Published in final edited form as:

Stat Med. 2018 April 15; 37(8): 1259–1275. doi:10.1002/sim.7559.

The effect of risk factor misclassification on the partial population attributable risk

Benedict H.W. Wong¹, Sarah B. Peskoe¹, and Donna Spiegelman²

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

²Departments of Biostatistics, Epidemiology and Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Abstract

The partial population attributable risk (pPAR) is used to quantify the population-level impact of preventive interventions in a multifactorial disease setting. In this paper, we consider the effect of nondifferential risk factor misclassification on the direction and magnitude of bias of pPAR estimands and related quantities. We found that the bias in the uncorrected pPAR depends nonlinearly and nonmonotonically on the sensitivities, specificities, relative risks, and joint prevalence of the exposure of interest and background risk factors, as well as the associations between these factors. The bias in the uncorrected pPAR is most dependent on the sensitivity of the exposure. The magnitude of bias varies over a large range, and in a small region of the parameter space determining the pPAR, the direction of bias is away from the null. In contrast, the crude PAR can only be unbiased or biased towards the null by risk factor misclassification. The semiadjusted PAR is calculated using the formula for the crude PAR but plugs in the multivariate-adjusted relative risk. Because the crude and semiadjusted PARs continue to be used in public health research, we also investigated the magnitude and direction of the bias that may arise when using these formulae instead of the pPAR. These PAR estimators and their uncorrected counterparts were calculated in a study of risk factors for colorectal cancer in the Health Professionals Follow-up Study, where it was found that because of misclassification, the pPAR for low folate intake was overestimated with a relative bias of 48%, when red meat and alcohol intake were treated as misclassified risk factors that are not modified, and when red meat was treated as the modifiable risk factor, the estimated value of the pPAR went from 14% to 60%, further illustrating the extent to which misclassification can bias estimates of the pPAR.

Keywords

attributable risk; bias; measurement error; risk factor misclassification

Correspondence: Donna Spiegelman, Department of Epidemiology, Harvard T.H. Chan School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA. stdls@hsph.harvard.edu.

ORCID

Benedict H.W. Wong <http://orcid.org/0000-0003-0619-8305>

Sarah B. Peskoe <http://orcid.org/0000-0002-1190-3606>

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

1 INTRODUCTION

The population attributable risk (PAR) is often defined as the percent reduction in disease incidence that would be observed if the exposure was to be entirely removed from the population, given the exposure distribution in the population to which the results are intended to apply. When the disease depends only on a single binary risk factor, the PAR is a simple, well-known function of the relative risk (RR) and the prevalence of the exposure.¹

Estimation of the PAR has increasingly become an important goal of research in cancer and other branches of epidemiology and public health, because it allows us to link the magnitude of estimated relative risks to the likely impact of public health interventions, which remove the harmful exposure or, at least, shift the distribution of the exposure so that it is less prevalent. Synonyms for the PAR include attributable risk, population attributable fraction, PAR proportion, and PAR percent.²

For example, it has become common in the genetic epidemiology of newly discovered gene variants and in meta-analysis of individual variants to report the PAR. Mocellin et al³ found that the PAR% for melanoma of the variant C allele of the XPD/ERCC2 single nucleotide polymorphism was 9.6%, where the $PAR\% = PAR \times 100\%$. Similarly, Hunter et al⁴ found that 4 single nucleotide polymorphisms in intron 2 of FGFR2 were associated with a PAR% of 16% for breast cancer. In addition to applications in genetic epidemiology, the PAR is an important summary statistic in quantification of the impact of environmental risk factors alone. For example, the PAR% in the Iowa Women's Health Study⁵ for smoking and the American Institute for Cancer Research recommendations for cancer prevention (World Cancer Research Fund Panel 1997) was 31% (95% confidence interval [CI], 19-37%) for cancer incidence and similar for cancer mortality. In our own Health Professionals Follow-up Study (HPFS), Platz et al⁶ reported that the partial PAR% for 6 modifiable colorectal cancer risk factors—obesity, physical inactivity, alcohol, early adulthood smoking, red meat, and low folic acid from supplements—was 71% (95% CI, 33-92%). More recently, in Harvard's Nurses' Health Study, Tamimi et al⁷ found that changing the risk factor profile to the lowest weight gain, no alcohol consumption, high physical activity, breast feeding, and no menopausal hormone therapy use was associated with a PAR% of 34.6 (95% CI, 22.7-45.4%) for breast cancer, with the PAR% for modifiable factors higher for estrogen receptor positive (ER+) (PAR% = 39.7) than estrogen receptor negative (ER-) (PAR% = 27.9) breast cancers.

In an evaluation of a preventive intervention in a multifactorial disease setting, interest is in the percentage of cases expected to be reduced with the exposures eliminated, when other risk factors, possibly nonmodifiable, exist but do not change as a result of the intervention. In the context of case-control studies, the partial population attributable risk (pPAR) was proposed to estimate this quantity.⁸ When the full PAR formula is incorrectly used in place of the pPAR formula, considering only the eliminated exposures but ignoring the unmodified background risk factors, the crude PAR is obtained. Whenever the risk factors are correlated, this will result in biased estimates as we will see later in this paper. There are also several instances where investigators have used relative risks adjusted for confounding, but applied

the crude PAR formula for the exposure of interest.^{9,10} We refer to this as the semiadjusted PAR.

Although it is widely acknowledged that the misclassification rates for genetic assays are quite low,¹¹ gene-environment interactions are of great interest, and environmental risk factors, notably dietary intake, physical activity, and environmental pollutants, are acknowledged to be measured with moderate to substantial error, with estimates of the correlation between the measured and true exposures of interest typically ranging between 0.4 and 0.6.¹² Hence, there is the need to develop methods to estimate pPAR values in multifactorial disease settings, which account for bias due to exposure misclassification.

It has been previously shown that under nondifferential exposure misclassification for a single binary exposure, the univariate PAR can either be estimated with no bias, underestimated, or in the opposite direction of the underlying true value.¹³ Bias due to exposure misclassification in the univariate risk factor setting has been studied widely,¹³⁻¹⁷ but not in the multifactorial setting where unmodified background risk factors exist and interventions may act on several risk factors simultaneously. In this paper, we explore the various parameters that affect the magnitude and direction of bias in the pPAR, considering misclassification in both the modifiable exposures of interest as well as nonmodified background risk factors.

We study bias in the uncorrected pPAR as a function of the sensitivities and specificities of the risk factors, the true relative risks, and the true prevalence of the risk factors. Because the crude and semiadjusted PARs continue to be used in public health research, we also investigate the magnitude and direction of the bias that may arise from using these formulae instead of the pPAR and/or from misclassification. In contrast to the univariate PAR, which can never be overestimated under nondifferential misclassification, we show that it is possible to overestimate the pPAR under certain conditions even with nondifferential misclassification, where “overestimate” in this paper refers to the fact that the theoretical value of the misclassified pPAR is greater than the true theoretical value of the pPAR.

