A correction for Levin's formula in the presence of confounding

John Ferguson, Alberto Alvarez, Martin Mulligan, Conor Judge, Martin O'Donnell

HRB Clinical Research Facility
University of Galway
Ireland

Abstract

In 1953, Morton Levin introduced a simple approach to estimate population attributable fractions (PAF) depending only on population risk factor prevalence and relative risk. This formula and its extensions are still in widespread use today, particularly to estimate PAF in populations where individual data is unavailable. Unfortunately, Levin's approach is known to be asymptotically biased for the PAF when the risk factor-disease relationship is confounded even if relative risks that are correctly adjusted for confounding are used in the estimator.

An alternative estimator, first introduced by Miettinen in 1972, is unbiased for the PAF provided the true relative risk is invariant across confounder strata. However, despite its statistical superiority, Miettinen's estimator is seldom used in practice since it requires an estimate of risk factor prevalence within disease cases, a quantity that appears harder to estimate than population risk factor prevalence.

Here we introduce a simple re-expression of Miettinen's estimand that depends on the causal relative risk, the unadjusted relative risk and the population risk factor prevalence. The associated estimator may be useful in estimating PAF in populations when individual data is unavailable provided estimated adjusted and unadjusted relative risks can be transported to the population of interest. The re-expression also generates novel analytic formulae for the relative and absolute bias in Levin's formula, solidifying earlier work by Darrow and Steenland that used simulations to investigate this bias. We extend all results to settings with non binary

valued risk factors and continuous exposures, and discuss the utility of our results in estimating PAF in practice.

Introduction

The Population Attributable Fraction (PAF), sometimes also referred to as the Population Attributable Risk and the Excess Fraction, is a central metric in epidemiology. It is useful both in ascertaining the importance of a risk factor in causing disease, as well as to identify the best risk factors to target in a health intervention. Applications of attributable fractions abound in the literature starting with Richard Doll's work estimating the burden of lung cancer due to smoking (Doll 1951), with recent examples including modifiable risk factors of gout in the US (Liu, Yao et al. 2022), depression risk factors in Brazil (Borelli, Leotti et al. 2022) and modifiable lifestyle risk factors of cancer in Denmark (Tybjerg, Friis et al. 2022), with these studies representing only a very small sampling of recent literature.

In its most basic formulation, PAF represents the fraction of prevalent disease cases that might have been avoided in some population if a risk factor were not present. If we express disease prevalence in the current population as $P(Y = 1)$, Y being a binary indicator for disease, and the prevalence of disease in a population where the risk factor were removed as $P(Y_0 = 1)$, the PAF is simply:

$$PAF = \frac{P(Y = 1) - P(Y_0 = 1)}{P(Y = 1)} \quad (1)$$

A vast literature has evolved on methods to estimate (1) in situations where individual level data exists on risk factors, disease outcomes and confounders. See (Ferguson 2022) for a recent review of methods for PAF estimation in the context of data at the individual level. However, it is sometimes necessary to derive estimates of attributable fractions in scenarios where collection of individual level data is impossible. As an example, the Global Burden of Disease project (Tran, Lang et al. 2022) estimates population attributable fractions at a country level for differing risk factor / disease combinations. For certain countries, no individual level data linking risk factors of interest to disease may be available. The most common approach in this situation is to substitute an estimated risk factor prevalence $\hat{\pi}$ and estimated relative risk of disease

(adjusted for confounders), $\hat{RR}_C$ into equation (2) below. Equation (2) was introduced by (Levin 1953), also in the context of estimating the PAF for smoking as a cause of lung cancer:

$$PAF_L = \frac{\pi(RR_C - 1)}{1 + \pi(RR_C - 1)}. \quad (2)$$

(2) is prominent throughout the literature on PAF estimation. However, as recognised by multiple authors (Khosravi and Mansournia 2019),(Rockhill, Newman et al. 1998, Darrow and Steenland 2011),(Hanley 2001) (2) will usually differ from (1) as a quantity. In particular (2) will equal the true PAF only under the unrealistic condition that the risk factor/outcome relationship is subject to no confounding (or alternatively competing sources of confounding 'cancel out', (Hernan and Robins 2023)). This condition implies that the relative risk unadjusted for covariates, RR_U, is identical to the 'causal' relative risk $RR_U = RR_C$. Here the causal relative risk can be defined as the ratio of disease prevalence comparing scenarios where the entire population was exposed and the entire population was unexposed to the risk factor in question.

Mittenen's definition of PAF

While (2) does not equal PAF in the absence of this 'no confounding' assumption, an alternative expression introduced by (Miettinen 1974) (3), which involves the prevalence of the exposure within cases, π_c, as opposed to the overall population prevalence and the causal relative risk, RR_C, does equal (1) irrespective of the degree of confounding provided the causal relative risk, RR_C, is constant in differing strata of the confounders. Miettinen's expression is given by

$$PAF = \pi_c \frac{RR_C - 1}{RR_C} \quad (3)$$

where π_c is now the proportion of disease cases that have the risk factor.

The usual advice is to use (3) when possible (Rockhill, Newman et al. 1998) and not (2) as a basis for estimating PAF. However, estimating π_c directly might be problematic for rare diseases, particularly in jurisdictions with undeveloped data-capture systems. For example the Global Burden of Disease project utilise (2) rather than (3) to estimate PAF most likely as a result of the difficulty in estimating π_c. This difficulty to estimate π_C is likely the reason for the popularity of Levin's formula.

