Misclassification in the Partial Population Attributable Risk

Misclassification in the Partial Population Attributable Risk

A dissertation presented

by

Hong Wen Benedict Wong

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

December 2017

Dissertation Advisor: Professor Donna Spiegelman Hong Wen Benedict Wong

Misclassification in the Partial Population Attributable Risk

Abstract

The population attributable risk (PAR) is often defined as the percent reduction in disease incidence that would be observed if the exposure were to be entirely removed from the population, given the exposure distribution in the population to which the results are intended to apply. The partial population attributable risk is used to quantify the population-level impact of preventive interventions in a multi-factorial disease setting.

In this dissertation, we considered the effect of non-differential risk factor misclassification on the direction and magnitude of bias of partial population attributable risk (pPAR) estimands and related quantities in Chapter 1. We also developed methods for estimation and inference in Chapters 2 and 3, using both a likelihood-based approach and a Bayesian approach to correct this bias, under different study designs. We evaluated the performance of the methods through extensive simulation studies, and applied them to the Health Professionals Follow-Up Study for risk factors of colorectal cancer.

Contents

Title page i
Abstract iii
Table of Contents iv

Contents **iv**

Acknowledgments vii

1 The Effect of Risk Factor Misclassification in the Partial Population Attributable Risk **1**

1.1 Introduction 3
1.2 Notation 5
1.2.1 The full PAR 6
1.2.2 The partial PAR and crude PAR 6
1.2.3 The semi-adjusted PAR 9
1.3 The effect of misclassification on PAR estimands 10
1.3.1 Theoretical findings 11
1.3.2 Numerical results 11
1.3.3 Graphical exploration 15
1.4 Bias in the semi-adjusted and crude PAR estimands 18
1.5 Illustrative Example 21
1.5.1 Risk Factor Misclassification 21
1.5.2 Methods and Results 22
1.6 Discussion 24

2 Correction for Risk Factor Misclassification in the Partial Population Attributable Risk **31**

2.1 Introduction . . . 32

2.2 Notation . . . 34

2.2.1 The Disease and Misclassification Models . . . 36

2.2.2 The Likelihood . . . 37

2.2.3 Interval estimates . . . 39

2.3 Simulation Study . . . 39

2.4 Application . . . 42

2.5 Discussion . . . 45

3 A Bayesian Approach to Correcting for Risk Factor Misclassification **47**

3.1 Introduction . . . 48

3.2 Notation . . . 49

3.2.1 The partial PAR . . . 50

3.2.2 The model, prior and posterior probability distributions . . . 51

3.2.3 The posterior distribution for the MS/IVS design . . . 53

3.2.4 The first stage . . . 55

3.2.5 The second stage . . . 56

3.2.6 Sampling from the posterior distribution when the validation study is external . . . 57

3.3 Application . . . 58

3.4 Discussion . . . 60

4 Summary **61**

References **65**

A Supplementary Materials for Chapter 1 **65**

A.1 Derivation of the cell probabilities: . . . 65

A.2 An expression for the crude relative risk when 2 binary risk factors . . . 67

A.3 Expression for the misclassified model 67
A.4 Proving Independence of bias from p_d 68
A.5 Proof of Zero Bias when $\theta_S = \theta_T = \phi_T = 1$: 69
A.6 Proof that the pPAR in the two-factor bivariate case reduces to the univariate case under the assumption of independence between risk factors and multiplicity of the relative risks: 70

B Published Manuscript for Chapter 1 **71**

Acknowledgments

I am extremely grateful to my advisor, Donna Spiegelman, for her guidance and encouragement through this journey. I would like to thank my dissertation committee members, Raymond Carroll, Molin Wang, and Paige Williams, for taking time off their busy schedules to provide feedback, comments and insights at my committee meetings. I thank my fellow students for their friendship and support, as well as other members of the department who have helped me over the years. This dissertation is dedicated to my family.

The Effect of Risk Factor Misclassification in the Partial Population Attributable Risk

Benedict Wong, Sarah Peskoe and Donna Spiegelman

Department of Biostatistics

Harvard T.H. Chan School of Public Health

This chapter has been accepted for publication as a journal paper in *Statistics in Medicine*. The paper, as it appears in journal form, is included as Appendix C. The paper is also available online at https://doi.org/10.1002/sim.7559

Chapter 1 Abstract

In this chapter, we consider the effect of non-differential risk factor misclassification on the direction and magnitude of bias of partial population attributable risk (pPAR) estimands and related quantities. The partial population attributable risk is used to quantify the population-level impact of preventive interventions in a multi-factorial disease setting. We found that the bias in the uncorrected pPAR depends non-linearly and non-monotonically on the sensitivities, specificities, relative risks and joint prevalences 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 semi-adjusted PAR is calculated using the formula for the crude PAR but plugs in the multivariate-adjusted relative risk. Because the crude and semi-adjusted 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 due to 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.

1.1 Introduction

The population attributable risk (PAR) is often defined as the percent reduction in disease incidence that would be observed if the exposure were 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 (Levin, 1952).

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, population attributable risk proportion and population attributable risk percent (Rothman and Greenland, 1998).

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 al. (Mocellin et al., 2009) found that the PAR% for melanoma of the variant C allele of the XPD/ERCC2 single nucleotide polymorphism (SNP) was 9.6%, where the PAR% = PAR $\times 100\%$. Similarly, Hunter et al. (Hunter et al., 2007) found that 4 SNPs 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 Study (Cerhan et al., 2004) 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 al. 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%) (Platz et al., 2000). More recently, in Har-

vard's Nurses' Health Study, Tamimi et al. 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 (Tamimi et al., 2016).

In an evaluation of a preventive intervention in a multi-factorial disease setting, interest is in the percentage of cases expected to be reduced with the exposures eliminated, when other risk factors, possibly non-modifiable, 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 (Bruzzi et al., 1985). 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 (Cole and MacMahon, 1971; Morgenstern, 1982). We refer to this as the semi-adjusted PAR.

Although it is widely acknowledged that the misclassification rates for genetic assays are quite low (Govindarajulu et al., 2006), 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 to 0.6 (Blair et al., 2007). 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 non-differential 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 (Hsieh and Walter, 1988). Bias due to exposure misclassification in the univariate risk factor setting has been

studied widely (Vogel et al., 2002; Hsieh and Walter, 1988; Hsieh, 1991; Vogel et al., 2005; Walter et al., 2007), but not in the multi-factorial 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 non-modified 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 prevalences of the risk factors. Because the crude and semi-adjusted 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 1.2 even with nondifferential misclassification. In Sections 1.3 and 1.4, we report the results of numerical studies of the bias over the multi-dimensional 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 1.5, in a study of the pPAR of risk factors for colorectal cancer in the Health Professionals Follow-Up Study (Platz et al., 2000). Section 1.6 concludes this paper with a summary and recommendations for future research.

1.2 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.

1.2.1 The full PAR

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

$$par_F = 1 - \frac{1}{\sum_{s=0}^{S} p_s rr_s}, \tag{1.1}$$

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 $s = 0, ..., S$ (Hanley, 2001). 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.

