Cancer Causes Control
DOI 10.1007/s10552-006-0090-y

ORIGINAL PAPER

# Point and interval estimates of partial population attributable risks in cohort studies: examples and software

**D. Spiegelman · E. Hertzmark · H. C. Wand**

Received: 8 July 2006 / Accepted: 28 October 2006
Springer Science+Business Media B.V. 2007

**Abstract** The concept of the population attributable risk (PAR) percent has found widespread application in public health research. This quantity describes the proportion of a disease which could be prevented if a specific exposure were to be eliminated from a target population. We present methods for obtaining point and interval estimates of partial PARs, where the impact on disease burden for some presumably modifiable determinants is estimated in, and applied to, a cohort study. When the disease is multifactorial, the partial PAR must, in general, be used to quantify the proportion of disease which can be prevented if a specific exposure or group of exposures is eliminated from a target population, while the distribution of other modifiable and non-modifiable risk factors is unchanged. The methods are illustrated in a study of risk factors for bladder cancer incidence (Michaud DS et al., New England J Med 340 (1999) 1390). A user-friendly SAS macro implementing the methods described in this paper is available via the worldwide web.

D. Spiegelman · E. Hertzmark
Departments of Epidemiology, School of Public Health,
Harvard University, 677 Huntington Avenue,
Boston, MA 02115, USA

D. Spiegelman (✉)
Departments of Biostatistics, School of Public Health,
Harvard University, 677 Huntington Avenue,
Boston, MA 02115, USA
e-mail: stdls@channing.harvard.edu

H. C. Wand
National Center in HIV Epidemiology and Clinical
Research, University of New South Wales, 376 Victoria
Street, Darlinghurst, NSW 2010, Australia

**Keywords** Population attributable risk · Relative risk · Epidemiologic methods · Cohort studies · Statistics · Burden of disease

## Introduction

What percent of cases would be prevented if it were possible to eliminate one or more exposures from a particular target population? The population attributable risk (PAR) answers this question. The PAR provides information about the public health significance of one or more exposures on the burden of disease in a population by accounting for both the strength of the association on the outcome and the prevalence of the exposure in the population to which the PAR is applied. The PAR was first formulated for a single binary exposure [1] and subsequently extended to the multivariate setting [2]. To calculate the PAR, one must estimate the relative risks for the risk factor(s) of interest as well as those for additional risk factors which may be potential confounders for the disease outcome in a multivariate model. In addition, prevalences must be estimated from the target population. A variety of names for the PAR have been used in the literature. According to a recent survey [3], the most common are attributable risk (AR) [1], etiologic fraction [4], attributable risk percentage [5] and attributable fraction [6]. A unified approach for the calculation of the attributable risk using multivariate models in case-control studies has been given, in which the concept of the partial PAR was first introduced [7]. A comprehensive overview of these methods, which discussed the issues to consider in correctly implementing PAR estimation techniques and interpreting the results was given later

[8]. Most of the literature has focused on point and interval estimation of the PAR in case-control studies.

In this paper, we derive the variance of the *partial* PAR where both the relative risks and population prevalences are estimated from the same cohort study. In a multifactorial disease setting, at least some key risk factors, such as age and family history, are not modifiable. This limits the practical utility of the *full* PAR, so we do not consider it further here. An example is given from a cohort study of risk factors for bladder cancer incidence in the Health Professionals' Follow-up Study (Michaud et al., 1999). The use of publicly available software in SAS is illustrated in this example.

## Full and partial PAR for cohort studies

The population attributable risk (PAR) is formulated as a function of relative risk(s) and the prevalence(s) of the risk factor(s). In its simplest form (Eq. 1), there is one exposure at two levels (exposed versus unexposed)

$$PAR = \frac{p(RR-1)\leftarrow}{p(RR-1)+1} = 1 - \leftarrow\frac{1}{p(RR-1)+1} = 1 - \frac{1}{\sum_{s=1}^{2} p_s RR_s} \quad (1)$$

where $RR$ is the relative risk, $p$ is the prevalence of the exposure in the population and $s$ indexes the two strata determined by the value of the risk factor.

Equation 1 was generalized to the multifactorial setting (Eq. 2), when there are multiple exposures at multiple levels, as

$$PAR_F = \frac{\sum_{s=1}^{S} p_s(RR_s-1)\leftarrow}{1+\sum_{s=1}^{S} p_s(RR_s-1)\leftarrow} = 1 - \frac{1}{\sum_{s=1}^{S} p_s RR_s} \quad (2)$$

In Eq. 2, $RR_s$ and $p_s$, $s = 1, \ldots, S$, are the relative risks and the prevalences in the target population for the $s$th combination of the risk factors. Eq. 2 evaluates the proportional reduction expected in the number of diseased individuals if all the known risk factors were eliminated from the target population. We will refer to this as the full $PAR$ ($PAR_F$). In an evaluation of a preventive intervention in a multifactorial disease setting, the interest is in the percent of cases associated with the exposures to be modified, when other risk factors, possibly non-modifiable, exist but do not change as a result of the intervention. The partial $PAR(PAR_p)$ was proposed [7] to estimate this quantity. The term *partial* here evokes the partial correlation coefficient in linear regression theory, involving the effect of a group of variables on an outcome after adjusting for the effects of another group. The $PAR_p$ is preferred over $PAR_F$ when the set of risk factors of interest includes some factors which cannot be modified (even theoretically), such as age and family history of the disease. Under the assumption of no interaction of the index exposure effects with the background risk factors, the $PAR_p$ is formulated as

$$PAR_p = \leftarrow\frac{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st}RR_{1s}RR_{2t} - \leftarrow\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st}RR_{2t}}{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st}RR_{1s}RR_{2t}} = 1 - \frac{\sum_{t=1}^{T} p_{.t}RR_{2t}}{\sum_{s=1}^{S}\sum_{t=1}^{T} p_{st}RR_{1s}RR_{2t}} \quad (3)$$

where $t$ denotes a stratum of unique combinations of levels of all background risk factors which are not under study, $t = 1, \ldots, T$ and $RR_{2t}$ is the relative risk in combination $t$ relative to the lowest risk level, where $RR_{2,1} = 1$. As previously, $s$ indicates an index exposure group defined by each of the unique combinations of the levels of the index risk factors, that is, those risk factors to which the $PAR_p$ applies, $s = 1, \ldots, S$, and $RR_{1s}$ is the relative risk corresponding to combinations relative to the lowest risk combination, $RR_{1,1} = 1$. The joint prevalence of exposure group $s$ and stratum $t$ is denoted by $p_{st}$, and $p_{.t} = \Sigma_{s=1}^{S} p_{st}$.