A re-expression of Miettenen's PAF

A fact that has been under appreciated in the literature is that equation (3) can be re-expressed to be a function of three quantities: the prevalence of the risk factor in the population, π, the causal relative risk, RR_C, and the unadjusted relative risk, RR_U in the population:

(4) $$PAF = [\frac{\pi RR_U}{1 + \pi(RR_U - 1)}]\frac{RR_C - 1}{RR_C}$$

The proof that (4) equals the PAF is a relatively simple application of Bayes' rule (as given in the Supplementary material). Despite this, the only reference for (4) in the context of PAF that we could find in the literature was (Khosravi, Nazemipour et al. 2021), based on a similar observation in (Strain, Brage et al. 2020) for the prevented fraction. However, the expression in (Khosravi, Nazemipour et al. 2021) is incomplete in that it doesn't distinguish between RR_U and RR_C and as a result may mislead researchers. While a somewhat simple re-expression of (3), (4) may be practically useful in estimating PAF with summary data as unadjusted relative risks are often reported together with adjusted relative risks. It is useful to notice that under no confounding ($RR_U = RR_C$), the expression (4) simplifies to Levin's formula (a fact which can be used to show that Levin's formula does recover the true PAF under confounding). However, formulae (2) and (4) will otherwise differ.

Analysis of bias in Levin's estimate

For illustration regarding both the biases at play from Levin's approach and the potential of (4) to provide better estimates of PAF than (2) when individual-level data is unavailable, we will consider results from (Schuch, Vancampfort et al. 2018) who conducted meta analyses to estimate the causal effect of physical inactivity on depression. Their analyses estimated both unadjusted $\hat{RR}_U$ [1.69 95%CI (1.18-1.96)] and adjusted $\hat{RR}_C$ 1.20 95%CI (1.10, 1.32) risk ratios for physical inactivity. The percentage of physical inactivity (according to a certain definition) may vary greatly over differing populations, depending on factors such as culture and age structure. Let's say that for a particular population an estimate of the percentage of physical inactivity was $\hat{\pi} = 0.5$. Plugging in the estimated adjusted relative risk into Levin's formula results in an estimated PAF of 9.1%, whereas using (4), the (correctly) estimated PAF is actually 10.5%. Suppose we wish instead to estimate the PAF in a more active population with an estimate of physical inactivity: $\hat{\pi} = 0.2$. In this case, the bias is smaller but the relative bias is larger: Levin's formula gives an estimate of 3.8% whereas (4) generates an estimate of 5.0%. These are examples of the general behaviour we would expect of the bias as we will show below.

(Darrow and Steenland 2011) investigated the bias of Levin's formula using simulation, and showed that as the degree of confounding, specified by $\max\{C, \frac{1}{C}\)$ where $C = \frac{RR_C}{RR_U}$ gets larger (with no confounding being represented by $C = 1$), the magnitude of *relative* bias in Levin's formula will increase. When $C < 1$, one would expect that $PAF_L < PAF$ (as is the case in the effect of physical inactivity on depression) Their simulations also indicate that this relative bias will be larger for smaller risk factor prevalences (as again observed in the previous example). However, having an explicit formula for relative bias allows more rigorous analysis of limiting behaviour (and subsequent maximum extent of bias) than was possible in (Darrow and Steenland 2011). Given the formula (4), an expression for the relative bias can be trivially derived by simply taking the ratio of (3) and (4). After some simplification this results in:

(5) $$\frac{PAF_L}{PAF} = \frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)}\times C$$

Observing that the partial derivative $\frac{\delta}{\delta C}\frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)}\times C$ is strictly positive (fixing RR_u and π) it follows that (5) is an increasing function of C. Given that (5) equals 1 at $C=1$ (that is Levin's formula is unbiased when there is no confounding) this indicates that the degree of relative bias gets larger as the degree of confounding increases in either direction $C>1 : C\uparrow\infty$ or $C<1 : C\downarrow 0$. It is useful to analyse the relative (and absolute bias) separately in these two scenarios $C>1$ and $C<1$. See the supplementary material for more detail on these analyses).

Bias in Levin's formula when $C>1$

For values of $C>1$, Levin's formula is biased above with the relative bias increasing toward a limit of $1+\frac{1-\pi}{\pi}RR_U$ as $C\to\infty$ (note that this analysis assumes π and RR_U are fixed). As one would expect under no confounding $C=1, \frac{PAF_L}{PAF}=1$ (that is the Levin estimand and true PAF are equal). Provided $RR_U>1$, the absolute bias: PAF_L-PAF also increases as C increases (from $PAF_L-PAF=0$ when $C=1$ to $1-\frac{\pi RR_U}{1+\pi(RR_U-1)}$ as $C\to\infty$). Note that Levin's formula, (2), converges to 1 as $C\to\infty$ irrespective of the values of π and RR_U, whereas the true PAF converges to $\frac{\pi RR_U}{1+\pi(RR_U-1)}$ as C gets very large. In practice, imagining what the bias might be for extremely large values of C is interesting theoretically in it demonstrates the maximum possible bias in Levin's approach; but whether one would see such extreme confounding

in a real example is debatable. For instance, a scenario where extreme values of C is possible is if there is a second risk factor X^* that acts as a confounder for $X \rightarrow Y$, with a strong causal influence on X (so much so that $X^* \sim X$ in the population), and such that the direct effect of X^* (not mediated through X) is to strongly reduce the likelihood of disease (by a factor of $\frac{1}{C}$) but the effect of X on Y is to strongly increase the likelihood of disease by factor of C. Such scenarios are likely implausible in practical situations.