The remainder of the paper is as follows: We present the expressions for the quantities of interest in Section 2 even with nondifferential misclassification. In Sections 3 and 4, we report the results of numerical studies of the bias over the multidimensional parameter space determining it, and observe how the bias varies as a function of these different parameters, when the other parameters are held constant. An illustrative example is presented in Section 5, in a study of the pPAR of risk factors for colorectal cancer in the HPFS.⁶ Section 6 concludes this paper with a summary and recommendations for future research.

2 NOTATION

In this section, we define alternative PAR estimands, with the goal of deriving their functional relationships to one another. Previously, we introduced the concept of the PAR and the pPAR. We will now define the various estimands in the absence and presence of misclassification.

2.1 The full PAR

When there are multiple exposures or a single exposure at multiple levels, let $s, s = 0, \dots, S$, indicates one of the unique combinations of the exposure levels. Without loss of generality, let $s = 0$ denotes the exposure combination with the lowest risk of disease. The full PAR is

$$par_F = 1 - \frac{1}{\sum_{s=0}^S p_s rr_s}, \quad (1)$$

where p_s is the proportion of the target population with the s th exposure combination and rr_s is the relative risk of disease for the s th exposure combination relative to the exposure combination with the lowest risk of disease, for $s = 0, \dots, S$.¹⁸ Note that $rr_0 = 1$ and $\sum_{s=0}^S p_s = 1$. Hence, par_F is interpreted as the expected proportional reduction in the number of diseased cases if all exposures were eliminated from the target population. The subscript F signifies the full or complete elimination of all exposures.

2.2 The partial PAR and crude PAR

In the presence of confounding due to background risk factors that are possibly nonmodifiable or that are not the target of a particular public health intervention, the above formula needs to be revised. Suppose instead that there are $S + 1$ combinations of the modifiable exposures and $T + 1$ combinations of the nonmodified background risk factors. Then, under the assumption of no interaction effects between the modifiable exposures and the nonmodified background risk factors, the pPAR is

$$par_P = 1 - \frac{\sum_{t=0}^T p_{.t} rr_{2t}}{\sum_{s=0}^S \sum_{t=0}^T p_{st} rr_{1s} rr_{2t}}, \quad (2)$$

where s denotes a stratum of unique combinations of levels of all modifiable exposures that are eliminated, t denotes a stratum of unique combinations of levels of all nonmodified background risk factors, p_{st} indicates the proportion of the population with respective risk levels s and t , with 0 again indexing the lowest risk strata, rr_{1s} is the relative risk of disease for stratum s relative to the lowest risk stratum, $s = 0, \dots, S$, and rr_{2t} is the relative risk in stratum t relative to the lowest risk stratum, $t = 0, \dots, T$. Note that $rr_{10} = rr_{20} = 1$ and $p_{.t} = \sum_{s=0}^S p_{st}$ for all t .¹⁹ Each of the relative risk estimands has 2 subscripts, the first of which indicates whether it corresponds to the set of modifiable exposures or the set of nonmodified background risk factors, and the second subscript indicates to which stratum it belongs. Note that the term “exposures” refers to the risk factors of primary interest, which must be “modifiable,” and we refer to the others as the “nonmodified” background risk factors.

In the scenario where there is only one modifiable exposure with 2 risk levels, and 1 nonmodified background risk factor with 2 levels, so that $S = T = 1$, a derivation for the joint prevalence $\{p_{st} \text{ for } (s, t) \in \{0, 1\}^2\}$ is given in Appendix S1.

Using the formula for par_F in Equation 1 in the presence of confounding due to background risk factors, as is nearly always the case in observational research, would give us a crude estimate of the PAR that is invalid.²⁰ Here, the crude PAR is

$$par_c = 1 - \frac{1}{\sum_{s=0}^S p_{s.} rr_{1s}^{(c)}}, \quad (3)$$

where the $rr_{1s}^{(c)}$, $s = 0, \dots, S$, are the crude relative risks, unadjusted for the background nonmodified risk factors, with $rr_{10}^{(c)} = 1$, and $p_{s.} = \sum_{t=0}^T p_{st}$ for all s .

When we have one modifiable exposure and one nonmodified background risk factor, with both of them being binary such that $S = T = 1$, Equations 2 and 3 reduce respectively to

$$par_P = 1 - \frac{\sum_{t=0}^1 p_{.t} rr_{2t}}{\sum_{s=0}^1 \sum_{t=0}^1 p_{st} rr_{1s} rr_{2t}} = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11}) rr_{2T}}{p_{00} + p_{10} rr_{1S} + p_{01} rr_{2T} + p_{11} rr_{1S} rr_{2T}} \quad (4)$$

and

$$par_c = 1 - \frac{1}{\sum_{s=0}^1 p_{s.} rr_{1s}^{(c)}} = 1 - \frac{1}{(p_{00} + p_{01}) + (p_{10} + p_{11}) rr_{1S}^{(c)}}, \quad (5)$$

where rr_{1S} and rr_{2T} are the relative risks for the disease given the modifiable exposure and nonmodified background risk factor, respectively, p_{st} is the proportion of the population with level s of the modifiable exposure and level t of the nonmodified background risk factor, and $rr_{1S}^{(c)}$ is the crude relative risk of having the disease given the modifiable exposure,

disregarding the presence or absence of the nonmodified background risk factor. In Appendix S2, we provide an expression for $rr_{1S}^{(c)}$ as a function of the marginal prevalence $p_{S.}$ and $p_{.T}$, the relative risks rr_{1S} and rr_{2T} , and the relative risk of having the modifiable exposure given the nonmodified background risk factor.

Without loss of generality, we can drop the subscripts S and T from the relative risk estimands in Equation 4 when the modifiable exposure and the nonmodified background risk factor are both binary and rewrite the equation as

$$par_P = 1 - \frac{p_{.0} + p_{.1}rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2} = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2}. \quad (6)$$

In Equation 6, we assumed that there was no interaction effect between the modifiable exposure and the nonmodified background risk factor. However, in practice, this assumption may not always hold. Hence, the above equation can be written more generally as

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_3}, \quad (7)$$

where rr_3 is the relative risk of disease between the population stratum who have both risk factors compared with the stratum with neither risk factor and rr_3 is not necessarily equal to rr_1rr_2 . In the presence of misclassification in the modifiable exposure and/or the nonmodified background risk factor, the uncorrected pPAR is given by

$$PAR_P = 1 - \frac{(P_{00} + P_{10}) + (P_{01} + P_{11})RR_2}{P_{00} + P_{10}RR_1 + P_{01}RR_2 + P_{11}RR_3}, \quad (8)$$

where uppercase letters are used throughout this paper to indicate that the parameters are misclassified and hence need to be corrected using the sensitivity and specificity values obtained from a validation study.

Equations 7 and 8 give the expressions for 2 estimands, the true pPAR (par_P) and uncorrected pPAR (PAR_P).

The purpose of this paper is to understand the parametric relationship between par_P , the quantity of interest, and PAR_P , the convergent value, or estimand, of $\widehat{PAR}_P$ when the modifiable and/or nonmodified risk factors are misclassified. Throughout this paper, we use the term “bias” to refer to the absolute and relative differences between the estimands $\widehat{PAR}_P$ and $\widehat{par}_P$, under nondifferential misclassification. Misclassification is said to be differential if the misclassification parameters, namely, the sensitivity and specificity of the risk factors, differ by disease status. If X is an indicator variable denoting the presence of a risk factor and Z is the indicator variable for the surrogate risk factor, then the risk factor sensitivity is defined as $\theta = P(Z = 1|X = 1)$, and its specificity is defined as $\phi = P(Z = 0|X = 0)$.