The partial PAR, as given by Eq. 3, represents the difference between the number of cases expected in the original cohort and the number of cases expected if all subsets of the cohort who were originally exposed to the modifiable risk factor(s) had eliminated their exposure(s) so that their relative risk compared to the unexposed was 1, divided by the number of cases expected in the original cohort.

To estimate $PAR_p$ or $PAR_F$ in a cohort study, one must first estimate the relative risks for the exposure(s) of interest and for the confounders, typically but not necessarily with a multiplicative model for the incidence rate of disease, $I(\boldsymbol{E}, \boldsymbol{C})$, such as

$$I(\boldsymbol{E},\boldsymbol{C}) = exp\{\boldsymbol{\beta}_1'\boldsymbol{E} + \boldsymbol{\beta}_2'\boldsymbol{C}\} \quad (4)$$

using a Poisson or pooled logistic regression model [9], where $E$ is a row vector of index exposure variables, and may include one or more binary or polytomous exposures and their interactions, $C$ is a row vector of background risk factors, usually including a row vector of indicator variables for age groups considered homogeneous with respect to disease risk, and may also include one or more binary or polytomous risk factors and their interactions. These models should include all higher order interactions suggested by the data, as usual, or the resulting $\widehat{PAR}$ will be biased.

There is a 1–1 relationship between $RR_{1s}$ and a reparameterization of Eq. 4. We define $\boldsymbol{E}$ as a vector of $p_1$ categorical variables, $\boldsymbol{E} = (E_1, \ldots, E_{p_1})$, which have $(S_1, \ldots, S_{p_1})$ levels. $\boldsymbol{C}$ is a vector of $p_2$ categorical variables, $\boldsymbol{C} = (C_1, \ldots, C_{p_2})$, which have $(T_1, \ldots, T_{p_2})$ levels. Without loss of generality, we assume that the reference levels in the set of binary indicator variables generated to represent $(\boldsymbol{E}', \boldsymbol{C}')'$, which must be the levels with lowest risk, are the first levels. We then generate the binary indicator variables $\boldsymbol{e} = (e_{12}, \ldots, e_{1S_1}, e_{22}, \ldots, e_{2S_2}, \ldots, e_{p_1 2}, \ldots, e_{p_1 S_{p_1}})'^{\leftarrow}$ and $\boldsymbol{c} = (c_{12}, \ldots, c_{1T_1}, c_{22}, \ldots, c_{2T_2}, \ldots, c_{p_2 2}, \ldots, c_{p_2 T_{p_2}})'$ of which the model

$$I(\boldsymbol{e}, \boldsymbol{c}) = exp\{\boldsymbol{\beta}_{\boldsymbol{1}}'\boldsymbol{e} + \boldsymbol{\beta}_{\boldsymbol{2}}'\overleftarrow{\boldsymbol{c}}\} \qquad (5)\leftarrow$$

is a function. Each unique set of possible values for $\boldsymbol{e}$ can be assigned a subscript $s$, $s = 1, \ldots, S$, where $S = \prod_{u=1}^{p_1} S_u$, and each unique set of possible values for $\boldsymbol{c}$ can be assigned a subscript $t$, $t = 1, \ldots, T$, where $T = \prod_{u=1}^{p_2} T_u$. For each $\boldsymbol{e_s}$, the corresponding relative risk for the index exposure variables, $RR_{1s}$, is $RR_{1s} = \exp\{\overleftarrow{\sum}_{j=1}^{S-p_1} \beta_{1j} e_{sj}\}$ and for each $\boldsymbol{c_t}$, the corresponding relative risk for the index background risk factors, $RR_{2t}$, is $RR_{2t} = \exp\{\overleftarrow{\sum}_{j=1}^{T-p_2} \beta_{2j} c_{tj}\}$. Following the conditions for confounding of the $\widehat{PAR}$ derived previously [10], unless age is either not a risk factor for the outcome of interest or is unassociated with the index exposure(s), the relative risks for age must be incorporated into the estimators given by Eqs. 2 and 3 [11, 12]. Hence, the Cox model, which in many epidemiologic applications conditions out the relative risks for age and assumes in its standard implementation no interactions with other model covariates [13], is typically not useful in this setting, unless age is jointly unassociated with all other risk factors in model, Eq. 5.

The prevalences for the combinations of background and index risk factors to be considered are estimated as multinomial probabilities from the person-time under follow-up in the cohort as the empirical fraction of person-time of follow-up among cohort members in each unique level of index exposures and background risk factors, and denoted $\hat{p}_{st}, s = 1, \ldots, S$; $t = 1, \ldots, T$. These are substituted into Eqs. 2 and 3. The asymptotic variance of $\widehat{PAR}_p$ is derived in Appendix 1 using the multivariate delta method for the cohort study setting, as given previously in a more general form [14, 15]. Appendix 2 illustrates the calculation of the $\widehat{PAR}_p$ and its 95% confidence limits with our user-friendly, fully-documented, publicly available macro (`http://www.hsph.harvard. edu/faculty/spiegelman/par.html`).

