail MH. Sources of differences in estimates of obesity-associated deaths from first National Health and Nutrition Examination Survey (NHANES I) hazard ratios. *Am J Clin Nutr* 2010; **91**: 519–27.
6 Greenland S. Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates. *Stat Med* 1984; **3**: 131–41.
7 Morgenstern H. (2008) Attributable fractions. In: Boslaugh S, ed. *Encyclopedia of Epidemiology*, Vol. 1. Thousand Oaks, CA, USA: Sage Publications, 55–63.
8 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**: 904–14.
9 Calle EE, Rodriguez C, Walker-Thurmond K, Thun MJ. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. *N Engl J Med* 2003; **348**: 1625–38.
10 Flegal KM, Graubard BI, Williamson DF, Gail MH. Excess deaths associated with underweight, overweight, and obesity. *JAMA* 2005; **293**: 1861–7.
11 Reeves GK, Pirie K, Beral V, Green J, Spencer E, Bull D. Cancer incidence and mortality in relation to body mass index in the Million Women Study: cohort study. *BMJ* 2007; **335**: 1134–45.
12 Levine BJ. The other causality question: estimating attributable fractions for obesity as a cause of mortality. *Int J Obesity* 2008; **32**: S4–7.
13 Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol* 1987; **125**: 761–8.
14 Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*, 3rd edn. Philadelphia, PA: Lippincot, Williams & Wilkins, 2008: 851 p.
15 Williamson DF. Population attributable risk of incident hypertension in women (letter). *JAMA* 2009;**302**:2550. doi:10.1001/jama.2009.1826.
16 Forman JP. Population attributable risk of incident hypertension in women (reply to Williamson). *JAMA* 2009;**302**:2550–1. doi:10.1001/jama.2009.1826.
17 Hanley JA. A heuristic approach to the formulas for population attributable fraction. *J Epidemiol Community Health* 2001; **55**: 508–14.

Appendix Recalculation of the fraction of new cases of diabetes in the United States population attributable to weight gain ≥ 5 kg originally reported in ref. (1, tables 2 and 4).

Weight gain exposure categories (kg)	N	No. diabetes cases	Adjusted RR (RR_A)	Proportion of total N (P_T)	Proportion of cases (P_D)	Incorrect PAF†	Correct PAF‡	Bias
<5('unexposed')*	5720	271	1	0.67	0.59			
5 to < 8	1240	81	2.11	0.15	0.18			+6 ppts
8 to < 11	701	31	1.19	0.08	0.07	27%	21%	
11 to < 20	706	50	2.66	0.08	0.11			+29%
> 20	148	24	3.84	0.02	0.05			
Total	8515	457		1	1			

*The unexposed category was originally defined as a loss of <5 to a gain of <5 kg. Two additional exposure categories, loss of ≥ 11 and loss of 5 to <11 kg, were also defined. When these two additional exposure categories were combined, their adjusted RR for diabetes was statically indistinguishable from 1.0. Hence, all three categories were combined and assigned the referent category RR = 1. †For exposures with multiple categories, the crude formula for PAF = $1 - (1/(\Sigma P_T\ RR_C))$; (8, formula 4, p. 906), where RR_C is the crude RR and the Σ indicates summation across all categories of the exposure. We did not report the crude RR in our study. We mistakenly put the adjusted RR (RR_A) in the above formula resulting in a PAF estimate of 27% that we originally reported. ‡The analogous formula that should be used with the adjusted RR is PAF = $1 - \Sigma\ (P_D/RR_A)$; (8, formula 9, p. 907). Using this formula, the correct PAF estimate is 21%.