1.2.2 The partial PAR and crude PAR

In the presence of confounding due to background risk factors that are possibly non-modifiable 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 non-modified background risk factors. Then, under the assumption of no interaction effects between the modifiable exposures and the non-modified background risk factors, the partial PAR (pPAR) is

$$par_P = 1 - \frac{\sum_{t=0}^{T} p_{.t} rr_{2t}}{\sum_{s=0}^{S} \sum_{t=0}^{T} p_{st} rr_{1s} rr_{2t}}, \tag{1.2}$$

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 non-

modified 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, $s = 0, ..., S$ and rr_{2t} is the relative risk in stratum t relative to the lowest risk stratum, $t = 0, ..., T$. Note that $rr_{10} = rr_{20} = 1$ and $p_{.t} = \sum_{s=0}^{S} p_{st}$ for all t (Spiegelman et al., 2007). 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 non-modified 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 'non-modified' background risk factors.

In the scenario where there is only one modifiable exposure with two risk levels, and one non-modified background risk factor with two levels, so that $S = T = 1$, a derivation for the joint prevalences $\{p_{st}$ for $(s, t) \in \{0, 1\}^2\}$ is given in the Appendix under Section A.1. Using the formula for par_F in Equation (1.1) 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 (Rockhill et al., 1998). Here, the crude PAR is

$$par_c = 1 - \frac{1}{\sum_{s=0}^{S} p_{s.} rr_{1s}^{(c)}}, \tag{1.3}$$

where the $rr_{1s}^{(c)}$, $s = 0, ..., S$ are the crude relative risks, unadjusted for the background non-modified 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 non-modified background risk factor, with both of them being binary such that $S = T = 1$, Equations (1.2) and (1.3) reduce respectively to

$$par_P = 1 - \frac{\sum_{t=0}^{1} p_{.t} rr_{2t}}{\sum_{s=0}^{1} \sum_{t=0}^{1} p_{st} rr_{1s} rr_{2t}} = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11}) rr_{2T}}{p_{00} + p_{10} rr_{1S} + p_{01} rr_{2T} + p_{11} rr_{1S} rr_{2T}}, \tag{1.4}$$

and

$$par_c = 1 - \frac{1}{\sum_{s=0}^{1} p_{s.} rr_{1s}^{(c)}} = 1 - \frac{1}{(p_{00} + p_{01}) + (p_{10} + p_{11}) rr_{1S}^{(c)}}, \tag{1.5}$$

where rr_{1S} and rr_{2T} are the relative risks for the disease given the modifiable exposure and non-modified background risk factor respectively, p_{st} is the proportion of the population with level s of the modifiable exposure and level t of the non-modified 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 non-modified background risk factor. In Section A.2 of the Appendix, we provide an expression for $rr_{1S}^{(c)}$ as a function of the marginal prevalences $p_{S.}$ and $p_{.T}$, the relative risks rr_{1S} and rr_{2T}, and the relative risk of having the modifiable exposure given the non-modified background risk factor.

Without loss of generality, we can drop the subscripts S and T from the relative risk estimands in Equation (1.4) when the modifiable exposure and the non-modified background risk factor are both binary, and rewrite the equation as

$$par_P = 1 - \frac{p_{.0} + p_{.1}rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2} = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2}, \quad (1.6)$$

In Equation (1.6), we assumed that there was no interaction effect between the modifiable exposure and the non-modified background risk factor. However, in practice, this assumption may not always hold. Hence, the above equation can be written more generally as

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_3}, \quad (1.7)$$

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 non-modified background risk factor, the uncorrected pPAR is given by

$$PAR_P = 1 - \frac{(P_{00} + P_{10}) + (P_{01} + P_{11})RR_2}{P_{00} + P_{10}RR_1 + P_{01}RR_2 + P_{11}RR_3}, \quad (1.8)$$

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 (1.7) and (1.8) give the expressions for two 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 non-modified 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 non-differential 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.

1.2.3 The semi-adjusted PAR

In the context of multi-factorial 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, e.g. (Cole and MacMahon, 1971; Morgenstern, 1982). We refer to this as the semi-adjusted PAR:

$$par_{semi} = 1 - \frac{1}{(1 - p_{1.}) + p_{1.} rr_1^{(a)}} \tag{1.9}$$

where $rr_1^{(a)}$ is the multivariate-adjusted relative risk of the modifiable exposure. If the relative risks of the modifiable exposure and the non-modified background risk factor are multiplicative as in Equation (1.6) in the absence of an interaction effect, then $rr_1^{(a)} = rr_1$. In practice, this may not always hold true.

In Equation (1.9), the crude relative risk of Equation (1.3) 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 (Greenland and Morgenstern, 1983), with the direction of the bias going both ways and the magnitude of the bias varying according to the parameters of the model (Flegal et al., 2004).

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

$$PAR_{semi} = 1 - \frac{1}{(1 - P_{1.}) + P_{1.}RR_1^{(a)}} \tag{1.10}$$

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

1.3 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 non-modified 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 $\theta_S, \phi_S, \theta_T, \phi_T, p_{S.}, p_{.T}, rr_{1S}, rr_{2T}, rr_{S|T}, p_d$, where p_d is the prevalence or cumulative incidence of the disease and $rr_{S|T}$ is the relative risk of having the modifiable exposure given the non-modified background risk factor, defined by $rr_{S|T} = \frac{P(X_S=1|X_T=1)}{P(X_S=1|X_T=0)}$ with $rr_{S|T} > 1$ indicating a positive association between the modifiable exposure and the background risk factor. While p_d has no effect on the bias (see Section 1.3.1), the other 9 parameters have non-linear non-monotonic effects on the magnitude and direction of the bias. In Sections 1.3.1-1.3.2, we investigate general trends surrounding the behavior of the bias, and in Section 1.3.3, 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 in-

teraction 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 (Greenland, 1980) and we allow for this phenomenon when deriving the expression for the uncorrected pPAR, PAR_P. In Equation A.2 in the Appendix, Section A.3, 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 non-linear.

1.3.1 Theoretical findings

Although p_d is one of the 10 parameters of the pPAR, we show in Section A.4 of the Appendix 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 $\theta_S = rr_{S|T} = 1$. These conditions are further described in this section.

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 $\theta_S, \phi_S, \theta_T, \phi_T$ 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}, rr_{1S}, rr_{2T} took values in {1.25, 2.5, 5}, and $rr_{S|T}$ {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 comprised 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 the risk factors. The values for the marginal prevalences 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 $bias_A = PAR_P - par_P$, and $bias_R = \frac{PAR_P - par_P}{par_P}$ respectively. Please note that because we were able to derived 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.

1.3.2.1 When does $PAR_P = par_P$?

We found that there is no bias when $\theta_S = 1$, and when either $\theta_T = \phi_T = 1$ or $rr_{S|T} = 1$ 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 1.1, we see that the curves for $bias_R$ cross the line for no bias at three distinct values of θ_S for the three different values of ϕ_S shown. The conditions for no bias depend on the values of the other parameters, which are in this Figure are set to θ_T = ϕ_T = $p_{S.}$ = $p_{.T}$ = 0.75, rr_{1S} = 1.25, rr_{2T} = 3, $rr_{S|T}$ = 1.1. These default parameters were chosen to be similar to the parameters observed in the Health Professionals Follow-up Study of risk factors for colorectal cancer disucssed in Section 5 (Rimm et al., 1991).

When $\theta_S = \theta_T = \phi_T = 1$, we provide a proof in Section A.5 of the Appendix which shows that the bias is exactly 0.