When $C > 1$, $\frac{\delta}{\delta\pi}\frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)} \times C < 0$ (fixing RR_U and C), indicating that as risk factor prevalence, π, decreases risk factor relative bias, (5), increases. Given that relative bias is larger than 1 when $C > 1$, this indicates that relative bias worsens as risk factor prevalence gets smaller, although the situation for the degree of absolute bias is more complicated (absolute bias will converge to 0 as $\pi \downarrow 0$ and as $\pi \uparrow 1$).

Bias in Levin's formula when $C \leq 1$

For $C < 1$, Levin's formula is biased below and the relative bias becomes progressively worse, eventually converging towards 0 as $C \downarrow 0$ (a realm in which risk factors are protective and attributable fractions are negative). However, usually we imagine a risk factor coding such that $RR_C \geq 1$. Assuming that $RR_U > 1$, the absolute bias $PAF_L - PAF$ is 0 under no causal effect, $RR_C = 1$, (or equivalently for $C = \frac{1}{RR_U}$) in addition to being 0 when $C = 1$ (that is under no confounding). By continuity of the expressions (2), (3) and (4) as functions of C we can argue that when the true causal effect is very small, $RR_C \approx 1$, the absolute bias in Levin's formula will be negligible, while the relative bias will be approximately $\frac{1+\pi(RR_U-1)}{(1-\pi)RR_U} < 1$, although arguably the relative bias is not so unimportant in this case. The lesson here is that Levin's formula

can be appropriate to use in the scenario that the true causal effect is very small or when there is no confounding).

When $C < 1$, $\frac{\delta}{\delta\pi}\frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)} \times C > 0$ (fixing RR_U and C), indicating that as risk prevalence decreases, relative bias decreases. Given that the relative bias function is smaller than 1 when $C < 1$, this indicates again that relative bias gets worse as risk factor prevalence gets smaller, although as in the situation where $C > 1$ absolute bias converges to 0 as π converges to 0 or as π converges to 1.

PAF for exposures with more general distributions.

When the distribution of exposure is multi-category or continuous, PAF can be defined as the fraction of prevalent (or incident) disease cases that would have been avoided in a population where the value of the exposure was fixed at a particular value (or within a range of values) that is known to minimise the likelihood of disease (this value is known as the minimum risk exposure value or MREV). While with individual level data it is possible to estimate the MREV (Ferguson, Maturo et al. 2020), when deriving estimates of PAF from summary estimates one usually assumes a pre-determined MREV (Tran, Lang et al. 2022)) For example, in the simple case of an exposure such as air pollution which can be eliminated, the MREV would be presumed to be 0 (that is elimination), with non-zero values being appropriate for exposures like blood pressure or sodium consumption. The causal relative risk is now a function, $RR_C(x)$ comparing the increased prevalence of disease if the population were all exposed to exposure level x relative to disease prevalence if the same were all exposed to the MREV. Under the assumption that for each possible exposure value, x, $RR_C(x)$ is constant within confounder strata, Miettinen's formula (4) can be extended as follows:

(6) $$PAF = \frac{P(Y=1)-P(Y_{MREV}=0)}{P(Y=1)} = E_{X|Y=1}[\frac{RR_C(X)-1}{RR_C(X)}]$$

where $E_{X|Y=1}[f(X)]$ denotes the average of a function, $f(X)$, of the exposure, X within the population of individuals with disease (see the Supplmentary material for a proof). Note that this formula is valid for both continuous exposures (such as air pollution) and multi-category exposures (such as a three level coding of smoking in terms of current smokers/former smokers and never-smokers), and simplifies to (3) in the setting of a binary exposure.

As shown in the supplementary material, (6) can be re-expressed using Bayes' Rule as the follows:

$$(7) \qquad PAF = \frac{E_X(RR_U(X)\frac{RR_C(X)-1}{RR_C(X)})}{E_X(RR_U(X))} = \frac{\int RR_U(x)\frac{RR_C(x)-1}{RR_C(x)}dF(x)}{\int RR_U(x)dF(x)}$$

, with $R_U(x)$ representing the unadjusted relative risk function that compares prevalence of disease in the strata of the population exposed to the value x of the risk factor with the the prevalence in the strata exposed to the MREV, and E_X is the population expectation operator. The right hand side of (7) indicates that one can calculate PAF by numerical integration of the quantities $RR_U(x)\frac{RR_C(x)-1}{RR_C(x)}$ and $RR_U(x)$ with respect to the population distribution, $F(x)$, of the exposure. In the special case that the exposure is discrete, but with more than two levels: $X \in \{0,1,...,l\}$, with the population prevalences of each level: $\pi(0), \ldots, \pi(l)$, and MREV=0, equation (7) can be re-expressed in a more palatable form:

$$(8) \qquad PAF = \frac{\sum_{x=1}^{x=l} p(x)RR_U(x)\frac{RR_C(x)-1}{RR_C(x)}}{\pi(0)+\sum_{x=1}^{x=l} p(x)RR_U(x)}$$