We use θ_S and θ_T to denote the sensitivities of the modifiable and modifiable binary risk factors (X_S and X_T), respectively, and likewise ϕ_S and ϕ_T for the respective specificities.

2.3 The semiadjusted PAR

In the context of multifactorial models for disease incidence, there are a number of examples where investigators have used relative risk values that have been adjusted for confounding, but applied the crude PAR formula for the exposure of interest, eg, Cole and MacMAHON⁹ and Morgenstern.¹⁰ We refer to this as the semiadjusted PAR:

$$par_{\text{semi}} = 1 - \frac{1}{(1 - p_{1.}) + p_{1.} rr_1^{(a)}}, \quad (9)$$

where $rr_1^{(a)}$ is the multivariate-adjusted relative risk of the modifiable exposure. If the relative risks of the modifiable exposure and the nonmodified background risk factor are multiplicative as in Equation 6 in the absence of an interaction effect, then $rr_1^{(a)} = rr_1$. In practice, this may not always hold true.

In Equation 9, the crude relative risk of Equation 3 is replaced by the adjusted relative risk $rr_1^{(a)}$. However, the formula for par_F has still been used, rather than par_P as needed. This method has been shown to be biased,²¹ with the direction of the bias going both ways and the magnitude of the bias varying according to the parameters of the model.²²

Although biased, because this method continues to be used, we will briefly explore the impact of misclassification on it. The formula for the uncorrected semiadjusted PAR is given by

$$PAR_{\text{semi}} = 1 - \frac{1}{(1 - P_{1.}) + P_{1.} RR_1^{(a)}}, \quad (10)$$

where $RR_1^{(a)}$ is the uncorrected multivariate-adjusted relative risk of the surrogate exposure in the presence of risk factor misclassification.

3 THE EFFECT OF MISCLASSIFICATION ON PAR ESTIMANDS

In this section, we explore the bias in the various PAR estimands when there is misclassification in the modifiable exposure and the nonmodified background risk factor. The plural term “risk factors” is used to refer to both the modifiable exposure and the background risk factor.

In the scenario where both risk factors are binary, there are 10 parameters that can affect the bias of PAR_P . They are θ_S , ϕ_S , θ_T , ϕ_T , p_S , p_T , rr_{1S} , rr_{2T} , rr_{ST} , p_d where p_d is the prevalence or cumulative incidence of the disease and rr_{ST} is the relative risk of having the modifiable exposure given the nonmodified background risk factor, defined by

$$rr_{ST} = \frac{P(X_S = 1 | X_T = 1)}{P(X_S = 1 | X_T = 0)} \text{ with } rr_{ST} > 1 \text{ indicating a positive association between the}$$

modifiable exposure and the background risk factor. While p_d has no effect on the bias (see Section 3.1), the other 9 parameters have nonlinear nonmonotonic effects on the magnitude and direction of the bias. In Sections 3.1 and 3.2, we investigate general trends surrounding the behavior of the bias, and in Section 3.3, we provide graphs to illustrate the individual effects of each of these 9 parameters while holding the other 8 constant.

In this paper, we assume that misclassification of the modifiable exposure is independent of the misclassification of the background risk factor. We also assume that there is no interaction between the exposure and the background risk factor, so that the relative risks are multiplicative in the definition of the true pPAR, par_P . The appearance of a multiplicative interaction may, however, be induced when one or both risk factors have been misclassified,²³ and we allow for this phenomenon when deriving the expression for the uncorrected pPAR, PAR_P . In Equation 12 in Appendix S3, the expression for the disease probability as a function of the misclassified risk factors is derived. It is apparent that following misclassification in the risk factors, this expression for the probability is highly nonlinear.

3.1 Theoretical findings

Although p_d is one of the 10 parameters of the pPAR, we show in Appendix S4 that the various PAR estimands are independent of p_d . In addition, we identified 2 conditions under which PAR_P is equal to par_P even in the presence of some misclassification, the first of which is when $\theta_S = \theta_T = \phi_T = 1$ and the second is when $\theta_S = \pi_{ST} = 1$. These conditions are further described in Section 3.2.1.

3.2 Numerical results

We conducted a numerical bias evaluation over the 9-dimensional parameter space, with the goal of answering the question: “In which region of the parameter space are we more likely to observe bias that is away from the null?” As the parameter space is continuous for each of the 9 parameters, we numerically evaluated the theoretical convergent values of the 6 estimands at fixed points within this hyper-grid and determined the magnitude and direction of the bias in the incorrect estimands relative to the true pPAR. Each of θ_S , ϕ_S , θ_T , ϕ_T took values in $\{0.55, 0.70, 0.85, 1.00\}$, while p_S , p_T took values in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, π_{1S} , π_{2T} took values in $\{1.25, 2.5, 5\}$, and π_{ST} took values in $\{0.70, 0.85, 1.00, 1.15, 1.30\}$. This range of parameter values corresponds to those typically seen in practice, and we chose the values to be evenly spaced within the grid, resulting in a grid composed of 266 335 points. The range of values for the misclassification parameters, namely, the sensitivities and the specificities, were chosen to include a value just greater than 0.50, because values below 0.50 indicate that the risk factors were more likely to be misclassified than correctly classified, which is possible but unlikely to be seen in practice. We chose the range of values for the relative risks to reflect both low and high degrees of association between the disease and risk factors. The values for the marginal prevalence ranged from 0.1 to 0.9 so that our numerical bias evaluation would include both uncommon risk factors as well as highly prevalent risk factors, in addition to moderately prevalent risk factors.

For each set of parameters, we calculated the values for the true and the misclassified pPARs, par_P and PAR_P . From these, we calculated the absolute bias and the relative bias,

given by $\text{bias}_A = \text{PAR}_P - \text{par}_P$ and $\text{bias}_R = \frac{\text{PAR}_P - \text{par}_P}{\text{par}_P}$, respectively. Please note that because

we were able to derive closed-form expressions for each estimand of interest, there are no simulation experiments in this paper. Rather, in what follows below, we conduct a systematic comparison of the estimands, that is, the convergent values of a number of alternative estimators for the pPAR, to their intended target, the pPAR.

3.2.1 When does $\text{PAR}_P = \text{par}_P$?—We found that there is no bias when $\theta_S = 1$ and when $\theta_T = \phi_T = 1$ or $\pi_{ST} = 1$ or both. These are sufficient but not necessary conditions for zero bias, as there are other sets of parameter values that also result in the PAR_P being equal to par_P in the presence of risk factor misclassification. For instance, in Figure 1, we see that the curves for bias_R cross the line for no bias at 3 distinct values of θ_S for the 3 different values of ϕ_S shown. The conditions for no bias depend on the values of the other parameters, which are in this figure set to $\theta_T = \phi_T = p_S = p_T = 0.75$, $\pi_{1S} = 1.25$, $\pi_{2T} = 3$, $\pi_{ST} = 1.1$. These default parameters were chosen to be similar to the parameters observed in the HPFS of risk factors for colorectal cancer discussed²⁴ in Section 5.