We also showed that when $rr_{S|T} = 1$, that is, when the two risk factors are independent and the relative risks are multiplicative, the bivariate pPAR reduces to the univariate PAR,

as shown in Section A.6 of the Appendix. Using Maple, we computed the expression for the bias and verified that when $\theta_S = rr_{S|T} = 1$, the bias is exactly 0. These results on the relative risk scale are in agreement with previously published findings for the full PAR (Hsieh and Walter, 1988).

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 was less than 10%. There were 56566 such occurrences among the 266335 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% to 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, $rr_{S|T}$. 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%).

1.3.2.2 When is $PAR_P > par_P$?

In the range of the parameter space explored, positive bias only occured when $rr_{S|T} > 1$, 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 two 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 $rr_{S|T} > 1$ 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

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

Table 1.1: Percentage of cases where $bias_R \leq -10\%$, $-10\% < bias_R < 0$, $bias_R = 0$, $0 < bias_R < 10\%$, and $bias_R \geq 10\%$, over a numerical bias evaluation in the 9-dimensional parameter space

these two conditions were met.

By dichotomizing $bias_R$ into 2 categories, positive vs. non-positive, and performing a logistic regression of the bias on the 9 parameters among the 266335 scenarios evaluated, we observed that the two most important parameters leading to positive bias were θ_S and $rr_{S|T}$, with coefficients 15 to 7 times greater than the next two 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 $rr_{S|T}$, and also with lower values of θ_T and ϕ_T.

We present the results of the numerical bias evaluation in Table 1.1, 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%).

1.3.2.3 General observations

Over the range of the 9 background parameters explored, it was apparent from Table 1.1 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

Table 1.2: 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* $\frac{\partial bias_R}{\partial \psi} < 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} = 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} > 0$
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%
$p_{.T}$	45%	10%	45%
$rr_{S\|T}$	11%	0	89%
rr_{1S}	44%	11%	45%
rr_{2T}	44%	11%	45%

word 'overestimate' is used when the theoretical values of the estimands are greater than the theoretical values of par_P.

In Table 1.2, $\frac{\partial bias_R}{\partial \phi_S}$ was calculated by taking the difference between $bias_R$ values, obtained from two adjacent points with different ϕ_S but identical values for all other parameters, divided by the difference of the two ϕ_S values, to obtain a finite difference approximation to the partial derivatives. From Table 1.2, it is apparent that the bias of the PAR_P depends non-linearly 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.

1.3.3 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 1.3.2 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 Health Professionals Follow-up Study (HPFS) of risk factors for colorectal cancer disucssed in Section 1.5 (Rimm et al., 1991). They are: $p_{S.} = p_{.T} = \phi_S = \theta_T = \phi_T = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3.0$, $rr_{S|T} = 1.1$.

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.

1.3.2.4 Effect of the misclassification parameters on PAR_P

In Figure 1.1, 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 $\theta_S \leq 0.95$, 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 1.2 corroborated the first observation in that $\frac{\partial bias_R}{\partial \phi_S}$, 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 $\frac{\partial bias_R}{\partial \phi_S}$ 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 1.2, the only exceptions being those cases with $rr_{1S} = 1.25$, $rr_{2T} = 5$, and $rr_{S|T} = 0.7$.

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 1.2 are consistent with data from the most comprehensive results from the numerical bias evaluation for $rr_{S|T} > 1$, 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 1.1 and 1.2, 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 $\frac{\partial bias_R}{\partial \theta_S} > 0$ more than 99% of the time. The cases where $\frac{\partial bias_R}{\partial \theta_S} < 0$ happened at low values of $rr_{S|T}$ 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.

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

In Figure 1.3, 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 non-linear. 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 $\frac{\partial bias_R}{\partial p_{S.}} > 0$ in 4% of the cases, $\frac{\partial bias_R}{\partial p_{S.}} = 0$ 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 1.4 were very much alike and were only distinguishable from one another as $\theta_S \to 1$. 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 $\frac{\partial bias_R}{\partial p_{.T}} = 0$ in 10% of the cases, and in half of the remaining cases, $\frac{\partial bias_R}{\partial p_{.T}}$ was positive.

When $rr_{S|T} = 1$ and $\theta_T = \phi_T = 1$, $\frac{\partial bias_R}{\partial rr_{1S}} = 0$. Otherwise, $\frac{\partial bias_R}{\partial rr_{1S}} > 0$ when $rr_{S|T} < 1$ and $\frac{\partial bias_R}{\partial rr_{1S}} < 0$ when $rr_{S|T} > 1$.

Turning our attention to rr_{2T}, we found that $\frac{\partial bias_R}{\partial rr_{2T}} = 0$ when either $rr_{S|T} = \theta_S = 1$ or $\theta_S = \theta_T = \phi_T = 1$. When neither of the above two conditions hold, and when $rr_{S|T} < 1$, then $\frac{\partial bias_R}{\partial rr_{2T}} < 0$ in a majority of cases studied. Likewise, when $rr_{S|T} > 1$, then $\frac{\partial bias_R}{\partial rr_{2T}} > 0$ in a

majority of the cases studied. Overall, when data from our numerical bias evaluation was aggregated, the frequency for which $\frac{\partial bias_R}{\partial rr_{2T}} > 0$ was approximately equal to the frequency for which $\frac{\partial bias_R}{\partial rr_{2T}} < 0$.

1.4 Bias in the semi-adjusted 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 1.1. 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 1.4, it is apparent that the bias in par_{semi} changed non-linearly and non-monotonically with respect to 4 parameters upon which it depends, p_S, p_T, rr_{1S}, and rr_{2T}. In all cases considered in the numerical bias evaluation, the bias in par_{semi} decreased as $rr_{S|T}$ 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 non-linearly and non-monotonically with respect to the other 7 parameters. Bias in par_C was greater than that for par_P and par_{semi} (Table 1.1), 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 $rr_{S|T}$ (Table 1.4). The bias in par_c varied non-linearly and non-monotonically with p_S, p_T, rr_{1S}, and rr_{2T}. The bias in PAR_C varied similarly with respect to these parameters, and non-monotonically and non-linearly with respect to ϕ_S, θ_T, and ϕ_T, but $\frac{\partial bias_R}{\partial \theta_S}$ 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 semi-adjusted estimators graphically. As in Section 1.3, the default values of the parameters, when held constant, were chosen to be similar to the parameters observed in the Health Professionals Follow-up Study of risk factors for colorectal cancer (Rimm et al., 1991). They were: $p_{S\cdot} = p_{\cdot T} = \phi_S = \theta_T = \phi_T = 0.75$, $\theta_S = 0.97$, $rr_{1S} = 1.25$, $rr_{2T} = 3.0$, $rr_{S|T} = 1.1$. In each graph, we

Table 1.3: 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* $\frac{\partial bias_R}{\partial \psi} < 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} = 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} > 0$
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%
$rr_{S\|T}$	par_{semi}	100%	0%	0%
	PAR_{semi}	19%	0%	81%
rr_{1S}	par_{semi}	35%	17%	48%
	PAR_{semi}	53%	4%	43%
rr_{2T}	par_{semi}	42%	20%	38%
	PAR_{semi}	41%	16%	43%

Table 1.4: 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* $\frac{\partial bias_R}{\partial \psi} < 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} = 0$	*Frequency of* $\frac{\partial bias_R}{\partial \psi} > 0$
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%
$rr_{S\|T}$	par_c	0%	0%	100%
	PAR_c	0%	0%	100%
rr_{1S}	par_c	44%	17%	39%
	PAR_c	54%	3%	43%
rr_{2T}	par_c	39%	19%	42%
	PAR_c	41%	15%	44%

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 six graphs with more interesting phenomena. We investigate the relative bias comparing the crude PAR, par_c, and the semi-adjusted 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 semi-adjusted PAR, PAR_{semi}, to 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 $bias_R = 0$ is given by the green dashed line.