Bias from Levin's approach for more general exposure distributions

The generalisation of Levin's formula is to multi-category and continuous exposure is given by:

$$(9) \qquad PAF_L = \frac{E_X(RR_C(X)) - 1}{E_X(RR_C(X))} = \frac{\int RR_C(x)dF(x) - 1}{\int RR_C(x)dF(x)}$$

Comparing equations (7) and (9), it follows immediately that under no confounding, that is $RR_U(x) = RR_C(x)$ for every exposure value x, Levin's formula is again unbiased. However, analysis of bias in Levin's formula is more complicated when the distribution of the exposure is non-binary, although under certain special settings, one can analyse bias in the same way as in the binary exposure case. For instance, suppose that the $MREV = 0$ and the prevalence of the minimum risk exposure level in the general population is $1 - \pi$ where $1 > 1 - \pi > 0$. Assuming that the confounding ratio $C(X) = \frac{R_C(X)}{R_U(X)} = C$ at all exposure levels not equal to the MREV, the relative bias in the Levin estimate is:

$$(10) \quad \frac{PAF_L}{PAF} = \frac{1 + \pi(E_{X|X>0}RR_U(X) - 1)}{1 + \pi(C \times E_{X|X>0}RR_U(X) - 1)} \times C,$$

, an expression very similar to (5) and which indicates that larger relative biases are expected under larger levels of confounding (that is $C << 1$ or $C >> 1$). For continuous exposures, the assumption that a non-zero proportion of the population is exposed to the minimum exposure level may well be implausible. However a similar equation to (10) will hold appropriately if there is a range of exposure values, R, having non zero probability in the population which approximately minimise counterfactual risk: $\max_{x \in R} RR_C(x) \approx 1$, where $0 < \pi = P(X \notin R) < 1$ and $C(x) = \frac{R_C(x)}{R_U(x)} = C$ when $x \notin R$.

The modified equation then would be:

(11) $\frac{PAF_L}{PAF} \sim \frac{1+\pi(E_{X|X\notin R}RR_U(X)-1)}{1+\pi(C\times E_{X|X\notin R}RR_U(X)-1)} \times C$. The proofs of equations (10) and (11) are given as supplementary material

Discussion

We use this final section mostly to discuss caveats and limitations to our suggested approaches. First, we have suggested that where possible equation (4) should be used instead of Levin's formula, (2), to estimate PAF when individual level data is unavailable. Equation (4) requires estimates $\hat{RR}_U$ and $\hat{RR}_C$ of the unadjusted and causal relative risks. To correctly estimate PAF these estimates need to be transportable to the target population. While transportability of the causal relative risk is an issue both for equation (4) and Levin's approach, (2), transportability of the unadjusted relative risk is an additional assumption and there is no good reason to expect it to hold.

Fortunately, often plugging in an incorrect estimated unadjusted relative risk into (4) will represent an improvement over Levin's formula (3), provided that confounding results in the same direction of bias in the source population (where the unadjusted and adjusted relative risks are estimated) and the target population (where these estimates are transported to derive an estimate of PAF). For instance, if the true population relative risks are RR_U and RR_C and we plug in any incorrect unadjusted estimate, $\hat{RR}_U$ such that: $RR_U \geq \hat{RR}_U < RR_C$ if $C > 1$(or such that $RR_C < \hat{RR}_U \leq RR_U$ if $C < 1$) into (4), so that:

$\hat{PAF} = [\frac{\pi\hat{RR}_U}{1+\pi(\hat{RR}_U-1)}]\frac{RR_C-1}{RR_C}$, the absolute error in estimating PAF is guaranteed to improve compared to applying Levin's formula directly, that is: $|\hat{PAF}-PAF| < |PAF_L-PAF|$. In more general, provided the patterns of confounding in the source and target population are reasonably similar, one might expect $\hat{RR}_U$ and RR_U to be reasonably similar and equation (4) to be more accurate than equation (2). After applying equation (4), one also perform sensitivity analyses to determine the range of true unadjusted relative risks (in the population of interest) where

the derived estimate would represent an improvement of the Levin estimate (in absolute value error). As an example, in the example with physical inactivity and depression described earlier, we estimated $\hat{RR}_U = 1.69$, which leads to an estimated PAF of 10.5% (assuming the correct prevalence $\pi = 0.5$ and the correct causal relative risk $RR_C = 1.2$ are known). Substituting $\hat{RR}_U$ into equation (4) in this context leads to a smaller absolute error in PAF provided the true unadjusted relative risk in the target population $RR_U > 1.42$.