When $\theta_S = \theta_T = \phi_T = 1$, we provide a proof in Appendix S5, which shows that the bias is exactly 0.

We also showed that when $\pi_{ST} = 1$, that is, when the 2 risk factors are independent and the relative risks are multiplicative, the bivariate pPAR reduces to the univariate PAR, as shown in Appendix S6. Using Maple, we computed the expression for the bias and verified that when $\theta_S = \pi_{ST} = 1$, the bias is exactly 0. These results on the relative risk scale are in agreement with previously published findings for the full PAR.¹³

It is also of interest to know when PAR_P is approximately equal to par_P . Using the values from the numerical bias evaluation, we identified the points in the parameter space where the relative bias had a magnitude less than 10%. There were 56 566 such occurrences among the 266 335 unique combinations of the parameters determining par_P and PAR_P explored, indicating that severe relative bias of at least 10% in either direction is found in 79%, or more than three-quarters, of the parameter space studied. Using a logistic regression model for small bias, defined as relative bias between -10% and 10%, to identify which of the 9 parameters were most strongly determining bias, we found that θ_S was by far the strongest determinant with an impact 27 times greater than the next most important parameter, π_{ST} . When θ_S was close to 0.50, the relative bias is highly negative on average. However, as θ_S increases towards 1, there were more scenarios where the relative bias was within (-10%, 10%).

3.2.2 When is $\text{PAR}_P > \text{par}_P$?—In the range of the parameter space explored, positive bias only occurred when $\pi_{ST} > 1$ and $\theta_T + \phi_T < 2$. The latter condition is satisfied almost all the time, except in very rare occasions when the background risk factor is perfectly measured, in which case $\theta_T + \phi_T = 2$. From the numerical bias evaluation, it was apparent that these conditions are necessary but not sufficient for positive bias, which tends to happen more often at larger values of θ_S . The first condition is met when the 2 risk factors are positively associated. The second condition is met when there is at least some

misclassification in the background risk factor, either imperfect sensitivity or imperfect specificity or both. Although we were unable to prove that $rr_{ST} > 1$ and $\theta_T + \phi_T < 2$ are necessary but not sufficient conditions for the bias in par_P to be positive, this phenomenon was apparent from the numerical bias evaluation; that is, there were no instances of positive bias unless these 2 conditions were met.

By dichotomizing $bias_R$ into 2 categories, positive vs nonpositive, and performing a logistic regression of the bias on the 9 parameters among the 266 335 scenarios evaluated, we observed that the 2 most important parameters leading to positive bias were θ_S and rr_{ST} , with coefficients 15 to 7 times greater than the next 2 most important determinants of positive bias, θ_T and ϕ_T . The occurrence of positive bias was strongly associated with larger values of θ_S and rr_{ST} and also with lower values of θ_T and ϕ_T .

We present the results of the numerical bias evaluation in Table 1, which shows how often the various estimands are biased towards or away from the null, relative to par_P . We see that the cell percentage numbers for the misclassified estimands are similar to each other, and we also see that the bias in par_{semi} is concentrated mainly within the interval $(-10\%, 10\%)$.

3.2.3 General observations—Over the range of the 9 background parameters explored, it was apparent from Table 1 that most of the time PAR_P would lead to estimates that would converge to values moderately or greatly underestimating the true $pPAR$. In just a smaller region of the parameter space, would PAR_P lead to an overestimated value, although interestingly this occurred in our motivating data example. With a slight abuse of terminology, we use the word “underestimate” in this paper to refer to the fact that the theoretical values of PAR_P and other estimands are lower than the theoretical value of par_P . Likewise, the word “overestimate” is used when the theoretical values of the estimands are greater than the theoretical values of par_P .

In Table 2, $\frac{\partial bias_R}{\partial \phi_S}$ was calculated by taking the difference between $bias_R$ values, obtained from 2 adjacent points with different ϕ_S but identical values for all other parameters, divided by the difference of the 2 ϕ_S values, to obtain a finite difference approximation to the partial derivatives. From Table 2, it is apparent that the bias of the PAR_P depends nonlinearly on each of the 9 parameters and that bias can both increase and decrease as each background parameter increases, depending on the values of the other background parameters.

3.3 Graphical exploration

We next demonstrate graphically how the relative bias varies with θ_S at different levels of one other parameter while the remaining parameters are held constant at a chosen set of parameter values. We chose θ_S to be one of the parameters simultaneously varied because the findings in Section 3.2 suggest that the relative bias of PAR_P may go from negative to positive as θ_S increases. The default values of the parameters, when held constant, were chosen to be similar to the parameters observed in the HPFS of risk factors for colorectal cancer discussed²⁴ in Section 5. They are $p_S = p_T = \phi_S = \theta_T = \phi_T = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3.0$, $rr_{ST} = 1.1$.

In each graph, we observed the change in relative bias as θ_S is varied over the interval (0.5, 1.0) and at selected values of another parameter. We selected only a handful of graphs examined in the course of this research, where the more interesting phenomena were observed. We plotted the relative bias in the pPAR against θ_S , exposure sensitivity, for different values of the other parameters, and observe the behavior of the bias.

3.3.1 Effect of the misclassification parameters on PAR_P—In Figure 1, the different curves correspond to different values of ϕ_S . These curves converge as θ_S approaches 1. Positive bias, corresponding to overestimation of the pPAR, is observed only when θ_S is relatively high and depends also on ϕ_S . When $\theta_S = 0.95$, the pPAR is underestimated to a greater extent at lower ϕ_S levels. Recall that “underestimate” in this paper refers to the fact that the theoretical values of PAR_P and other estimands are lower than the theoretical value of par_P. Results from our numerical bias evaluation in Table 2

corroborated the first observation in that $\frac{\partial \text{bias}_R}{\partial \phi_S}$, the change in the relative bias with respect

to ϕ_S , is zero when θ_S is 1. The numerical bias evaluation also showed that $\frac{\partial \text{bias}_R}{\partial \phi_S}$ is positive

more often than it is negative or zero, indicating that bias, which is generally negative, tends to decrease in magnitude as ϕ_S increases. This was seen in 99.97% of the cases explored in Table 2, the only exceptions being those cases with $rr_{1S} = 1.25$, $rr_{2T} = 5$, and $rr_{3T} = 0.7$.

Now, holding ϕ_S and all other parameters but ϕ_T constant at the default values, we see that the pPAR is again only overestimated at high values of θ_S , where “overestimate” in this paper refers to the fact that the theoretical values of PAR_P and other estimands are greater than the theoretical value of par_P. Nearly half of the time, an increase in ϕ_T resulted in a greater negative bias, whereas at higher levels of θ_S , an increase in ϕ_T resulted in a smaller positive bias. The trends in Figure 2 are consistent with data from the most comprehensive results from the numerical bias evaluation for $rr_{3T} > 1$, and similar results were obtained when the roles of ϕ_T and θ_T were switched, so that θ_T is now the parameter being varied, along with θ_S , while ϕ_T and the other parameters were kept constant. In both Figures 1 and 2, we observed that the relative bias became either less negative or more positive as θ_S increased. This too was corroborated by the results of the numerical bias evaluation, where we found that the $\frac{\partial \text{bias}_R}{\partial \theta_S} > 0$ more than 99% of the time. The cases where $\frac{\partial \text{bias}_R}{\partial \theta_S} < 0$

happened at low values of rr_{3T} and rr_{1S} and high values of rr_{2T} . We also saw that for a given level of ϕ_S , θ_T , or ϕ_T , the value of θ_S for which there is no bias differs.