In Figure 1.5, 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 1.6 as well, although the magnitude of the gradient is much smaller for ϕ_S than for θ_S. In Figure 1.7, we see that the relative bias of PAR_{semi} is very small but positive for $\theta_T \leq 0.75$, 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 1.8 suggests that as $rr_{S|T}$ increases, par_{semi} decreases but the other 3 estimands increase. This is largely corroborated by the results of the numerical bias evaluation shown in Table 1.3.

When $p_{S\cdot}$ 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 $p_{S\cdot} \leq 0.9$ and was slightly biased towards the null for $p_{S\cdot} > 0.9$, as is evident from Figure 1.9. When $p_{\cdot T}$ was varied over the same interval, as seen in Figure 1.10, the relative bias of the crude PARs and the uncorrected semi-adjusted 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 $\frac{\partial bias_R}{\partial p_{.T}} = 0$ 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.

1.5 Illustrative Example

We demonstrate the use of the pPAR in the Health Professionals Follow-up Study (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 (Rimm et al., 1991). Following (Platz et al., 2000), 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 semi-adjusted PARs continue to be reported in public health research, e.g. (Cole and MacMahon, 1971; Morgenstern, 1982; Lee et al., 2012), 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 non-differential risk factor misclassification, under the additional assumption that the misclassification of each risk factor is independent of other risk factors.

1.5.1 Risk Factor Misclassification

HPFS began in 1986 when 51,529 U.S. male health professionals, aged 40-75, 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 two diet records obtained from a sub-sample of 127 HPFS participants (Feskanich et al., 1993).

The data from the FFQ in the main study was used to compute the prevalences $(\hat{p}_Z)$ 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 1.5 that the risk factors were measured with moderate to

Table 1.5: Parameter estimates for calculating PARs for colorectal cancer in the Health Professionals Follow-up Study

Modifiable Exposure	$\hat{\theta}$	$\hat{\phi}$	$\hat{p}_Z$	$\hat{p}_X$	$\widehat{RR}$	$\widehat{rr}$
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 1.6: Pairwise odds ratios between risk factors in the Health Professionals Follow-up Study, after correcting for misclassification

Risk Factor 1	*Risk Factor 2*	$\widehat{or}$
Low Folate	Alcohol	2.1
Low Folate	Red Meat (main dish)	2.2
Red Meat (main dish)	Alcohol	4.8

substantial error.

The relative risks $(\widehat{RR})$ in Table 1.5 were obtained from the previously published papers on these risk factors in relation to colorectal cancer incidence in HPFS (Giovannucci et al., 1994, 1995). Using these values, along with the risk factor prevalences and the sensitivities and specificities, we calculated the corrected marginal prevalences $(\hat{p}_X)$ and relative risks $(\widehat{rr})$ using the well-known algebra of misclassification (Kleinbaum and Kupper, 1982). With the marginal prevalences reported in Table 1.5, and the reclassification matrices estimated from the validation study, we calculated the corrected joint prevalences from the uncorrected joint prevalences, and from there the PAR estimates given in Table 1.7. Additionally, in Table 1.6, 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.

1.5.2 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 two background risk factors unchanged, unless the 3 risk factors are each pairwise uncorrelated, the $\widehat{PAR_P}$ must be used for valid estimation. We report the pPAR values for three

Table 1.7: Population attributable risks of colorectal cancer due to alcohol, red meat and folate intake, Health Professionals Follow-up Study(HPFS). Note: * denotes the estimates for the full PARs, $\widehat{PAR}_F$ and $\widehat{par}_F$, $^{\#}$ indicates the valid pPAR that should be used in the presence of confounding

Modifiable Exposure	*Background Risk Factor(s)*	$\widehat{PAR}_P$	$\widehat{par}_P^{\#}$	$\widehat{PAR}_s$	$\widehat{par}_s$	$\widehat{PAR}_c$	$\widehat{par}_c$
Low Folate	Alcohol, Red Meat	0.21	0.14	0.24	0.14	0.22	0.25
Alcohol	Low Folate, Red Meat	0.24	0.73	0.24	0.71	0.25	0.81
Red Meat	Low Folate, Alcohol	0.14	0.60	0.14	0.59	0.15	0.65
Low Folate, Alcohol	Red Meat	0.38	0.76	0.43	0.75	0.40	0.87
Alcohol, Red Meat	Low Folate	0.38	0.89	0.42	0.89	0.38	0.89
Low Folate, Red Meat	Alcohol	0.34	0.65	0.35	0.65	0.34	0.68
All Three		0.50^*	0.90^*	NA	NA	NA	NA

scenarios where each of the modifiable dietary risk factors is eliminated one at a time while the distributions of the other two are left unchanged. We also report the crude and semi-adjusted PAR values for these three scenarios, as well as the theoretical values of the uncorrected pPAR estimands under misclassification. These are shown in the top part of Table 1.7. We also report the various PAR values for the three scenarios where the modifiable exposures are eliminated two at a time while keeping the distribution of the third one unmodified.

When low folate intake is prevented while the distribution of the other two 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 1.4, namely Figures 1.2, 1.3, 1.4, 1.7, and 1.8, 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 (Hsieh and Walter, 1988). 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.

1.6 Discussion

In this paper, we have extended the work of Hsieh and Walter (Hsieh and Walter, 1988) to consideration of the pPAR, taking into account the multifactorial disease setting common in chronic disease epidemiology and public health research. We studied the behaviour of the pPAR, the semi-adjusted PAR, and the crude PAR, as well as their uncorrected estimands. Over a wide range of parameteres 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 non-differential misclassification.

Based on 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 semi-adjusted PAR, par_{semi}, fell within the interval for low bias nearly all of the time. Furthermore, we observed that θ_S and $rr_{S|T}$ 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 $rr_{S|T}$. The change in $bias_R$ with respect to a change in any of the six other parameters depended non-monotonically and non-linearly on the values of the other parameters, with $rr_{S|T}$ often having the greatest influence.

There is no bias in the pPAR when there is nearly perfect classification, i.e. $\theta_S = \theta_T = \phi_T = 1$, 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 non-differential 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 semi-adjusted PARs were slightly biased towards the null when compared with the pPARs. In most cases, the mis-

classified 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 prevalences 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 magntitude and direction of the bias, because the pPAR depends in a complex, non-linear manner on 9 parameters. It is therefore necessary to correct for misclassification in order to ensure that we neither underestimate nor overestimate the pPAR. In the subsequent sections, we developed methods to correct point and interval estimates for the pPAR in various main study/validation study designs.