As noted by other authors, the bias from Levin's formula is generally quite small (Darrow and Steenland 2011) and likely insubstantial compared to other biases involved in estimating a PAF. For instance, in the example regarding the relationship between physical inactivity and depression, there are additional questions about consistency of ascertainment and measurement of both the exposure (inactivity) and outcome (depression). The prevalence of defined inactivity (and the associated relative risk) will change depending on what the investigator considers to constitute inactivity, and the definition may differ across different studies. Similar issues arise in consistency of ascertainment and measurement of depression. These issues reduce the real world meaning and actionability of any estimated PAF. On a related point, physical activity should really be measured on a continuum; a binary definition of inactivity will likely underestimate associated disease burden. For instance, a better approach to calculating PAF might determine an optimal level of physical activity (that perhaps differs dependent on age and other characteristics of a person) and estimate disease prevalence in a hypothetical population where everybody had at least this optimal level of activity; however, implementation of such an estimator may be practically challenging. Finally, the assumption that the causal relative risk is constant across differing confounding strata will usually be dubious (individuals exposed or not exposed to differing patterns of confounding variables have differing baseline probabilities of disease making the same relative effect of a risk factor unlikely). This assumption is necessary to prove the equality of Miettinen's formula (3) and true PAF. If (as expected) the causal relative risk varies over differing strata of confounders, the correct extensions of Miettinen's formulae for binary and general exposure distributions are as follows:

(12) $$E_{C|Y=1}[P(X=1\,|\,C,Y=1)\frac{RR_C(C)-1}{RR_C(C)}]$$

(13) $$E_{C|Y=1}[E_{X|C,Y=1}[\frac{RR_C(X,C)-1}{RR_C(X,C)}]]$$

where $RR_C(c)$ is regarded as the causal relative risk within confounder stratum $C=c$ in (12) and $RR_C(x,c)$ as the causal relative risk comparing exposure levels x and the MREV in confounder stratum $C=c$ in (13). These formulae (proven in the Supplmentary Material) are equivalent to (and represent extensions of in the case of (13)) the formulae suggested in (Bruzzi, Green et al. 1985), and require individual level data (incorporating effect modification between risk factors and confounders) to implement. When relative risks vary over confounder strata, the marginal causal relative risk, RR_C (defined as the ratio of disease probabilities if everyone were exposed and everyone were exposed to the risk factor) will be a weighted average of the causal relative risks, $RR_C(c)$ in confounder strata (this follows since the relative risk is a collapsible risk measure). As a result, one would not expect large differences between (4) and (12), at least under moderate levels of effect modification.

While equivalent results to equation (4) exist for multi-category and continuous exposures (equations (7)-(8)) these may be not as useful in practice as equation (4) as they require specification of unadjusted relative risks comparing many levels of exposure to baseline. As a result, Levin's formula (despite its bias) may forever be the method of choice for estimating PAFs and impact fractions for continuous exposures with summary data. However, the biases in Levin's formulae are often dismissed in this setting. If the formula is to be used it is important to recognise its bias and have some awareness of the likely extent of the bias. In this regard, if there is a range of values of the exposure which approximately minimise risk and it's plausible that the confounding parameter: $C(X)=\frac{RR_C(X)}{RR_U(X)}$ is approximately constant outside of this range, equation (11) may be useful in determining the likely error from using Levin's approach.

Finally, we'd like to make some comments regarding Levin's formula itself that may be causing some confusion in the literature. Some authors have stated that Levin's formula is only valid when relative risks are unadjusted (Khosravi and Mansournia 2019), (Rockhill, Newman et al. 1998), (Flegal, Panagiotou et al. 2015). This may give an erroneous impression that Levin's formula should be applied to unadjusted relative risks, which has perhaps encouraged some researchers to apply Levin's formula with unadjusted relative risks or odds ratios (Abreo, Gebretsadik et al. 2018), (Lee, Whitsel et al. 2022). However, substituting unadjusted relative risks into Levin's formula is not recommended and is likely to lead to a more egregious error than using relative risks appropriately adjusted for confounding in the same formula (particularly when the true causal relative risk is close to 1). For example in the example regarding physical inactivity and depression we discussed earlier this leads to an estimated PAF of 25.7%, likely a huge overestimate for the true PAF. A second more technical point is that Levin's formula does not require an assumption of no effect modification as some authors have stated (Khosravi and Mansournia 2019). The sole assumption for Levin's formula to exactly represent PAF is that there *is* no confounding (see the Supplementary material). In the unlikely circumstance that there is no confounding, the correct relative risk to use in Levins' formula would be the marginal unadjusted relative risk over the population: $RR_U = \frac{P(Y=1\,|\,X=1)}{P(Y=1\,|\,X=0)}$, but this marginal relative risk will be a weighted average of covariate specific relative risks that may well differ for differing covariate strata. However, we reiterate our advice from earlier in the paragraph: if Levin's formula is to be used in situations where adjusted and unadjusted relative risks differ substantially, the version using the adjusted relative risk is likely to be less biased and should be preferred to the version with the unadjusted relative.

Bibliography

Abreo, A., et al. (2018). "The impact of modifiable risk factor reduction on childhood asthma development." Clinical and translational medicine **7**(1): 1-12.

Borelli, W. V., et al. (2022). "Preventable risk factors of dementia: Population attributable fractions in a Brazilian population-based study." The Lancet Regional Health-Americas **11**: 100256.

Bruzzi, P., et al. (1985). "Estimating the population attributable risk for multiple risk factors using case-control data." American journal of epidemiology **122**(5): 904-914.

Darrow, L. A. and N. K. Steenland (2011). "Confounding and bias in the attributable fraction." Epidemiology: 53-58.

Doll, R. (1951). "On the aetiology of cancer of the lung." Acta Unio Int Contra Cancrum. **7(1 Spec. No.)**: 39-50.

Ferguson, J., et al. (2020). "Population attributable fractions for continuously distributed exposures." Epidemiologic Methods **9**(1).