As seen in Eqs. 2–3, the $PAR$ is a function of the relative risks and the prevalences of the exposures and confounders. When the $PAR$ is estimated in a case-control study where the target population is the study base from which the cases arose, $Cov\left(\hat{p}_{st}, \widehat{RR}_{uv}\right)$ is non-zero when $s = u$ and $t = v$, and 0 otherwise. We show in Appendix 1 that, asymptotically, in a cohort study, $cov\left(\hat{p}_{st}, \widehat{RR}_{uv}\right) = 0, (s, u) = 1, \ldots, S, (t, v) = 1, \ldots, T$, as was given more generally previously [15].

The $PAR$ is not strictly additive. Additivity concerns the relationship between the $PAR$ for two or more risk factors to the sum of $PAR$s for each of these risk factors separately. The sum of the crude $PAR$s for each factor of interest obtained by collapsing over all other factors is generally less than the joint $PAR_F$ for the risk factors taken together [16]. However, the sum of the individual $PAR_p$s representing the effect of removing one risk factor while keeping other factors unchanged will generally be more than the $PAR_F$ for all the risk factors taken together [17].

Another important property of the $PAR$ is its distributivity [18]. The crude $PAR_F$ from a multilevel exposure equals the $PAR_F$ calculated from combining those categories into a single exposed category [2, 18, 19]. Insofar as the distributive property may hold approximately when there are several multilevel exposures, it may be statistically and computationally efficient to collapse categories, since even a modest number of multilevel exposures may create a very large number of joint levels with sparse information, leading to unstable prevalence estimates that will destabilize the overall $PAR_F$ or $PAR_p$. However, it should be noted that the distributive properly strictly holds only for the $\widehat{PAR}_F$, and will be an approximation for the $\widehat{PAR}_P$ [18].

## The role of fluid intake and cigarette smoking in bladder cancer prevention [20]

In the Health Professionals' Follow-up Study, fluid intake and cigarette smoking were the strongest modifiable risk factors for bladder cancer. We selected these two risk factors to examine the proportion of bladder cancer that could be prevented by certain public health interventions in 45,253 members of the Health Professionals' Follow-up Study, a cohort of male health professionals, who were followed between 1986 and 1996 for the incidence of bladder cancer, during which time 238 cases occurred among 442,508 person-years with complete index exposure data. Further details on this study have been given previously [20]. Fluid intake was ascertained at baseline through the reported frequency of 22 beverages. Current smoking status (yes/no) was updated every 2 years, and pack-years were given at baseline. Pooled logistic regression models adjusted for age in 5 year age groups, calendar year of questionnaire return (five periods), geographic region (five regions), baseline energy intake (in quintiles) and baseline intake of fruits and vegetables (four groups) were fit to the data to estimate the relative risks of these background risk factors, as well as the relative risks of the index exposures: fluid intake (quintiles), current smoking status (yes/no), and pack-years of smoking (six categories). Table 1 gives the frequency distribution of each of the background risk factors and the index exposures, and the relative risks of the index exposures and background risk factors. Based on these data and the methods discussed above, we calculated the $\widehat{PAR}_p$s corresponding to interventions focused on smoking cessation or prevention and increasing fluid intake (Table 2). If all HPFS cohort members increased their fluid intake to more than 2.4 liters per day,

**Table 1** Prevalences and relative risks for study of risk factors for bladder cancer in the Health Professionals Follow-up Study ($n$ = 45, 253)

| Variable | | Prevalence[a] (%) | Relative risk (95%CI) |
|---|---|---|---|
| Fluid intake (ml/day) | Quintile 5 | 20 | 1.0 |
| | Quintile 4 | 20 | 1.57 (0.98–2.54) |
| | Quintile 3 | 20 | 2.07 (1.30–3.30) |
| | Quintile 2 | 20 | 1.88 (1.15–3.05) |
| | Quintile 1 | 20 | 2.29 (1.41–3.72) |
| Current smoking | No | 92 | 1.0 |
| | Yes | 8 | 1.48 (1.00–2.17) |
| Pack-years of cigarette smoking | None | 48 | 1.0 |
| | < 10 | 10 | 1.44 (0.84–2.48) |
| | 10– < 25 | 19 | 1.94 (1.31–2.86) |
| | 25– < 45 | 14 | 2.44 (1.67–3.58) |
| | 45– < 65 | 7 | 2.88 (1.85–4.49) |
| | 65+ | 3 | 3.79 (2.30–6.24) |
| Region | West | 21 | 1.0 |
| | Midwest | 27 | 1.36 (0.88–2.11) |
| | South | 27 | 1.68 (1.10–2.56) |
| | Northeast | 23 | 1.91 (1.25–2.91) |
| | Pacific, missing | 1 | 1.33 (0.32–5.57) |
| Age (years) | < 50 | 27 | 1.0 |
| | 50– < 55 | 16 | 2.81 (1.29–6.16) |
| | 55– < 60 | 15 | 4.04 (1.94–8.42) |
| | 60– < 65 | 15 | 6.00 (2.97–12.12) |
| | 65– < 70 | 13 | 9.55 (4.82–18.91) |
| | 70– < 75 | 9 | 14.29 (7.18–28.45) |
| | 75– < 80 | 4 | 14.55 (6.85–30.91) |
| | 80+ | 1 | 27.60 (10.50–72.55) |
| Fruit and vegetable intake (servings/day) | 7.5+ | 25 | 1.0 |
| | 5– < 7.5 | 25 | 1.28 (0.89–1.83) |
| | 3.5– < 5 | 25 | 1.09 (0.73–1.64) |
| | < 3.5 | 25 | 1.42 (0.93–2.15) |
| Total Energy Intake (kcal/day) | Quintile 1 | 20 | 1.0 |
| | Quintile 2 | 20 | 1.35 (0.92–1.98) |
| | Quintile 3 | 20 | 1.04 (0.68–1.59) |
| | Quintile 4 | 20 | 1.09 (0.69–1.70) |
| | Quintile 5 | 20 | 1.37 (0.87–2.17) |
| Calendar period | 1994–1995 | 20 | 1.0 |
| | 1992–1993 | 20 | 1.31 (0.85–2.01) |
| | 1990–1991 | 20 | 1.77 (1.17–2.69) |
| | 1988–1989 | 20 | 2.04 (1.34–3.10) |
| | 1986–1987 | 20 | 1.52 (0.96–2.41) |