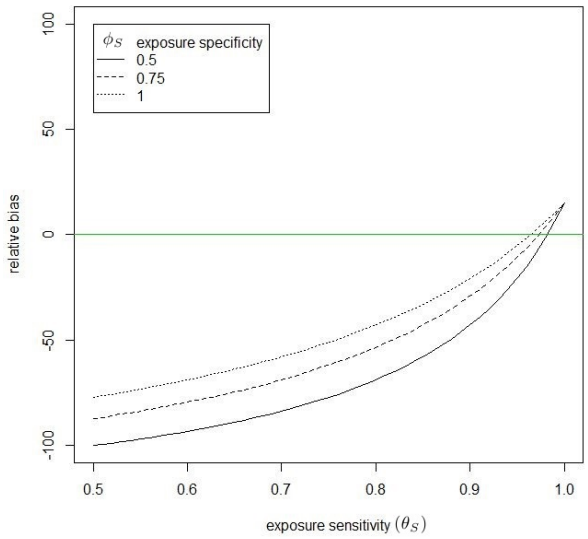

Figure 1.1: Effect of varying ϕ_S, the exposure specificity
($\theta_T = \phi_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$)

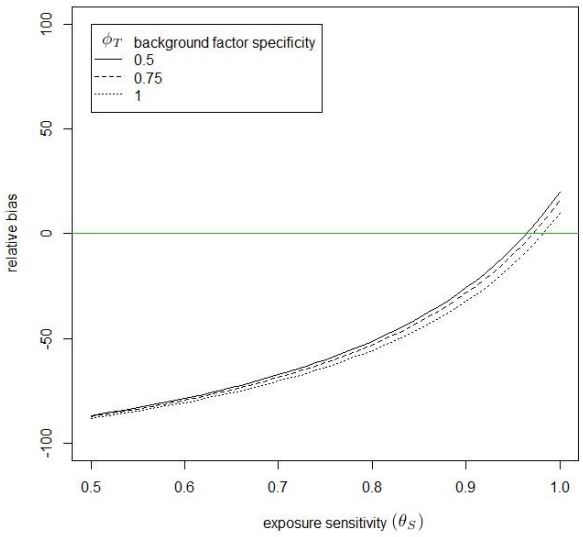

Figure 1.2: Effect of varying ϕ_T, the background specificity
($\phi_S = \theta_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$)

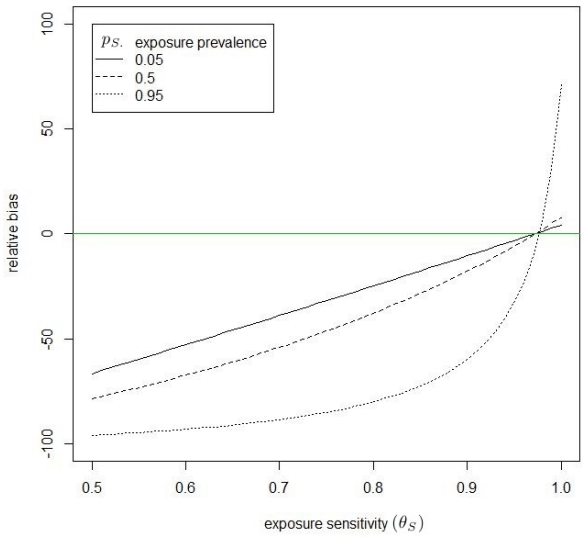

Figure 1.3: Effect of varying $p_{S\cdot}$ the exposure prevalence
($\phi_S = \phi_T = \theta_T = p_{\cdot T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$)

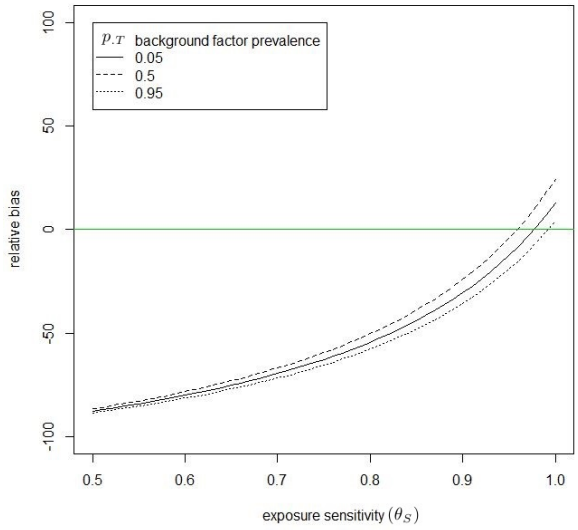

Figure 1.4: Effect of varying $p_{\cdot T}$, the background factor prevalence
($\phi_S = \phi_T = \theta_T = p_{S\cdot} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$)

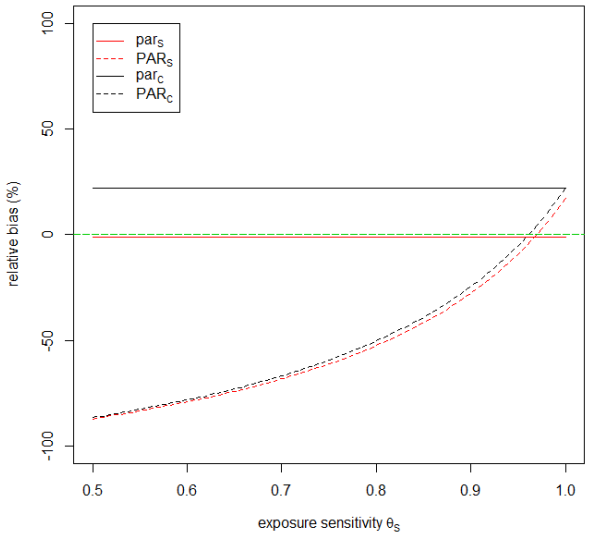

Figure 1.5: Effect of varying θ_S, exposure sensitivity, holding $\phi_S = \phi_T = \theta_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

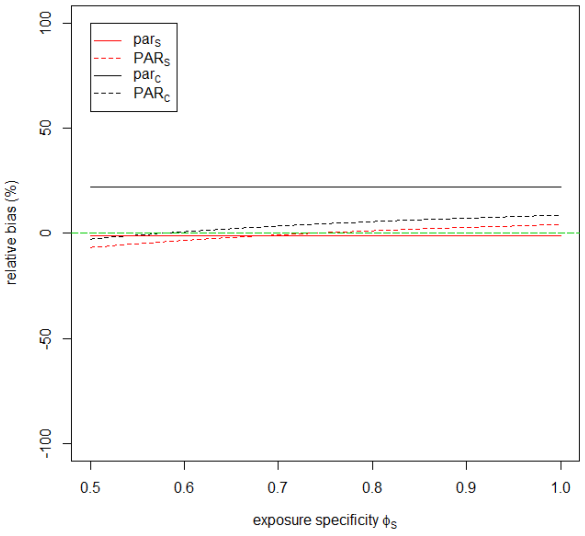

Figure 1.6: Effect of varying ϕ_S, exposure specificity, holding $\theta_S = \phi_T = \theta_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

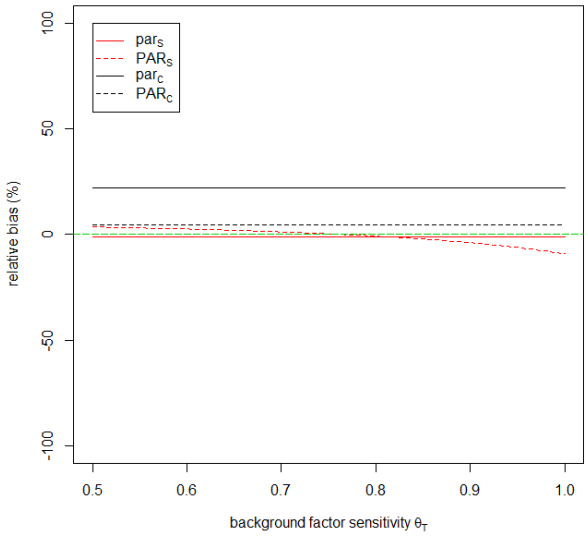

Figure 1.7: Effect of varying θ_T, background factor sensitivity, holding $\phi_S = \theta_S = \phi_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