Ferguson, J. O. C. M. (2022). "graphPAF: An R package to estimate and display population attributable fractions." from https://cran.r-project.org/web/packages/graphPAF/vignettes/graph_PAF_vignette.pdf.

Flegal, K. M., et al. (2015). "Estimating population attributable fractions to quantify the health burden of obesity." Annals of Epidemiology **25**(3): 201-207.

Hanley, J. (2001). "A heuristic approach to the formulas for population attributable fraction." Journal of Epidemiology & Community Health **55**(7): 508-514.

Hernan, M. A. and J. M. Robins (2023). Causal Inference, CRC Press.

Khosravi, A. and M. A. Mansournia (2019). "Issues with incorrect computing of population attributable fraction (PAF) in a global perspective on coal-fired power plants and burden of lung cancer." Environmental Health **18**(1): 1-2.

Khosravi, A. and M. A. Mansournia (2019). "Recommendation on unbiased estimation of population attributable fraction calculated in “prevalence and risk factors of active pulmonary

tuberculosis among elderly people in China: a population based cross-sectional study”." Infectious diseases of poverty **8**(1): 1-3.

Khosravi, A., et al. (2021). "Population attributable fraction in textbooks: Time to revise." Global Epidemiology **3**: 100062.

Lee, M., et al. (2022). "Variation in population attributable fraction of dementia associated with potentially modifiable risk factors by race and ethnicity in the US." JAMA network open **5**(7): e2219672-e2219672.

Levin, M. L. (1953). "The occurrence of lung cancer in man." Acta Unio int contra cancrum **9**: 531-941.

Liu, K., et al. (2022). Modifiable risk factors and incidence of gout: estimation of population attributable fraction in the US. Seminars in Arthritis and Rheumatism, Elsevier.

Miettinen, O. S. (1974). "Proportion of disease caused or prevented by a given exposure, trait or intervention." American journal of epidemiology **99**(5): 325-332.

Rockhill, B., et al. (1998). "Use and misuse of population attributable fractions." American journal of public health **88**(1): 15-19.

Schuch, F. B., et al. (2018). "Physical activity and incident depression: a meta-analysis of prospective cohort studies." American Journal of Psychiatry **175**(7): 631-648.

Strain, T., et al. (2020). "Use of the prevented fraction for the population to determine deaths averted by existing prevalence of physical activity: a descriptive study." The Lancet Global Health **8**(7): e920-e930.

Tran, K. B., et al. (2022). "The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019." The Lancet **400**(10352): 563-591.

Tybjerg, A. J., et al. (2022). "Updated fraction of cancer attributable to lifestyle and environmental factors in Denmark in 2018." Scientific Reports **12**(1): 1-11.

Supplementary Material

1. Proof that Miettinen's formula is unbiased for PAF if $RR_C = \frac{P(Y=1\,|\,X=1,C=c)}{P(Y=1\,|\,X=0,C=c)}$ is constant over confounder strata $C=c$

$$
\begin{aligned}
PAF &= \frac{P(Y=1)-P(Y_0=1)}{P(Y=1)} \\
&= \frac{P(Y=1)-E_C(P(Y=1\,|\,X=0,C))}{P(Y=1)} \\
&= \frac{E_{X,C}(P(Y=1\,|\,X,C))-E_C(P(Y=1\,|\,X=0,C))}{P(Y=1)} \\
&= \frac{E_C(P(X=1\,|\,C)P(Y=1\,|\,X=1,C)+P(X=0\,|\,C)P(Y=1\,|\,X=0,C))-E_C(P(Y=1\,|\,X=0,C))}{P(Y=1)} \\
&= \frac{E_C(P(X=1\,|\,C)P(Y=1\,|\,X=1,C)+P(X=0\,|\,C)P(Y=1\,|\,X=0,C)-P(Y=1\,|\,X=0,C))}{P(Y=1)} \\
&= \frac{E_C(P(X=1\,|\,C)P(Y=1\,|\,X=1,C)-P(X=1\,|\,C)P(Y=1\,|\,X=0,C))*}{P(Y=1)} \\
&= (RR_C-1)\frac{E_C(P(X=1\,|\,C)P(Y=1\,|\,X=0,C))}{P(Y=1)} \text{ since } RR_C = \frac{P(Y=1\,|\,X=1,C)}{P(Y=1\,|\,X=0,C)} \\
&= \frac{(RR_C-1)}{RR_C}\frac{E_C(P(X=1\,|\,C)P(Y=1\,|\,X=1,C))}{P(Y=1)} \text{ since } RR_C^{-1} = \frac{P(Y=1\,|\,X=0,C)}{P(Y=1\,|\,X=1,C)} \\
&= \frac{P(X=1,Y=1)}{P(Y=1)} \times \frac{RR_C-1}{RR_C} \text{ (since } P(X=1,Y=1)=E_C(P(X=1,Y=1\,|\,C)) \\
&= P(X=1\,|\,Y=1)\times\frac{RR_C-1}{RR_C} = \pi_c\frac{RR_C-1}{RR_C}
\end{aligned}
$$

Proof when there are no confounders (but possible effect modification)