[a] Note: prevalences may not add to 100% due to rounding

**Table 2** $\widehat{PAR}_p$ (95% CI) for several risk factors for bladder cancer in the Health Professionals Follow-up Study [20]

| Exposure | $\widehat{PAR}_F$ from crude model | $\widehat{PAR}_p$ from multivariate model | $\widehat{PAR}_p$ from collapsed multivariate model |
|---|---|---|---|
| Fluid intake | 0.41 (0.15, 0.62) | 0.43 (0.17, 0.63) | 0.40 (0.16, 0.59) |
| Current smoking | 0.08 (0.03, 0.13) | 0.05 (–0.02, 0.12) | 0.07 (0.01, 0.13) |
| Pack-years of cigarette smoking | 0.50 (0.32, 0.64) | 0.43 (0.21, 0.62) | 0.41 (0.25, 0.55) |
| Fluid intake + current smoking | 0.49 (0.23, 0.69) | 0.46 (0.17, 0.67) | 0.44 (0.18, 0.64) |
| Fluid intake + pack-years of cigarette smoking | 0.77 (0.55, 0.89) | 0.68 (0.36, 0.86) | 0.65 (0.40, 0.81) |
| Current smoking + pack-years of cigarette smoking | 0.50 (0.28, 0.67) | 0.45 (0.20, 0.65) | 0.44 (0.27, 0.59) |
| Fluid intake + current smoking + pack-years of cigarette smoking | 0.77 (0.53, 0.90) | 0.69 (0.36, 0.87) | 0.67 (0.40, 0.83) |
| Number of combinations of index exposure and background risk factors observed in the study (of total possible) | 60 (of 60) | 66,155 (of 240,000) | 16,793 (of 32,000) |

an estimated 43% (95% CI 17%–63%) of bladder cancer would be avoided. If all HPFS cohort members increased their fluid intake to more than 2.4 liters per day and quit smoking, an estimated 46% (95% CI 17%–67%) of the incident cases of bladder cancer would be avoided. If all HPFS cohort members increased their fluid intake to more than 2.4 liters per day and had never smoked at all, an estimated 69% (95% CI 36%–87%) would have been avoided. Appendix 2 illustrates the calculation of these $\widehat{PAR}_p$s with our publicly available macro (`http://www.hsph.harvard.edu/faculty/spiegelman/par.html`).

Although the additivity approximation worked for the combined effects of increased fluid intake and smoking cessation (43% + 5% = 48%), while the correctly calculated $\widehat{PAR}_p$ was 46%, the additive approximation broke down more substantially for the combined effects of increased fluid intake and lifetime smoking prevention (43% + 5% + 43% = 91%), while the correctly calculated $\widehat{PAR}_p$ was 69% (Table 2). The $\widehat{PAR}_p$ for fluid intake when modeled by quintiles of intake was 43%, but when we grouped those with low fluid intake (below the fifth quintile) together into a single exposed group, the $\widehat{PAR}_p$ was 40% (Table 2). Hence, as noted previously [19], the distributive property often holds approximately in multifactorial disease settings although it is strictly true only for the full $PAR$ given by Eqs. 2 and 3. Interestingly, not only did the point estimates and confidence bounds differ for index exposures to which the distributive property was applied, but they also differ for binary risk factors in the model which in a univariate setting would not be affected by this change. For example, the $\widehat{PAR}_p$ for smoking cessation went from 5% (95% CI 2%–12%) to 7% (95% CI 1%–13%) when the distributive property was applied to pack-years of smoking and fluid intake. From a comparison of the ratio of the standard errors of the $\widehat{PAR}_p$ to the point estimates, there was no obvious efficiency gain here in collapsing risk factors categories to apply the distributive property approximation.

Some authors have incorrectly suggested that a $PAR_p$ can be validly estimated by using the simple formula $\widetilde{PAR}_p = \frac{\sum_{s=1}^{S} \hat{p}_{s.}\left(e^{\hat{\beta}_s}-1\right)}{1+\sum_{s=1}^{S} \hat{p}_{s.}\left(e^{\hat{\beta}_s}-1\right)}$, where $e^{\hat{\beta}_s}$ is the multivariate-adjusted relative risk comparing the *s*th level of the exposure to the reference level obtained by fitting (Eq. 5) by Poisson or pooled logistic regression and $\hat{p}_{s.}$ is the marginal prevalence of level *s* of exposure in the cohort study. With a bit of algebra, some re-arrangement of Eq. 3 reveals that unless $RR_{2t} = 1$ for all $t = 1, \ldots, T$ , or unless the index exposures are not associated with the background risk factors, i.e. unless $p_{st} = p_{s.}\, p_{.t}$, this method of estimating $\widehat{PAR}_p$ will be biased, as has been shown previously as early as 1983 [11, 12]. For example, the $\widehat{PAR}_p$ correctly calculated from Eq. 3 was 5.0% for cessation of smoking; using this incorrect method, it was under-estimated by 26% as 3.7%. That is, an estimated 5% of the incident cases of bladder cancer would have been eliminated in the Health Professionals' Follow-up Study if all those currently smoking quit. The $\widehat{PAR}_p$ correctly calculated from Eq. 3 was 40% for low fluid intake, defined as below the fifth quintile, using the distributive approximation which appeared to be reasonable here; using the incorrect method described above, little difference was seen—it was estimated as 39%. The correlations between the index exposures, fluid intake, current smoking and pack-years, and the highest risk background factors in our data are low. The highest correlation observed with an important background risk factor is that between age and pack-years, 0.19. It should be noted that in simulation studies, presumably with much higher correlations between index and background risk factors, severe bias has been reported when this biased method is used [11].