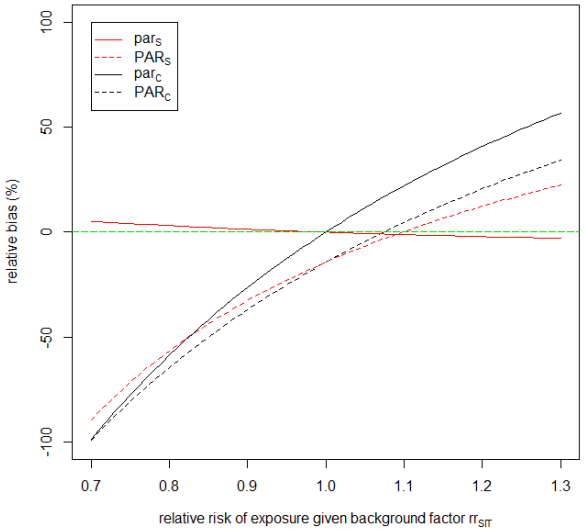

Figure 1.8: Effect of varying $rr_{S|T}$, association of S with T, holding $\phi_S = \theta_S = \phi_T = \theta_T = p_{S.} = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$

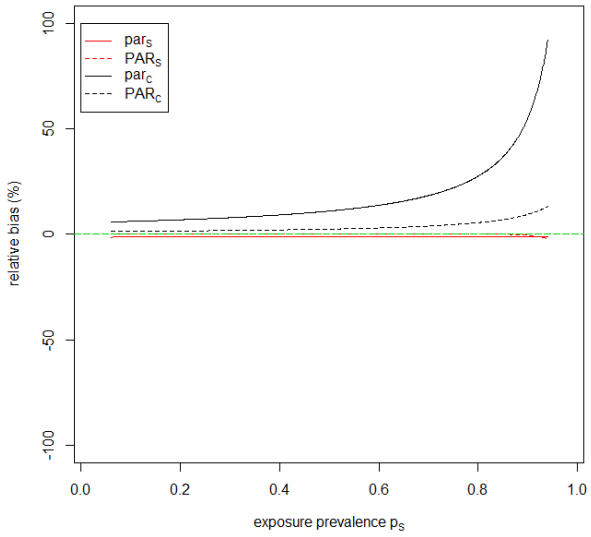

Figure 1.9: Effect of varying $p_{S.}$, holding
$\phi_S = \theta_S = \phi_T = \theta_T = p_{.T} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

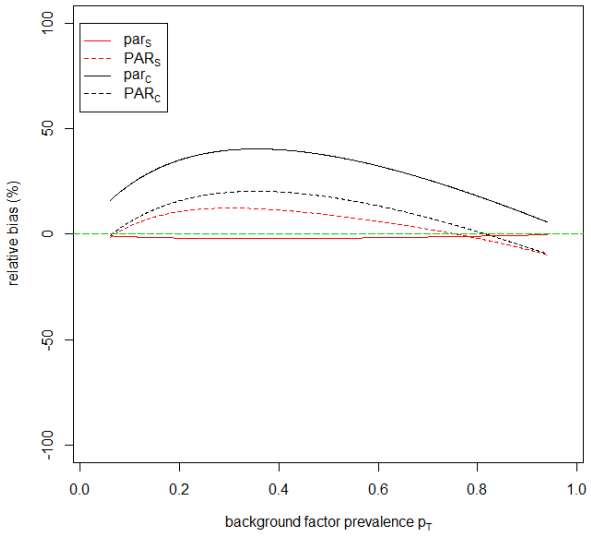

Figure 1.10: Effect of varying $p_{.T}$, holding
$\phi_S = \theta_S = \phi_T = \theta_T = p_{S.} = 0.75$, $rr_{1S} = 1.25$, $rr_{2T} = 3$, $rr_{S|T} = 1.1$

Correction for Risk Factor Misclassification in the Partial Population Attributable Risk

Benedict Wong, Donna Spiegelman and Molin Wang

Department of Biostatistics

Harvard T.H Chan School of Public Health

Chapter 2 Abstract

Estimation of the population attributable risk (PAR) has become an important goal in public health research, because it describes the proportion of disease cases that could be prevented if an exposure were entirely eliminated from a target population as a result of some intervention. In epidemiologic studies, categorical covariates are often misclassified. We present methods for obtaining point and interval estimates of the PAR and the partial PAR in the presence of misclassification, using a likelihood-based approach to estimate parameters in the logistic regression models for the disease and for the misclassification process, under two different study designs. We assess the performance of this method via a simulation study, and use it to obtain corrected point and interval estimates of the pPAR for high red meat intake in relation to colorectal cancer incidence in the Health Professionals Follow-Up Study of risk factors.

2.1 Introduction

The population attributable risk (PAR) is defined as the fraction of disease cases that would be prevented if an exposure were to be entirely eliminated from the population of interest. It has attracted much interest in cancer research, as well as in epidemiology and health policy, as it allows us to evaluate the impact of public health interventions which remove the harmful exposure. According to Mary Northridge, former editor-in-chief of the American Journal of Public Health, if the goal is to estimate the amount or proportion of cases of a disease attributable to a given risk factor, or to predict the impact of medical and public health interventions on the health status of a population, then PARs are particularly relevant (Northridge, 1995).

In a single exposure setting, the PAR is a function of the relative risk and the prevalence of the exposure (Levin, 1952). In the presence of confounding due to existing risk factors for the disease under study whose distribution is not affected by the interventions, the effect of the interventions can be evaluated using the partial population attributable risk (pPAR) (Bruzzi et al., 1985). The pPAR is also called the adjusted attributable risk (Benichou, 2001).

In epidemiologic studies, it is common for categorical variables to be misclassified, resulting in biased estimates for the exposure prevalences as well as the relative risks. This in turn affects the accuracy of the pPAR estimates, which are functions of the exposure prevalences and the relative risk estimates. The effect of non-differential risk factor misclassification on the PAR estimates in the single-exposure setting has received much attention. Misclassification is said to be non-differential when it is independent of disease status, that is, when exposure sensitivity and specificity are the same for both the disease cases and the non-cases (Johnson et al., 2014). Hsieh and Walter (Hsieh and Walter, 1988) showed that when there is non-differential misclassification in a single binary exposure and when there is imperfect sensitivity, both the PAR and the disease-exposure odds ratio will be underestimated. They also showed that when there is perfect sensitivity, the odds ratio is again underestimated but the PAR is unbiased. On the other hand, when misclassification is differential, the bias in PAR estimates can be in either direction (Copeland et al., 1977). Misclassification is said to be differential when exposure sensitivity and specificity differ between the disease cases and the non-cases (Johnson et al., 2014), and can arise through the dichotomization of a continuous exposure which is subject to non-differential measurement error (Dalen et al., 2009). Other studies have also examined the effect of outcome misclassification on the PAR (Hsieh, 1991; Vogel et al., 2005), and a single paper has examined the effect of exposure misclassification on the pPAR in the two-exposure setting (Wong et al., 2018). In the latter study, it was shown that in the presence of exposure misclassification, the bias in the pPAR can be in either direction, unlike the bias in the single-exposure PAR which can only be toward the null. In addition, these authors found that the magnitude of the bias is most dependent on the sensitivity of the exposure being eliminated. These findings motivate the need for developing tools that can help researchers estimate unbiased pPARs and confidence intervals in the presence of misclassification.