When there are no confounders, the proof is much simpler:

$$PAF = \frac{P(Y=1) - P(Y_0=1)}{P(Y=1)} = \frac{P(Y=1) - P(Y=1\,|\,X=0)}{P(Y=1)}$$

$$= \frac{P(Y=1\,|\,X=0)P(X=0) + P(Y=1\,|\,X=1)P(X=1) - P(Y=1\,|\,X=0)}{P(Y=1)}$$

$$= \frac{P(X=1)(P(Y=1\,|\,X=1) - P(Y=1\,|\,X=0))}{P(Y=1)}$$

$$= \frac{P(X=1)P(Y=1\,|\,X=1)(1 - RR_C^{-1})}{P(Y=1)} = \pi_c(1 - RR_C^{-1})$$

Where $RR_C = \frac{P(Y=1\,|\,X=1)}{P(Y=1\,|\,X=0)}$ is the marginal unadjusted relative risk (which has a causal interpretation due to the lack of confounders). However, note that the causal relative risk might still differ when stratified by particular covariate patterns - that is this proof still holds true under general patterns of effect modification. An application of Bayes' Rule shows this is also Levin's formula which indicates Levin's formula does not require an assumption of no effect modification.

2. Proof of generalisation of Miettinen's formula when the causal relative risk $RR_C(c)$ varies over covariate strata:

If $RR_C(C) = \dfrac{P(Y=1 \mid X=1, C)}{P(Y=1 \mid X=0, C)}$ is not constant the result follows from the line with the asterix in (1) using similar algebra:

After the asterix, the proof proceeds as:

$$\frac{E_C(P(Y=1, X=1 \mid C)\frac{RR_C(C)-1}{RR_C(C)})}{P(Y=1)}$$

$$= \frac{E_C(P(Y=1 \mid C)P(X=1 \mid Y=1, C)\frac{RR_C(C)-1}{RR_C(C)})}{P(Y=1)}$$

$$= E_{C \mid Y=1}(P(X=1 \mid Y=1, C)\frac{RR_C(C)-1}{RR_C(C)})$$

Where the last line follows since if $f(c)$ is the density of the confounders C $f(c \mid Y=1) = f(c)P(Y=1 \mid c)/P(Y=1)$ is the density of the confounders given $Y=1$

3. Proof that Miettinen's formula can be re-expressed as $\frac{\pi RR_u}{1+\pi(RR_u-1)}\frac{RR_C-1}{RR_C}$

The result will be true if $\pi_c = \frac{\pi RR_u}{1+\pi(RR_u-1)}$; this can be proven using Bayes' Rule.

Proof:

$$\begin{aligned}
\pi_C &= P(X=1\,|\,Y=1) \\
&= \frac{P(X=1,Y=1)}{P(Y=1)} \\
&= \pi\frac{P(Y=1\,|\,X=1)}{\pi P(Y=1\,|\,X=1)+(1-\pi)P(Y=1\,|\,X=0)} \\
&= \pi\frac{1}{\pi+(1-\pi)RR_U^{-1}} \\
&= \frac{\pi RR_U}{1+\pi(RR_U-1)}
\end{aligned}$$

This implies that:

$$PAF = [\frac{\pi RR_U}{1+\pi(RR_U-1)}]\frac{RR_C-1}{RR_C}$$

4. Derivation of Formula (5) in the main manuscript for relative bias

$$
\begin{aligned}
B &= \frac{PAF_L}{PAF} \\
&= \frac{(RR_C - 1)(1 + \pi(RR_U - 1))}{RR_U(1 + \pi(RR_C - 1))} \times \frac{RR_C}{RR_C - 1} \\
&= \frac{1 + \pi(RR_U - 1)}{1 + \pi(RR_C - 1)} \times \frac{RR_C}{RR_U} \\
&= \frac{1 + \pi(RR_U - 1)}{1 + \pi(C \times RR_U - 1)} \times C
\end{aligned}
$$

5. Proof of generalisations of Miettinen's formula for continuous (or multi-category) exposure distributions

First for a strata of confounders, $C = c$ we will assume that

$RR_C(x,c) = \dfrac{P(Y=1 \mid X=x, C=c)}{P(Y=1 \mid X=0, C=c)} = RR_C(x)$ is constant over all confounder strata

$C = c$ for each exposure value x. We without loss of generality, that the Minimum risk exposure value is 0 (MREV = 0)

Then defining

$$
\begin{aligned}
PAF &= \frac{P(Y=1) - P(Y_0=1)}{P(Y=1)} \\
&= \frac{P(Y=1) - E_C(P(Y=1 \mid X=0, C))}{P(Y=1)} \\
&= \frac{E_{X,C}(P(Y=1 \mid X, C)) - E_C(P(Y=1 \mid X=0, C))}{P(Y=1)} \\
&= \frac{E_C(\int f(x \mid C)(P(Y=1 \mid X=x, C) - P(Y=1 \mid X=0, C))dx)}{P(Y=1)} \\
&= \frac{E_C(\int f(x \mid C)\frac{RR_C(x,C)-1}{RR_C(x,C)} P(Y=1 \mid X=x, C)dx)}{P(Y=1)} \\
&= E_{C \mid Y=1}(\int f(x \mid C, Y=1)\frac{RR_C(x,c)-1}{RR_C(x,c)}dx) \\
&= \int f(x \mid Y=1)\frac{RR_C(x)-1}{RR_C(x)}dx
\end{aligned}
$$

The second last line follows by expanding out the integral into a double integral over values of x and c and noting that:

$$f(x \mid C, Y=1) = \frac{P(Y=1 \mid C, X=x)f(C)f(x \mid C)}{f(C)P(Y=1 \mid C)} = \frac{P(Y=1 \mid C, X=x)f(x \mid C)}{P(Y=1 \mid C)}$$

so that $f(x \mid C, Y = 1)P(C \mid Y = 1) = f(x \mid C = c)P(Y = 1 \mid X = x, C = c)$ and $f(C \mid Y = 1) = \dfrac{f(C)}{p(Y = 1)}$, with the symbol f denotes a generic density function, with its arguments determining the precise density specified.