## Discussion

The variance of the partial $\widehat{PAR}$ was derived for cohort studies. It was noted that Poisson or pooled logistic regression models, rather than the Cox model, are needed to estimate the relative risks, because estimates of the relative risks of all background risk factors and index exposures are necessary, including those of age, if unbiased estimates of the $PAR_p$ are to be obtained. Investigators can switch from the perhaps more standard Cox regression analysis of their cohort study to Poisson or pooled logistic regression analysis by transforming the data into counting process format [21] (also known as person-time format), if it is not already in that form, from the one record per person structure, and by grouping the primary time variable, typically age, into a series of suitable indicator variables to be entered into the new model.

The methods were applied to a study of bladder cancer incidence in relation to increased fluid intake and smoking cessation and prevention. Publicly-available user-friendly software using a newly developed SAS macro was illustrated. Since to our knowledge, no other such software is publicly available, this addresses a significant need, as noted in Benichou's recent review [8]. It should be noted that Mezzetti et al. [22] provided a SAS macro for the point and interval estimation of the partial $\widehat{PAR}$ in case-control studies, based upon formulas given by [7, 23], where the estimated exposure prevalences are correlated with the estimated relative risks. In cohort studies, as shown in Appendix 1, these asymptotic correlations are 0, and hence the variance formula is not valid in this setting. In addition, the formula for the point estimate for the partial $\widehat{PAR}$ used in case-control studies uses an estimate of the proportion of cases that are exposed [24], rather than an estimate of the exposure prevalence in the study basis as the cohort study version does. In a cohort study, the latter quantity can be estimated with substantially more data, and hence, the estimator which uses estimates of exposure in the cases alone, although valid, is likely to be inefficient. To be certain of this conjecture, this issue would need further study.

As always, the estimated $PAR_p$ and its estimated confidence bounds will be valid only when the assumptions used to estimate it are valid. The relative risk model (Eq. 4), and consequently, its reparameterization, Eq. 5, must be correctly specified, and the risk factors not included in the intervention to be evaluated should not be intermediates on the causal pathways of any of the index exposures. As always, the relative risk and prevalence estimates are assumed to be unbiased estimates of their underlying parameters. For this to be true, it is assumed that no information bias, residual or unmeasured confounding, or selection bias is present.

Although the relative risks for the background risk factors and index exposures can typically be most validly estimated in a well designed epidemiologic cohort study, for the evaluation of public health interventions it is often of greatest interest to estimate the joint prevalences of the risk background factors and index exposures in a more general population to which these interventions may be applied, such as complex population-based surveys such as NHANES [25] or NHIS [26]. The variance of the $\widehat{PAR}_p$ for this situation has been derived [27] and some SAS code has been provided, although enhanced user-friendliness is needed for broader applicability. The specific derivation of the variance and implementation of software for $\widehat{PAR}_p$s calculated in cohort studies, which allow for interventions on some but not all of a polytomous index exposure (e.g. eliminating both under-weight and overweight in the prevention of ovulatory infertility) [28], and for $\widehat{PAR}_p$s which consider interventions that alter the prevalences of the index exposures without entirely eliminating high risk groups, also called the generalized impact fraction [15, 29, 30], are also needed. The partial population attributable risk can be a useful tool for translating the results of analytic epidemiology to public health practice.

**Acknowledgments** Supported by a grant from the National Institutes of Health (CA55075)

## Appendix 1: Derivation of the $Var\left(\widehat{PAR}_p\right)$

$$
\begin{aligned}
Var\left(\widehat{PAR}_p\right) &= Var\left(\frac{\sum_{t=1}^{T}\hat{p}_{.t}\widehat{RR}_{2t}}{\sum_{s=1}^{S}\sum_{t=1}^{T}\hat{p}_{st}\widehat{RR}_{1s}\widehat{RR}_{2t}}\right)\\
&= Var(f(\hat{\boldsymbol{p}},\widehat{\boldsymbol{RR_1}},\widehat{\boldsymbol{RR_2}})) \quad \left(\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial \boldsymbol{p}}\right)'_{\hat{p},\widehat{RR}}\\
&\quad Var\left(\hat{\boldsymbol{p}}\right)\left(\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial \boldsymbol{p}}\right)_{\hat{p},\widehat{RR}}\\
&\quad + \left[\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial\left(\boldsymbol{RR_1'},\boldsymbol{RR_2'}\right)'}\right]'_{\hat{p},\widehat{RR}} Var\left[\left(\widehat{\boldsymbol{RR}}_{\boldsymbol{1}}',\widehat{\boldsymbol{RR}}_{\boldsymbol{2}}'\right)'\right]\\
&\quad \left[\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial\left(\boldsymbol{RR_1'},\boldsymbol{RR_2'}\right)'}\right]_{\hat{p},\widehat{RR}}
\end{aligned}
\qquad (6)
$$

where

$$
\begin{aligned}
\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial p_{st}} &= \frac{b\,RR_{2t} - a\,RR_{2t}RR_{1s}}{b^2},\\
\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial RR_{1s}} &= -\frac{a\sum_{t=1}^{T}p_{st}RR_{2t}}{b^2},\\
\frac{\partial f(\boldsymbol{p},\boldsymbol{RR_1},\boldsymbol{RR_2})}{\partial RR_{2t}} &= \frac{bp_{.t} - a\sum_{s=1}^{S}p_{st}RR_{1s}}{b^2},\\
a = \sum_{t=1}^{T}p_{.t}RR_{2t}, &\quad b = \sum_{s=1}^{S}\sum_{t=1}^{T}p_{st}RR_{1s}RR_{2t},
\end{aligned}
$$

where $\Sigma_s\, p_{st} = p_{.t}$, and $\boldsymbol{RR_1} = \left(RR_{1,1}, RR_{1,2},\ldots,RR_{1S}\right)'$ and $\boldsymbol{RR_2} = \left(RR_{2,1}, RR_{2,2},\ldots,RR_{2T}\right)'$ are the vectors of the relative risks corresponding to the modifiable and unmodifiable risk factors respectively.