In Section 2.2, we describe the methods for correcting the exposure-misclassification induced bias in pPAR estimates and confidence intervals. We assess the performance of our method via an extensive simulation study in Section 2.3, and we apply our methods to estimate the PARs for colorectal cancer (CRC) in the Health Professionals Follow-Up

Study (HPFS) (Rimm et al., 1991) in Section 2.4. Section 2.5 concludes this paper.

2.2 Notation

For a single binary exposure, the PAR is defined as the proportion of disease cases that would have been avoided if the exposure were completely eliminated from the population. It is given by

$$par = 1 - \frac{I_0}{(1-p)\cdot I_0 + p \cdot I_1} = 1 - \frac{1}{(1-p) + p \cdot rr}, \tag{2.1}$$

where p is the prevalence of the exposure of interest in the population, I_1 is the probability of disease among the exposed, I_0 is the probability of disease among the non-exposed, and rr is the relative risk representing the exposure-disease association, defined by $\frac{I_1}{I_0}$. Hence, the PAR is the expected percentage reduction in disease cases resulting from a public health intervention that removes this exposure.

If there exists another exposure that is either non-modifiable or is not the target of the public health intervention, the above formula needs to be revised. Let X_1 be the modifiable exposure and X_2 be the non-modified exposure. When X_1 and X_2 are binary, taking values in $\{0, 1\}$, then, under the assumption of no interaction between the modifiable and non-modified exposures in the model for the disease outcome, the partial PAR (pPAR) is

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2}, \tag{2.2}$$

(Spiegelman et al., 2007), where $p_{st} = P(X_1 = s, X_2 = t)$, for $s, t = 0, 1$, $rr_1 = \frac{P(Y=1|X_1=1)}{P(Y=1|X_1=0)}$, with Y being the binary value of the disease, and $rr_2 = \frac{P(Y=1|X_2=1)}{P(Y=1|X_2=0)}$.

In practice, the assumption of no interaction may not hold, although it usually does. When not, there is an ineraction, the above equation can be written as

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_3}, \tag{2.3}$$

where $rr_3 = \frac{P(Y=1|X_1=X_2=1)}{P(Y=1|X_1=X_2=0)}$, with rr_1 and rr_2 redefined such that $rr_1 = \frac{P(Y=1|X_1=1,X_2=0)}{P(Y=1|X_1=X_2=0)}$ and $rr_2 = \frac{P(Y=1|X_1=0,X_2=1)}{P(Y=1|X_1=X_2=0)}$.

When the exposures are subject to non-differential misclassification in standard study designs, we observe values for the surrogate exposures rather than the true exposures. An example of a surrogate exposure in epidemiologic research is self-reported dietary intake, which is a commonly used, inaccurate substitute for actual dietary intake, the true exposure. Let X_k be the true exposure and Z_k be the corresponding surrogate exposure, for $k = 1, 2$. The uncorrected pPAR is given by

$$PAR_P = 1 - \frac{(P_{00} + P_{10}) + (P_{01} + P_{11})RR_2}{P_{00} + P_{10}RR_1 + P_{01}RR_2 + P_{11}RR_3}, \tag{2.4}$$

where uppercase letters indicate that the parameter values are associated with the surrogate exposures and not the true exposures, with $P_{st} = P(Z_1 = s, Z_2 = t)$, for $s, t = 0, 1$, $RR_1 = \frac{P(Y=1|Z_1=1,Z_2=0)}{P(Y=1|Z_1=Z_2=0)}$, $RR_2 = \frac{P(Y=1|Z_1=0,Z_2=1)}{P(Y=1|Z_1=Z_2=0)}$, and $RR_3 = \frac{P(Y=1|Z_1=Z_2=1)}{P(Y=1|Z_1=Z_2=0)}$.

We can extend the expressions in (2.3) and (2.4) for the scenarios where there may be multiple binary exposures and/or categorical exposures with more than two risk levels. Suppose there are S possible unique combinations from the set of modifiable exposures, and T possible unique combinations from the set of non-modified exposures. Without the assumption of no interaction between the modifiable exposures and the non-modified exposures, the pPAR can be expressed as:

$$par_P = 1 - \frac{\sum_{t=0}^{T-1} p_{.t} rr_{2t}}{\sum_{s=0}^{S-1} \sum_{t=0}^{T-1} p_{st} rr_{3st}}, \tag{2.5}$$

where s denotes a stratum of unique combinations of levels of all exposures that are modifiable, t denotes a stratum of unique combinations of levels of all background risk factors which are not modified, p_{st} indicates the proportion of the population with respective risk factor levels s and t, with 0 indexing the lowest risk levels, and $p_{.t} = \sum_{s=0}^{S-1} p_{st}$ for all t. Define $rr_{2t} = \frac{P(Y=1|X_1=0,X_2=t)}{P(Y=1|X_1=X_2=0)}$, and $rr_{3st} = \frac{P(Y=1|X_1=s,X_2=t)}{P(Y=1|X_1=X_2=0)}$, where $rr_{20} = 1$ and $rr_{300} = 1$. The uncorrected pPAR in the multi-factorial setting can be analagously defined.

As the pPAR is a function of the exposure prevalences as well as the relative risks, it is necessary to obtain estimates for these parameters, in order to estimate the pPAR. In the subsequent sections, we present methods for obtaining bias-corrected estimates for the pPAR and its confidence intervals in the presence of misclassification.

2.2.1 The Disease and Misclassification Models

In a study with a total of K binary exposures, including both the modifiable and non-modified exposures, $S \times T = 2^K$. Define $\boldsymbol{X} = (X_1, ..., X_K)$ to be the vector of the true exposure values, and let $\boldsymbol{Z} = (Z_1, ..., Z_K)$ be the vector of the surrogate exposure values, with Z_k the surrogate of X_k for all $k = 1, ..., K$.

When the disease outcome is binary, it is common to use a logistic regression model for the exposure-outcome relationship:

$$f_1(Y|\boldsymbol{X};\boldsymbol{\beta}) = \frac{e^{Y(\beta_0+\boldsymbol{X}\boldsymbol{\beta})}}{1+e^{(\beta_0+\boldsymbol{X}\boldsymbol{\beta})}}, \tag{2.6}$$

where Y is the disease outcome, β_0 is the baseline log-odds of disease for individuals who do not have any exposures, and $\boldsymbol{\beta}$ is the k-vector of log-odds ratios representing the conditional $X - Y$ associations. Throughout this paper, we use f to denote probability functions.

The misclassification process can be modeled as:

$$f_2(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\psi}) = f_2(Z_1, ..., Z_K|X_1, ..., X_K;\boldsymbol{\psi}), \tag{2.7}$$

where ψ is the vector of parameters that characterize the relationship between $\boldsymbol{Z}$ and $\boldsymbol{X}$. When it is reasonable to assume that the misclassification process is conditionally independent across the K exposures, we can simplify the model to

$$f_2(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{\psi}) = \prod_{k=1}^{K} f_{2,k}(Z_k|X_k;\psi_k),$$

where $\psi_k = (\psi_{k,1}, \psi_{k,2})$ can be seen as a reparameterization of the exposure sensitivities, $\Pr(Z_k = 1|X_k = 1)$, and specificities, $\Pr(Z_k = 0|X_k = 0)$, for each (X_k, Z_k) pair, for $k = 1, ..., K$.