3.3.2 Effect of the risk factor prevalence, of the association of the modifiable exposure with the background risk factor, and of the relative risks on PAR_P

In Figure 3, for each of the p_S values considered, the relative bias decreased in magnitude as θ_S increased and then increased in the positive direction. For different values of p_S , the corresponding curves behaved differently. For example, the curve for $p_S = 0.05$ had a linear appearance, whereas the curve for $p_S = 0.95$ appeared nonlinear. The relative bias spanned nearly the whole interval of (−100, 100) for $p_S = 0.95$ as θ_S increased from 0.5 to 1.0, with all other parameters kept constant. From the graph, we also saw that the relative bias both

increased and decreased with p_S . Data from our numerical bias evaluation showed that $\frac{\partial \text{bias}_R}{\partial p_S} > 0$ in 4% of the cases, $\frac{\partial \text{bias}_R}{\partial p_S} = 0$ in 2% of the cases, and was negative otherwise.

There are interesting results when p_T is the parameter varied instead. The curves in Figure 4 were very much alike and were only distinguishable from one another as $\theta_S \rightarrow 1$. The relative bias were all positive at or before $\theta_S = 1$, with the highest relative bias observed when $p_T = 0.5$ and the lowest when $p_T = 0.95$. Data from our numerical bias evaluation showed that $\frac{\partial \text{bias}_R}{\partial p_T} = 0$ in 10% of the cases, and in half of the remaining cases, $\frac{\partial \text{bias}_R}{\partial p_T}$ was positive.

When $\pi_{S/T} = 1$ and $\theta_T = \phi_T = 1$, $\frac{\partial \text{bias}_R}{\partial \pi_{1S}} = 0$. Otherwise, $\frac{\partial \text{bias}_R}{\partial \pi_{1S}} > 0$ when $\pi_{S/T} < 1$ and $\frac{\partial \text{bias}_R}{\partial \pi_{1S}} < 0$ when $\pi_{S/T} > 1$.

Turning our attention to π_{2T} , we found that $\frac{\partial \text{bias}_R}{\partial \pi_{2T}} = 0$ when either $\pi_{S/T} = \theta_S = 1$ or $\theta_S = \theta_T = \phi_T = 1$. When neither of the above 2 conditions hold, and when $\pi_{S/T} < 1$ then $\frac{\partial \text{bias}_R}{\partial \pi_{2T}} < 0$ in a majority of cases studied. Likewise, when $\pi_{S/T} < 1$, then $\frac{\partial \text{bias}_R}{\partial \pi_{2T}} > 0$ in a majority of the cases studied. Overall, when data from our numerical bias evaluation were aggregated, the frequency for which $\frac{\partial \text{bias}_R}{\partial \pi_{2T}} > 0$ was approximately equal to the frequency for which $\frac{\partial \text{bias}_R}{\partial \pi_{2T}} < 0$.

4 BIAS IN THE SEMIADJUSTED AND CRUDE PAR ESTIMANDS

The magnitude and direction of the bias in par_{semi} , PAR_{semi} , par_C , and PAR_C over the 266 335 point numerical bias evaluation is given in Table 1. Although par_{semi} did show some bias, bias was within $\pm 10\%$ in all but 1% of the parameter space explored. This is reassuring, as this measure remains commonly used in the literature. Once misclassification was introduced, the pattern of bias of PAR_{semi} was quite similar to that of PAR_P . From Table 3, it is apparent that the bias in par_{semi} changed nonlinearly and nonmonotonically with respect to 4 parameters upon which it depends, p_S , p_T , π_{1S} , and π_{2T} . In all cases considered in the numerical bias evaluation, the bias in par_{semi} decreased as $\pi_{S/T}$ increased. The bias in PAR_{semi} nearly always increased as θ_S and to a somewhat less extent similarly as ϕ_S increased. The bias in PAR_{semi} changed nonlinearly and nonmonotonically with respect to the other 7 parameters. Bias in par_C was greater than that for par_P and par_{semi} (Table 1), and the bias in PAR_C was large and similar in direction to that of PAR_{semi} and PAR_P . The bias in both par_C and PAR_C increased with $\pi_{S/T}$ (Table 4). The bias in par_C varied nonlinearly and nonmonotonically with p_S , p_T , π_{1S} , and π_{2T} . The bias in PAR_C varied similarly with respect

to these parameters, and nonmonotonically and nonlinearly with respect to ϕ_S , θ_T , and ϕ_T , but $\frac{\partial \text{bias}_R}{\partial \theta_S}$ for PAR_C was positive for all values considered in the numerical bias evaluation.

This means that the absolute bias decreases as long as the bias itself is negative, but after the bias becomes positive, the absolute bias begins increasing.

Next, we explored the behavior of the bias in the crude and semiadjusted estimators graphically. As in Section 3, the default values of the parameters, when held constant, were chosen to be similar to the parameters observed in the HPFS of risk factors for colorectal cancer.²⁴ They were $p_S = p_T = \phi_S = \theta_T = \phi_T = 0.75$, $\theta_S = 0.97$, $rr_{1S} = 1.25$, $rr_{2T} = 3.0$, $rr_{3T} = 1.1$. In each graph, we studied the effect of varying one parameter at a time, while holding the rest constant at their default values. For this paper, we presented the 6 graphs with more interesting phenomena. We investigate the relative bias comparing the crude PAR, par_C , and the semiadjusted PAR, par_{semi} to the true pPAR, par_P . These curves for the relative bias are represented by the solid black and red lines, respectively. We also compare the uncorrected crude PAR, PAR_C , and the uncorrected semiadjusted PAR, PAR_{semi} with the par_P . The curves for the relative bias of these values are represented by the dashed black and red lines respectively. In all the figures, the line $\text{bias}_R = 0$ is given by the green dashed line.

In Figure 5, we observed that the gradient of the relative bias for PAR_C with respect to θ_S is positive, and this was corroborated by the numerical bias evaluation, so that whenever θ_S is increased, the relative bias of PAR_C must either become more positive or less negative. This means that the absolute bias decreases as long as the bias itself is negative, but after the bias becomes positive, the absolute bias begins increasing. The numerical bias evaluation also revealed that more often than not, the gradient of the relative bias for PAR_C with respect to ϕ_S is positive, and we see this manifested in Figure 6 as well, although the magnitude of the gradient is much smaller for ϕ_S than for θ_S . In Figure 7, we see that the relative bias of PAR_{semi} is very small but positive for $\theta_T = 0.75$ and increases in the negative direction as θ_T increases from 0.75. A similar trend was observed for ϕ_T (figure not shown). However, results from the numerical bias evaluation showed that these phenomena were specific to the parameters chosen in this example and are not true in general. Figure 8 suggests that as rr_{3T} increases, par_{semi} decreases but the other 3 estimands increase. This is largely corroborated by the results of the numerical bias evaluation shown in Tables 3 and 4.