Under the proportional hazards model, $RR_{1s} = e^{\boldsymbol{\beta_1'e_s}}$, where $\boldsymbol{e_s}$ is the vector of values of the binary indicators corresponding to the $s$th combination of modifiable exposure variables, of which there are $S$ combinations, and $RR_{2t} = e^{\boldsymbol{\beta_2'c_t}}$ where $\boldsymbol{c_t}$ is the vector of values of the $t$th combination of unmodifiable background risk, of which there are $T$ combinations. Then, $Var\left[\left(\widehat{\boldsymbol{RR}}_1',\widehat{\boldsymbol{RR}}_2'\right)'\right] = D\Sigma D'$, where $\Sigma = Var\left[\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}',\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}'\right)'\right]$, and $D = [(D_{uv}), u = 1,\ldots,S+T, v = 1,\ldots,p_1+p_2]$ where

$$
D_{uv} = \begin{cases}
\frac{\partial RR_{1,u}}{\partial\beta_{1,v}} & \text{if } u\leqslant S \text{ and } v\leqslant p_1\\
\frac{\partial RR_{2,u-S}}{\partial\beta_{2,v-p_1}} & \text{if } u>S \text{ and } v>p_1\\
0 & \text{if } u\leqslant S \text{ and } v>p_1\\
0 & \text{if } u>S \text{ and } v\leqslant p_1
\end{cases}
$$

Under the proportional hazards model, $\frac{\partial RR_{1,u}}{\partial\beta_{1,v}} = e_{uv}e^{\boldsymbol{\beta_1'e_u}}$, where $e_{uv}$ is the $v$th element of the vector $\boldsymbol{e}_u$, and $\frac{\partial RR_{2,u-S}}{\partial\beta_{2,v-p}} = c_{u-S,v-p_1}e^{\boldsymbol{\beta_2'c_{u-S}}}$, where $c_{u-S,v-p_1}$ is the $v-p_1$th element of the vector $\boldsymbol{c}_{u-S}$.

The variance of the $\widehat{PAR}_p$ is estimated by replacing, in Eq. 6, $(p, RR)$ with $\left(\hat{p},\widehat{RR}\right)$, $\Sigma$ with the estimated variance-covariance matrix of $\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{1}}',\hat{\boldsymbol{\beta}}_{\boldsymbol{2}}'\right)'$ obtained from the pooled logistic regression model or Poisson regression model used to fit (Eq. 5). In a cohort study, the multinomial distribution is used to estimate the variance-covariance matrix of $\hat{p}$, where $p = (p_{1,1}, p_{1,2},\ldots,p_{ST})$, and $Cov(\hat{p}_{st},\hat{p}_{uv}) = \begin{cases}\hat{p}_{st}(1-\hat{p}_{st})/n & \text{if } s = u \,\&\, t = v\\ -\hat{p}_{st}\hat{p}_{uv}/n & \text{if } s\neq u \text{ or } u\neq v\end{cases}$, and $n$ is the total number of units of person-time of follow-up observed.

In the spirit of transformation suggested by Leung and Kupper [31], to improve the asymptotic behavior of the 95% confidence intervals of $\widehat{PAR}_p$ and to ensure that the confidence intervals remain within the range of –100% to 100%, it is useful to calculate the confidence intervals using the Fisher's Z transformation, that is

$$
\widehat{Var}\left[Fisherz\left(\widehat{PAR}_p\right)\right] \approx \frac{1}{\left[\left(1+\widehat{PAR}_p\right)\left(1-\widehat{PAR}_p\right)\right]^2}\widehat{Var}\left(\widehat{PAR}_p\right)
$$

Then the 95% confidence interval for the $\widehat{PAR}_p$ is estimated as

$$
\frac{e^{2\left(\widehat{PAR}_p \pm 1.96\sqrt{\widehat{Var}\left[Fisherz\left(\widehat{PAR}_p\right)\right]}\right)}-1}{e^{2\left(\widehat{PAR}_p \pm 1.96\sqrt{\widehat{Var}\left[Fisherz\left(\widehat{PAR}_p\right)\right]}\right)}+1},
$$

where $Fisherz\left(\widehat{PAR}_p\right) = \log\left[\sqrt{\frac{1+\widehat{PAR}_p}{1-\widehat{PAR}_p}}\right]$.

In a cohort study, it can be shown $Cov\left(\hat{\boldsymbol{p}},\hat{\boldsymbol{\beta}}\right) = 0$ by a double expectation argument: The estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{p}}$ are the solutions of the following estimating equations,

$$
\boldsymbol{U}_\beta(\boldsymbol{\beta},\boldsymbol{p}) = \sum_{i=1}^{n}\frac{\partial g(\boldsymbol{e}_i,\boldsymbol{c}_i;\boldsymbol{\beta})}{\partial(\boldsymbol{\beta}',\boldsymbol{p}')'}\left[Y_i - \mathrm{E}(Y_i|g(\boldsymbol{e}_i,\boldsymbol{c}_i;\boldsymbol{\beta}))\right] = \boldsymbol{0}
$$