The last line follows by again expanding out the expression into a double integral over the values x and c, using Fubini's theorem to reverse the order of integration (so the inner integral is a function of c) and then noting that $E_{C|Y=1} f(x \mid C, Y = 1) = f(x \mid Y = 1)$ and $RR(x, C) = RR(x)$.

6. Proof of generalisation of Miettinen's formula for continuous (or multi-category) exposure distributions when causal relative risks vary over confounder strata:

The proof is essentially the same as the preceding proof. Note the penultimate line can be re-expressed as:

$$E_{C|Y=1}\left(\int f(x \mid C, Y=1)\frac{RR_C(x,c)-1}{RR_C(x,c)}dx = E_{C|Y=1}[E(X \mid C, Y=1)[\frac{RR_C(X,C)}{RR_C(X,C)}]\right]$$

which is the formula in question.

7. Proof of formula (7)

$$
\begin{aligned}
& f(x \mid Y=1) \\
&= \frac{f(x)P(Y=1 \mid x)}{\int_t f(t)P(Y=1 \mid t)} \text{ (Bayes' Rule)} \\
&= \frac{f(x)RR_u(x)}{\int f(t)RR_u(t)dt} \text{ (Divide above and below by } P(Y=1 \mid 0)) \\
&= \frac{f(x)RR_u(x)}{E_X(RR_u(X))}
\end{aligned}
$$

Implying that:

$$PAF = \int \frac{f(x)RR_U(x)}{E_X(RR_U(X))} \frac{RR_C(x)-1}{RR_C(x)} dx = \frac{E_X(RR_U(X)\frac{RR_C(X)-1}{RR_C(X)})}{E_X(RR_U(X))}$$

8. Analysis of the derivatives of the Relative Bias equation (4)

$$\frac{d}{dC}\frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)}\times C$$

$$=\frac{(1+\pi(RR_U-1))(1+\pi(C\times RR_U-1))-C\pi RR_U(1+\pi(RR_U-1))}{(1+\pi(C\times RR_U-1))^2}$$

$$=\frac{(1+\pi(RR_U-1))(1+\pi(C\times RR_U-1)-C\pi RR_U)}{(1+\pi(C\times RR_U-1))^2}$$

$$=\frac{(1+\pi(RR_U-1))(1-\pi)}{(1+\pi(C\times RR_U-1))^2}>\frac{(1-\pi)^2}{(1+\pi(C\times RR_U-1))^2}>0 \quad \text{for all C}$$

Derivative with respect to π:

$$\frac{d}{d\pi}\frac{1+\pi(RR_U-1)}{1+\pi(C\times RR_U-1)}\times C$$

$$=\frac{-(1+\pi(RR_u-1))(C\times RR_U-1))+(1+\pi(C\times RR_u-1))(RR_U-1)}{(1+\pi(C\times RR_u-1))^2}\times C$$

$$=\frac{RR_U(1-C)}{(1+\pi(C\times RR_u-1))^2}\times C$$

This derivative is negative for $C>1$ and positive for $C<1$

(In each case, smaller prevalences increase relative biases - but should decrease absolute biases)

$$\text{As } C \to \infty, \frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)} \times C \to 1+\frac{(1-\pi)}{\pi RR_U}$$

$$\text{As } C \to 0, \frac{1+\pi(RR_u-1)}{1+\pi(C\times RR_u-1)} \times C \to 0$$

9. Proof of Equation (10) - Relative bias for a positive continuous exposure with $MREV = 0$, where $0 < 1 - \pi = P(X = 0) < 1$ and $C = \dfrac{RR_C(X)}{RR_U(X)}$ is constant for $X > 0$

$$B = \frac{PAF_L}{PAF}$$

$$= \frac{E_X(RR_C(X)) - 1}{E_X(RR_C(X))} \frac{E_X(RR_U(X))}{E_X(RR_U(X)\frac{RR_C(X) - 1}{RR_C(X)})}$$

$$= \frac{1 - \pi + \pi E_{X|X>0}(RR_C(X)) - 1}{1 - \pi + \pi E_{X|X>0}(RR_C(X))} \frac{1 - \pi + \pi E_{X|X>0}(RR_U(X))}{\pi E_{X|X>0}(RR_U(X)\frac{RR_C(X) - 1}{RR_C(X)})}$$

$$= C \frac{1 - \pi + \pi E_{X|X>0}(RR_U(X))}{1 - \pi + \pi E_{X|X>0}(RR_C(X))}$$

$$= C \frac{1 + \pi(E_{X|X>0}(RR_U(X)) - 1)}{1 + \pi(C \times E_{X|X>0}(RR_U(X)) - 1)}$$

where the second last and last lines follow from the assumption that $RR_C(X) = C \times RR_U(X)$ when $X > 0$. The proof of equation (11) is almost identical.