When p_S was varied from 0.05 to 0.95, the crude PAR estimands were biased away from the null, and the par_{semi} was slightly biased towards the null. The PAR_{semi} appeared unbiased for most values of p_S . 0.9 and was slightly biased towards the null for $p_S > 0.9$, as is evident from Figure 9. When p_T was varied over the same interval, as seen in Figure 10, the relative bias of the crude PARs and the uncorrected semiadjusted PAR increased initially and later decreased. However, results from the numerical bias evaluation showed that all of these phenomena were specific to the parameters chosen and do not represent general phenomena.

We saw that $\frac{\partial \text{bias}_R}{\partial p_T} = 0$ in 15 to 20% of the scenarios explored, and the remainder split

between the positive and negative directions with neither being significantly more frequent than the other (Table 4).

5 ILLUSTRATIVE EXAMPLE

We demonstrate the use of the pPAR in the HPFS by examining the extent to which high red meat intake, high alcohol intake, and low supplemental folate intake explain the occurrence of colorectal cancer in the study population.²⁴ Following Platz et al,⁶ we defined high red meat intake as more than 2 servings of pork, beef, or lamb as a main dish each week, high alcohol intake as 15 or more grams of alcohol per day, and low supplemental folate intake as less than 100 μg of folic acid per day from supplements. Because the crude and semiadjusted PARs continue to be reported in public health research, eg, the previous studies,^{9,10,25} we also present these quantities, to show the extent of bias when they are incorrectly used when the pPAR is the quantity needed. We investigate how these quantities vary in the presence of nondifferential risk factor misclassification, under the additional assumption that the misclassification of each risk factor is independent of other risk factors.

5.1 Risk factor misclassification

Health Professionals Follow-up Study began in 1986 when 51 529 US male health professionals, aged 40 to 75 years, responded to a questionnaire, which included a semiquantitative Food Frequency Questionnaire (FFQ) with 131 food items plus vitamin and mineral supplement use. The validity of this FFQ was assessed using 2 diet records obtained from a subsample of 127 HPFS participants.²⁶

The data from the FFQ in the main study were used to compute the prevalence ($\hat{p}_Z$) of the surrogate risk factors. The sensitivities ($\hat{\theta}$) and specificities ($\hat{\phi}$) were computed by comparing the responses from the FFQ to those from the dietary records in the validation study. It can be seen in Table 5 that the risk factors were measured with moderate to substantial error. The relative risks ($\hat{RR}$) in Table 5 were obtained from the previously published papers on these risk factors in relation to colorectal cancer incidence in HPFS.^{27,28} Using these values, along with the risk factor prevalence and the sensitivities and specificities, we calculated the corrected marginal prevalence ($\hat{p}_X$) and relative risks ($\hat{r}$) using the well-known algebra of misclassification.²⁹ With the marginal prevalence reported in Table 5, and the reclassification matrices estimated from the validation study, we calculated the corrected joint prevalence from the uncorrected joint prevalence and from there the PAR estimates given in Table 7. Additionally, in Table 6, we estimated the pairwise odds ratios between the risk factors studied, after correcting for misclassification, to give a sense of the magnitude and direction of the pairwise associations between the risk factors. It is evident from these odds ratios of association that the risk factors are positively associated.

5.2 Methods and results

To validly estimate the proportion of colorectal cancer cases that would be prevented by eliminating an exposure from the population while keeping the distributions of 2 background risk factors unchanged, unless the 3 risk factors are each pairwise uncorrelated, the $\widehat{\text{PAR}}_p$ must be used for valid estimation. We report the pPAR values for 3 scenarios where each of the modifiable dietary risk factors is eliminated one at a time while the distributions of the other 2 are left unchanged. We also report the crude and semiadjusted PAR values for these

3 scenarios, as well as the theoretical values of the uncorrected pPAR estimands under misclassification. These are shown in the top part of Table 7. We also report the various PAR values for the 3 scenarios where the modifiable exposures are eliminated 2 at a time while keeping the distribution of the third one unmodified.

When low folate intake is prevented while the distribution of the other 2 factors is unchanged, the uncorrected PAR_P is 21%, overestimating the true par_P which is 14%, consistent with patterns evident in the graphs in Section 4, namely, Figures 2, 3, 4, 7, and 8, all of which illustrated that PAR_P was more likely to be overestimated when the modifiable exposure has high sensitivity and a low relative risk, while the background risk factors are moderately misclassified and have high relative risks, and the risk factors are positively associated. Recall that “overestimate” in this paper refers to the fact that the theoretical values of PAR_P and other estimands are greater than the theoretical value of par_P . It is interesting to observe overestimation in the pPAR, because in the univariate case, when there is only one risk factor, the PAR can never be overestimated.¹³ In all 6 scenarios, the values for par_S are similar to those for par_P or slightly biased towards the null, while the crude par_C is generally biased away from the null. Also, $\widehat{PAR}_{semi}$ and $\widehat{PAR}_C$ grossly underestimate par_P in all but one scenario.

6 DISCUSSION

In this paper, we have extended the work of Hsieh and Walter¹³ to consideration of the pPAR, taking into account the multifactorial disease setting common in chronic disease epidemiology and public health research. We studied the behavior of the pPAR, the semiadjusted PAR, and the crude PAR, as well as their uncorrected estimands. Over a wide range of parameters values used to evaluate the estimands in our numerical bias evaluation, we determined how frequently bias was zero, positive, and negative and how often the magnitude of the bias was low or high. We were also able to determine how often the change in relative bias of an estimand with respect to a change in a given parameter was zero, positive, or negative. Throughout, we have assumed that the outcome is correctly classified while the modifiable exposures and/or the background risk factors may be subject to nondifferential misclassification.

On the basis of the numerical bias evaluation, we saw that the frequency of positive vs negative bias, high vs low bias, was similar across all 3 uncorrected estimands and that the relative bias of the semiadjusted PAR, par_{semi} , fell within the interval for low bias nearly all of the time. Furthermore, we observed that θ_S and π_{ST} have the largest impacts on the uncorrected estimands and hence the direction and magnitude of bias. The relative bias of the uncorrected pPAR, $bias_R = \frac{PAR_P - par_P}{par_P}$, generally increased with θ_S , ϕ_S , and π_{ST} . The change in $bias_R$ with respect to a change in any of the 6 other parameters depended nonmonotonically and nonlinearly on the values of the other parameters, with π_{ST} often having the greatest influence.

There is no bias in the pPAR when there is nearly perfect classification, ie, $\theta_S = \theta_T = \phi_T = 1$, regardless of the value of ϕ_S . This is similar to what has been reported for the univariate

PAR where there is no bias in the PAR in the presence of perfect sensitivity under nondifferential misclassification. The bias is also zero at other sets of parameter values with no particular pattern observable as to when this occurs.