$$
\begin{aligned}
\boldsymbol{U}_p(\boldsymbol{\beta},\boldsymbol{p}) &= \begin{pmatrix}\boldsymbol{0}_{(S+T-p_1-p_2)\times 1}\\ \sum_{i=1}^{n}\left[\boldsymbol{I}(\boldsymbol{e}_i=\boldsymbol{E}_s\,\&\,\boldsymbol{c}_i=\boldsymbol{C}_t)-\mathrm{E}(\boldsymbol{I}(\boldsymbol{e}_i=\boldsymbol{E}_s\,\&\,\boldsymbol{c}_i=\boldsymbol{C}_t))\right]\end{pmatrix}\\
&= \boldsymbol{0},
\end{aligned}
$$

where $Y_i$ is 1 if the unit of person-time is a case and 0 otherwise, $g(\boldsymbol{e}_i, \boldsymbol{c}_i; \boldsymbol{\beta})$ will typically be the expit or exponential function, depending on whether pooled logistic regression or Poisson regression is used to estimate $\boldsymbol{\beta}$, $\mathrm{E}(\cdot)$ is the expectation operator, and $\boldsymbol{I}(\cdot)$ is an $S + T$ vector of indicator functions which take values 1 when the condition inside the parentheses is true and 0 otherwise. Because they are unbiased score functions, $\mathrm{E}\left[\boldsymbol{U}_{\beta i}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)\right] = \boldsymbol{0}$ and $\mathrm{E}\left[\boldsymbol{U}_{pi}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)\right] = \boldsymbol{0}$, $i = 1, \ldots, n$. This implies that

$$
\begin{aligned}
&Cov\left[\boldsymbol{U}_{\beta i}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right), \boldsymbol{U}_{pi}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)\right]\\
&= \mathrm{E}_{Y_i, \boldsymbol{c}_i, \boldsymbol{e}_i}\left[\boldsymbol{U}_{\beta i}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right) \boldsymbol{U}'_{pi}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)\right]\\
&= \mathrm{E}_{\boldsymbol{c},\boldsymbol{e}}\mathrm{E}_{Y|\boldsymbol{c},\boldsymbol{e}}\left[\boldsymbol{U}_{\beta i}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right) \boldsymbol{U}'_{pi}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{p}}\right)\right]\\
&= \mathrm{E}_{\boldsymbol{c},\boldsymbol{e}}\mathrm{E}_{Y|\boldsymbol{c},\boldsymbol{e}}\left[\left(\frac{\partial g(\boldsymbol{e}_i, \boldsymbol{c}_i; \boldsymbol{\beta})}{\partial(\boldsymbol{\beta}', \boldsymbol{p}')'}\right)[Y_i - \mathrm{E}(Y_i|g(\boldsymbol{e}_i, \boldsymbol{c}_i; \boldsymbol{\beta}))]\right.\\
&\left.\left(\begin{pmatrix} \boldsymbol{0}_{(S+T-p_1-p_2)\times 1} \\ \sum_{i=1}^{n}[\boldsymbol{I}(\boldsymbol{e}_i = \boldsymbol{E}_s \;\&\; \boldsymbol{c}_i = \boldsymbol{C}_t) - \mathrm{E}(\boldsymbol{I}(\boldsymbol{e}_i = \boldsymbol{E}_s \;\&\; \boldsymbol{c}_i = \boldsymbol{C}_t))] \end{pmatrix}\right)'\right]\\
&= \mathrm{E}_{\boldsymbol{c},\boldsymbol{e}}\left[\left(\frac{\partial g(\boldsymbol{e}_i, \boldsymbol{c}_i; \boldsymbol{\beta})}{\partial(\boldsymbol{\beta}', \boldsymbol{p}')'}\right)\right.\\
&\left.\left(\begin{pmatrix} \boldsymbol{0}_{(S+T-p_1-p_2)\times 1} \\ \sum_{i=1}^{n}[\boldsymbol{I}(\boldsymbol{e}_i = \boldsymbol{E}_s \;\&\; \boldsymbol{c}_i = \boldsymbol{C}_t) - \mathrm{E}(\boldsymbol{I}(\boldsymbol{e}_i = \boldsymbol{E}_s \;\&\; \boldsymbol{c}_i = \boldsymbol{C}_t))] \end{pmatrix}\right)'\right] \times\\
&\mathrm{E}_{Y|\boldsymbol{c},\boldsymbol{e}}[Y_i - \mathrm{E}(Y_i|g(\boldsymbol{e}_i, \boldsymbol{c}_i; \boldsymbol{\beta}))] = \boldsymbol{0}.
\end{aligned}
$$

## Appendix 2. Sample SAS code for calculating the $\widehat{PAR}_P$

Program:

```
title 'make variance-covariance matrix of beta
  coefficients';
proc logistic descending data=all covout
  outest=betas;
  model bladder=
    volrnk0 volrnk1           /* lowest 4 quintiles of
  volrnk2 volrnk3                fluid intake */
    region1 region2           /* geographic regions */
  region3 region4
    agegrp2 - agegrp8         /* 5-year age groups */
    smkc                      /* current smoking */
    packyr2-packyr6           /* categories of pack-
                                 years */
    period1 period2           /* calendar time periods
  period3 period4                */
    calor2-calor5             /* highest 4 quintiles of
                                 caloric intake */
```