In the HPFS study of risk factors for colorectal cancer considered here, the crude pPARs substantially overestimated the corrected pPAR, while the semiadjusted PARs were slightly biased towards the null when compared with the pPARs. In most cases, the misclassified pPARs appeared to be underestimates of the corrected pPAR values. Hence, the importance of correcting for misclassification to avoid biased estimates should be evident.

Because the uncorrected pPARs were mostly biased towards the null in the HPFS study, it appears that colorectal cancer is substantially more preventable than may be currently believed, with the pPAR being underestimated by as much as 77% (0.14 versus 0.60) when the modifiable exposure is high red meat intake and the background risk factors are high alcohol intake and low folate intake. However, the results of our numerical study and our data analysis have shown that it is in fact possible to overestimate the pPAR as well, so that in some other situations, the effect of a public health intervention in preventing diseases could also be smaller than may be believed.

The PAR is an important tool for the translation of etiologic epidemiologic research to the public health arena. As is evident from the figures and results in this paper, it is possible to dramatically underestimate the pPAR and also to overestimate it, depending on the extent of misclassification, the prevalence of the risk factors, their association with one another, and the relative risks. Except in a few situations, it is not possible to quantify the magnitude and direction of the bias, because the pPAR depends in a complex, nonlinear manner on 9 parameters. It is therefore necessary to correct for misclassification to ensure that we neither underestimate nor overestimate the pPAR. Future work will develop methods to correct point and interval estimates for the pPAR in main study/validation study designs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Grant to acknowledge is DP1 ES025459.

References

1. Levin ML. The occurrence of lung cancer in man. *Acta-Unio Internationalis Contra Cancrum*. 1952; 9(3):531–541.
2. Rothman KJ, Greenland S. *Modern epidemiology*. 1998
3. Mocellin S, Verdi D, Nitti D. DNA repair gene polymorphisms and risk of cutaneous melanoma: a systematic review and meta-analysis. *Carcinogenesis*. 2009; 30(10):1735–1743. [PubMed: 19706646]
4. Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genet*. 2007; 39(7):870–874. [PubMed: 17529973]

5. Cerhan JR, Potter JD, Gilmore JME, et al. Adherence to the AICR cancer prevention recommendations and subsequent morbidity and mortality in the Iowa Women's Health Study cohort. *Cancer Epidemiol Biomarkers Prev.* 2004; 13(7):1114–1120. [PubMed: 15247121]
6. Platz EA, Willett WC, Colditz GA, Rimm EB, Spiegelman D, Giovannucci E. Proportion of colon cancer risk that might be preventable in a cohort of middle-aged US men. *Cancer Causes Control.* 2000; 11(7):579–588. [PubMed: 10977102]
7. Tamimi RM, Spiegelman D, Smith-Warner SA, et al. Population attributable risk of modifiable and nonmodifiable breast cancer risk factors in postmenopausal breast cancer. *Am J Epidemiol.* 2016; 84(12):884–893.
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(5):904–914. [PubMed: 4050778]
9. Cole P, MacMAHON B. Attributable risk percent in case-control studies. *Br J Preventive Social Med.* 1971; 25(4):242–244.
10. Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health.* 1982; 72(12):1336–1344. [PubMed: 7137430]
11. Govindarajulu US, Spiegelman D, Miller KL, Kraft P. Quantifying bias due to allele misclassification in case-control studies of haplotypes. *Genet Epidemiol.* 2006; 30(7):590–601. [PubMed: 16830341]
12. Blair A, Stewart P, Lubin JH, Forastiere F. Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *Am J Ind Med.* 2007; 50(3):199–207. [PubMed: 17096363]
13. Hsieh C-C, Walter SD. The effect of non-differential exposure misclassification on estimates of the attributable and prevented fraction. *Stat Med.* 1988; 7(10):1073–1085. [PubMed: 3206003]
14. Vogel C, Gefeller O, et al. Implications of nondifferential misclassification on estimates of attributable risk. *Methods Inf Med.* 2002; 41(4):342–348. [PubMed: 12425247]
15. Hsieh C-C. The effect of non-differential outcome misclassification on estimates of the attributable and prevented fraction. *Stat Med.* 1991; 10(3):361–373. [PubMed: 2028120]
16. Vogel C, Brenner H, Pfahlberg A, Gefeller O. The effects of joint misclassification of exposure and disease on the attributable risk. *Stat Med.* 2005; 24(12):1881–1896. [PubMed: 15736279]
17. Walter SD, Hsieh CC, Liu Q. Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Stat Med.* 2007; 26(26):4833–4842. [PubMed: 17691081]
18. Hanley JA. A heuristic approach to the formulas for population attributable fraction. *J Epidemiol Community Health.* 2001; 55(7):508–514. [PubMed: 11413183]
19. Spiegelman D, Hertzmark E, Wand HC. Point and interval estimates of partial population attributable risks in cohort studies: examples and software. *Cancer Causes & Control.* 2007; 18(5):571–579. [PubMed: 17387622]
20. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health.* 1998; 88(1):15–19. [PubMed: 9584027]
21. Greenland S, Morgenstern H. Morgenstern corrects a conceptual error. *Am J Public Health.* 1983; 73(6):703–704.
22. Flegal KM, Graubard BI, Williamson DF. Methods of calculating deaths attributable to obesity. *Am J Epidemiol.* 2004; 160(4):331–338. [PubMed: 15286018]
23. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol.* 1980; 112(4):564–569. [PubMed: 7424903]
24. Rimm EB, Giovannucci EL, Willett WC, et al. Prospective study of alcohol consumption and risk of coronary disease in men. *The Lancet.* 1991; 338(8765):464–468.
25. Lee I-M, Shiroma EJ, Lobelo F, et al. Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The Lancet.* 2012; 380(9838):219–229.
26. Feskanich D, Rimm EB, Giovannucci EL, et al. Reproducibility and validity of food intake measurements from a semiquantitative Food Frequency Questionnaire. *J Am Dietetic Assoc.* 1993; 93(7):790–796.

27. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Willett WC. Intake of fat, meat, and fiber in relation to risk of colon cancer in men. *Cancer Res.* 1994; 54(9):2390–2397. [PubMed: 8162586]
28. Giovannucci E, Rimm EB, Ascherio A, Stampfer MJ, Colditz GA, Willett WC. Alcohol, low-methionine-low-folate diets, and risk of colon cancer in men. *J National Cancer Inst.* 1995; 87(4): 265–273.
29. Kleinbaum, DG., Kupper, LL. *Epidemiologic Research: Principles and Quantitative Methods.* John Wiley & Sons; New York: 1982.

FIGURE 1.

Effect of varying ϕ_S , the exposure specificity ($\theta_T = \phi_T = p_{S\cdot} = p_{\cdot T} = 0.75$, $\pi_{1S} = 1.25$, $\pi_{2T} = 3$, $\pi_{ST} = 1.1$) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 2.

Effect of varying ϕ_T , the background specificity ($\phi_S = \theta_T = p_S = p_{\cdot T} = 0.75$, $\pi_{1S} = 1.25$, $\pi_{2T} = 3$, $\pi_{ST} = 1.1$) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 3.