Continued

Program:

```
    fruv1-fruv3;          /* lowest 3 categories of
                             fruit-and-
                             vegetable intake */
title 'make dataset of joint prevalences of
  modifiable and unmodifiable risk
        factors';
proc sort data=all; by
    volrnk0 volrnk1 volrnk2 volrnk3
    region1 region2 region3 region4
    agegrp2 - agegrp8
    smkc
    packyr2-packyr6
    period1 period2 period3 period4
    calor2-calor5
    fruv1-fruv3;
run;
proc means noprint data=all; var bladder;
output out=phats n=fq;
by
    volrnk0 volrnk1 volrnk2 volrnk3
    region1 region2 region3 region4
    agegrp2 - agegrp8
    smkc
    packyr2-packyr6
    period1 period2 period3 period4
    calor2-calor5
    fruv1-fruv3;
run;
%par(bdata=betas, pdata=phats, n_or_p=n,
  n_or_pname=fq,
fixedvar=agegrp2 agegrp3 agegrp4 agegrp5 agegrp6
  agegrp7 agegrp8 period1
period2 period3 period4
region2 region3 region4 region5 calor2 calor3
  calor4 calor5
fruv862 fruv863 fruv861
modvar=smkc packyr2 packyr3 packyr4 packyr5
  packyr6
volrnk0 volrnk1 volrnk2 volrnk3);
```

Output:

```
option for the variance-covariance matrix of the
  prevalences is FIXED.
Partial PAR (95% CI) for
  modifiable vbls : VOLRNK0 VOLRNK1 VOLRNK2
  VOLRNK3 SMKC PACKYR2
PACKYR3 PACKYR4 PACKYR5 PACKYR6
  fixed vbls : AGEGRP2 AGEGRP3 AGEGRP4 AGEGRP5
  AGEGRP6 AGEGRP7 AGEGRP8
PERIOD1 PERIOD2 PERIOD3 PERIOD4 REGION2 REGION3
  REGION4 REGION5 CALOR2
CALOR3 CALOR4 CALOR5 FRUV862 FRUV863 FRUV861
          0.692 (0.366 , 0.869)
```

## References

1. Levin M (1953) The occurrence of lung cancer in man. Aca Unio Inter Contra Cancrum 9:531–541
2. Walter SD (1975) The distribution of Levin's measure of attributable risk. Biometrika 62:371–374
3. Uter W, Pfahlberg A (1999) The concept of attributable risk in epidemiological practice. Biom J 41(8):985–993

4. Miettinen O (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol 99(5):325–332
5. Cole P, Macmahon B (1971) Attributable risk percent in case-control studies. Br J Prevent Social Med 25(4):242
6. Last JM (1983) A dictionary of epidemiology. Oxford University Press, New York
7. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C (1985) Estimating the population attributable risk for multiple risk-factors using case-control data. Am J Epidemiol 122(5):904–913
8. Benichou J (2001) A review of adjusted estimators of attributable risks. Stat Methods Med Res 10:195–216
9. D'Agostino RB, Lee M-L, Belanger AJ, Cupples LA, Anderson K, Kannel WB (1990) Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham heart study. Stat Med 9:1501–1515
10. Walter SD (1980) Prevention for multifactorial diseases. Am J Epidemiol 112(3):409–416
11. Gefeller O (1992) Comparison of adjusted attributable risk estimators. Stat Med 11(16):2083–2091
12. Morgenstern H (1983) Morgenstern corrects a conceptual error (Letter). Am J Publ Health 73(6):703–703
13. Korn EL, Graubard BI, Midthune D (1997) Time-to-event analysis of longitudinal follow-up of a survey: choice of the time-scale. Am J Epidemiol 145(1):72–80
14. Basu S, Landis JR (1995) Model-based estimation of population attributable risk under cross-sectional sampling. Am J Epidemiol 142(12):1338–1343
15. Greenland S, Drescher K (1993) Maximum-likelihood-estimation of the attributable fraction from logistic-models. Biometrics 49(3):865–872
16. Benichou J, Chow WH, McLaughlin JK, Mandel JS, Fraumeni JF (1998) Population attributable risk of renal cell cancer in Minnesota. Am J Epidemiol 148(5):424–430
17. Wilson PD, Loffredo CA, Correa-Villasenor A, Ferencz C (1998) Attributable fraction for cardiac malformations. Am J Epidemiol 148(5):414–423
18. Wacholder S, Benichou J, Heineman E (1994) Attributable risk—advantages of a broad definition of exposure (vol 140, pg 303,1994). Am J Epidemiol 140(7):668–668
19. Benichou J (1991) Methods of adjustment for estimating the attributable risk in case-control studies; a review. Stat Med 10:1753–1773
20. Michaud DS, Spiegelman D, Clinton SK, Willett WC, Giovannucci EL (1999) Total fluid intake, specific beverages and bladder cancer risk in the health professional follow-up study. New England J Med 340:1390–1397
21. Therneau TM, Grambsch PM (2000) Modeling survival data: extending the cox model. Springer-Verlag, New York, New York
22. Mezzetti M, Ferraroni M, Decarli A, LaVecchia C, Benichou J (1996) Software for attributable risk and confidence interval estimation in case-control studies. Comput Biomed Res 29(1):63–75
23. Benichou J, Gail MH (1990) Variance calculations and confidence-intervals for estimates of the attributable risk based on logistic-models. Biometrics 46(4):991–1003
24. Miettinen OS (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention. Am J Epidemiol 99(5):325–332
25. Third National Health and Nutrition Examination Survey 1989–1994 (1996) NHANES II laboratory data file (CD-ROM). US Department of Health and Human Services National Center for Health Statistics, Hyattsville, MD
26. National Center for Health Statistics (1993) 1987 National health interview survey. Government Printing Office
27. Graubard BI, Fears TR (2005) Standard errors for attributable risk for simple and complex sample designs. Biometrics 61(3):847–855
28. Rich-Edwards JW, Spiegelman D, Garland M, Hertzmark E, Hunter DJ, Colditz GA et al (2002) Physical activity, body mass index and ovulatory disorder infertility. Epidemiology 13:184–190
29. Morgenstern H, Bursic E (1982) 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 7:292–309
30. Drescher K, Becher H (1997) Estimating the generalized impact fraction from case-control data. Biometrics 53(3):1170–1176
31. Leung HM, Kupper LL (1981) Comparisons Of Confidence-Intervals For Attributable Risk. Biometrics 37(2):293–302