Effect of varying p_S , the exposure prevalence ($\phi_S = \phi_T = \theta_T = p_T = 0.75$, $\pi_{1S} = 1.25$, $\pi_{2T} = 3$, $rr_{ST} = 1.1$) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 4.

Effect of varying $p.T$, the background factor prevalence ($\phi_S = \phi_T = \theta_T = p_S = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{ST} = 1.1$) [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 5.

Effect of varying θ_S , exposure sensitivity, holding $\phi_S = \phi_T = \theta_T = p_S = p_T = 0.75$, $\pi_{1S} = 1.25$, $\pi_{2T} = 3$, $\pi_{ST} = 1.1$

FIGURE 6.

Effect of varying ϕ_S , exposure specificity, holding $\theta_S = \phi_T = \theta_T = p_S = p_T = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{ST} = 1.1$

FIGURE 7.

Effect of varying θ_T , background factor sensitivity, holding $\phi_S = \theta_S = \phi_T = p_S = p_{\cdot T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{ST} = 1.1$

FIGURE 8.

Effect of varying rr_{SIT} , association of X_S with X_T , holding $\phi_S = \theta_S = \phi_T = \theta_T = p_S = p_T = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$

FIGURE 9.

Effect of varying p_S , holding $\phi_S = \theta_S = \phi_T = \theta_T = p_{\cdot T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

FIGURE 10.

Effect of varying p_T , holding $\phi_S = \theta_S = \phi_T = \theta_T = p_S = 0.75$, $r_{1S} = 1.25$, $r_{2T} = 3$, $r_{ST} = 1.1$

Percentage of cases where $\text{bias}_R = -10\%$, $-10\% < \text{bias}_R < 0$, $\text{bias}_R = 0$, $0 < \text{bias}_R < 10\%$, and $\text{bias}_R \geq 10\%$, over a numerical bias evaluation in the 9-dimensional parameter space

TABLE 1

	bias_R	-10%	$-10\% < \text{bias}_R < 0$	$\text{bias}_R = 0$	$0 < \text{bias}_R < 10\%$	$\text{bias}_R \geq 10\%$
PAR_P	77	7	7	7	2	2
par_{semi}	0	40	22	37	1	1
PAR_{semi}	77	8	5	8	2	2
par_c	12	25	22	28	13	13
PAR_c	77	7	6	7	3	3

TABLE 2

Direction of the change in bias of PAR_P with respect to each of the parameters studied, over a numerical bias evaluation in the 9-dimensional parameter space

Parameter (ψ)	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} < 0$	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} = 0$	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} > 0$
θ_S	0.04%	0%	99.96%
ϕ_S	7%	11%	82%
θ_T	46%	11%	43%
ϕ_T	46%	12%	42%
p_S	94%	2%	4%
$p_{\cdot T}$	45%	10%	45%
π_{ST}	11%	0%	89%
π_{1S}	44%	11%	45%
π_{2T}	44%	11%	45%

TABLE 3

Direction of the change in the bias of the semi-adjusted PAR estimands with respect to each of the parameters studied, over a numerical bias evaluation in the 9-dimensional parameter space

Parameter (ψ)	Estimand	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} < 0$ Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} = 0$ Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} > 0$		
θ_S	par_{semi}	0%	100%	0%
	PAR_{semi}	0.04%	0%	99.96%
ϕ_S	par_{semi}	0%	100%	0%
	PAR_{semi}	10%	5%	85%
θ_T	par_{semi}	0%	100%	0%
	PAR_{semi}	43%	18%	39%
ϕ_T	par_{semi}	0%	100%	0%
	PAR_{semi}	43%	18%	39%
p_S	par_{semi}	38%	19%	43%
	PAR_{semi}	84%	5%	11%
p_T	par_{semi}	35%	20%	45%
	PAR_{semi}	44%	16%	40%
$\pi_{S T}$	par_{semi}	100%	0%	0%
	PAR_{semi}	19%	0%	81%
π_{1S}	par_{semi}	35%	17%	48%
	PAR_{semi}	53%	4%	43%
π_{2T}	par_{semi}	42%	20%	38%
	PAR_{semi}	41%	16%	43%

TABLE 4

Direction of the change in bias of the crude PAR estimands with respect to each of the parameters studied, over a numerical bias evaluation in the 9-dimensional parameter space

Parameter (ψ)	Estimand	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} < 0$	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} = 0$	Frequency of $\frac{\partial \text{bias}_R}{\partial \psi} > 0$
θ_S	par_c	0%	100%	0%
	PAR_c	0%	0%	100%
ϕ_S	par_c	0%	100%	0%
	PAR_c	3%	19%	78%
θ_T	par_c	0%	100%	0%
	PAR_c	10%	80%	10%
ϕ_T	par_c	0%	100%	0%
	PAR_c	10%	80%	10%
p_S	par_c	38%	19%	43%
	PAR_c	85%	4%	11%
p_T	par_c	45%	20%	35%
	PAR_c	47%	15%	38%
π_{ST}	par_c	0%	0%	100%
	PAR_c	0%	0%	100%
π_{1S}	par_c	44%	17%	39%
	PAR_c	54%	3%	43%
π_{2T}	par_c	39%	19%	42%
	PAR_c	41%	15%	44%

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 5

Parameter estimates for calculating population attributable risks for colorectal cancer in the Health Professionals Follow-up Study

Modifiable exposure	$\hat{\theta}$	$\hat{\phi}$	$\hat{p}_Z$	$\hat{p}_X$	$\widehat{RR}$	$\hat{r}^*$
Red meat (main dish)	0.53	0.78	0.39	0.55	1.40	3.6
Alcohol	0.76	0.93	0.65	0.85	1.42	3.9
Low folate	0.97	0.68	0.75	0.66	1.18	1.24

TABLE 6

Pairwise odds ratios between risk factors in the Health Professionals Follow-up Study, after correcting for misclassification

Risk factor 1	Risk factor 2	$\hat{or}$
Low folate	Alcohol	2.1
Low folate	Red meat (main dish)	2.2
Red meat (main dish)	Alcohol	4.8

Population attributable risks of colorectal cancer due to alcohol, red meat, and folate intake, Health Professionals Follow-up Study

TABLE 7

Modifiable exposure	Background risk factor(s)				
	$\widehat{PAR}_P$	$\widehat{par}_P^{\#}$	$\widehat{PAR}_S$	$\widehat{par}_S$	$\widehat{par}_C$
Low folate					
Alcohol, red meat	0.21	0.14	0.24	0.14	0.22
Alcohol					
Low folate, red meat	0.24	0.73	0.24	0.71	0.25
Red meat					
Low folate, alcohol	0.14	0.60	0.14	0.59	0.15
Red meat					
Low folate, alcohol	0.38	0.76	0.43	0.75	0.40
Alcohol, red meat					
Low folate	0.38	0.89	0.42	0.89	0.38
Low folate, red meat					
Alcohol	0.34	0.65	0.35	0.65	0.34
All 3					
	0.50 [*]	0.90 [*]	NA	NA	NA

^{*} Denotes the estimates for the full $\widehat{PAR}_S$, $\widehat{PAR}_F$ and $\widehat{par}_F$.

[#] Indicates the valid pPAR that should be used in the presence of confounding.