As an alternative to modeling the misclassification process, we can model the reclassification process as:

$$f_3(\boldsymbol{X}|\boldsymbol{Z};\boldsymbol{\gamma}) = f_3(X_1, ..., X_K|Z_1, ..., Z_K;\boldsymbol{\gamma}) \tag{2.8}$$

where γ is the vector of parameters that relate $\boldsymbol{X}$ to $\boldsymbol{Z}$. Following (Spiegelman et al., 2000), we can decompose the above equation into a sequence of conditional models:

$$f_3(\boldsymbol{X}|\boldsymbol{Z};\boldsymbol{\gamma}) = f_{3,1}(X_1|Z_1,...,Z_K;\boldsymbol{\gamma_1})f_{3,2}(X_2|X_1,Z_1,...,Z_K;\boldsymbol{\gamma_2})...f_{3,K}(X_K|X_1,...,X_{K-1},Z_1,...,Z_K;\boldsymbol{\gamma_K})$$

When it is reasonable to assume that the reclassification process is conditionally independent across the K exposures, we can simplify the model to

$$f_3(\boldsymbol{X}|\boldsymbol{Z};\boldsymbol{\gamma}) = \prod_{k=1}^{K} f_{3,k}(X_k|Z_k;\gamma_k),$$

where $\gamma_k = (\gamma_{k,1},\gamma_{k,2})$ can be seen as a reparameterization of the positive predictive values, defined as $\Pr(X_k = 1|Z_k = 1)$, and the negative predictive values, defined as $\Pr(X_k = 0|Z_k = 0)$, for each (X_k, Z_k) pair, for $k = 1,...,K$.

Note that even if the misclassification process is conditionally independent across the K exposures, it is necessary that the exposures are also mutually independent, in order for the reclassification process to be conditionally independent across the K exposures.

2.2.2 The Likelihood

In a main study/internal validation study (MS/IVS) design, validation data is obtained from participants who are also part of the main study. All participants in the main study provide data on $\boldsymbol{Y}$ and $\boldsymbol{Z}$, and the participants in the internal validation study provide data on $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$. Then $i = 1,...,n_M$ indexes the participants providing only main study data and $i = n_M + 1,...,n_M + n_V$ indexes the participants additionally providing validation study data. For the i-th subject, $\mathbf{X}_i = (X_{i1},...,X_{iK})$ and $\mathbf{Z}_i = (Z_{i1},...,Z_{iK})$. The joint likelihood for the observed data in this MS/IVS design is

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\zeta}) &= \prod_{i=1}^{n_M} f_4(Y_i|\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma})f_5(\boldsymbol{Z}_i;\boldsymbol{\zeta}) \prod_{i=n_M+1}^{n_M+n_V} f_1(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})f_3(\boldsymbol{X}_i|\boldsymbol{Z}_i;\boldsymbol{\gamma})f_5(\boldsymbol{Z}_i;\boldsymbol{\zeta}) \\ &= \prod_{i=1}^{n_M} f_4(Y_i|\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) \prod_{i=n_M+1}^{n_M+n_V} f_1(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})f_3(\boldsymbol{X}_i|\boldsymbol{Z}_i;\boldsymbol{\gamma}) \prod_{i=1}^{n_M+n_V} f_5(\boldsymbol{Z}_i;\boldsymbol{\zeta}) \end{aligned} \tag{2.9}$$

for $i = 1, ..., n_M$ for the participants that provided only main study data, and $i = n_M + 1, ..., n_M + n_V$ for the participants in the validation study, and where

$$f_4(Y_i|\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{\boldsymbol{x}} f_1(Y_i|\boldsymbol{x};\boldsymbol{\beta}) f_3(\boldsymbol{x}|\boldsymbol{Z}_i;\boldsymbol{\gamma})$$

due to the surrogacy assumption that $f(Y|\boldsymbol{X},\boldsymbol{Z}) = f(Y|\boldsymbol{X})$. Assume

$$f_5(\boldsymbol{Z}_i;\boldsymbol{\zeta}) = \prod_{j=1}^{2^K} \zeta_j^{I(1+\sum_{k=1}^{K} Z_{ik}\cdot 2^{k-1}=j)}$$

where j denotes a stratum of unique combinations of levels of all the exposures, $j = 1, ..., 2^K$, and ζ_j is the proportion of individuals in the full study population who belong to the j-th stratum.

In a main study/external validation (MS/EVS) study design, the participants in the main study provide data on $\boldsymbol{Y}$ and $\boldsymbol{Z}$, and the participants in the external validation study provide only data on $\boldsymbol{X}$ and $\boldsymbol{Z}$, but not $\boldsymbol{Y}$. When using external validation data, it is necessary to assume that the misclassification process, for $\boldsymbol{Z}|\boldsymbol{X}$, between the main and validation studies are similar. This assumption is empirically unverifiable, however, it is necessary in order for the external validaton study data to be used for misclassification correction. Where appropriate, we can assume that the exposure prevalences in the validation study are similar to those in the main study. We refer to this as double transportability, because both the misclassification process and the exposure prevalences are assumed to be transportable from the validation study to the main study.

Under both assumptions, the joint likelihood for the observed data is

$$L(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\zeta}) = \prod_{i=1}^{n_M} f_4(Y_i|\boldsymbol{Z}_i;\boldsymbol{\beta},\boldsymbol{\gamma}) \prod_{i=n_M+1}^{n_M+n_V} f_3(\boldsymbol{X}_i|\boldsymbol{Z}_i;\boldsymbol{\gamma}) \prod_{i=1}^{n_M+n_V} f_5(\boldsymbol{Z}_i;\boldsymbol{\zeta}), \tag{2.10}$$

where the $f_1(.)$ term has been omitted from the expression in (2.9).

If the second assumption does not hold, i.e. the exposure prevalences in the validation study are different from those in the main study, then the joint likelihood for the observed data is

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta}) = \prod_{i=1}^{n_M} \sum_{\boldsymbol{x}} f_1(Y_i|\boldsymbol{x}; \boldsymbol{\beta}) f_2(\boldsymbol{Z}_i|\boldsymbol{x}; \boldsymbol{\psi}) f_6(\boldsymbol{x}; \pi) \prod_{i=n_M+1}^{n_M+n_V} f_2(\boldsymbol{Z}_i|\boldsymbol{X}_i; \boldsymbol{\psi}). \tag{2.11}$$

Depending on the study design and the assumptions made, we can maximize the likelihood expression in Equation (2.9), (2.10) or (2.11), to obtain the maximum likelihood estimates for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\zeta})$ or $(\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\pi})$. In a MS/IVS design, or in a MS/EVS design with double transportability, we can then obtain estimates for the pPAR, $\widehat{pPAR}$, by calculating the joint exposure prevalence estimates of $\boldsymbol{X}$ from the estimators $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\zeta}})$ and the relative risk estimates from the estimators $\widehat{\boldsymbol{\beta}}$. In a MS/EVS design where only the misclassification process is transportable, we can use the estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\pi}})$ to calculate the pPAR estimates.

2.2.3 Interval estimates

The variance of $\widehat{pPAR}$ can be obtained using the multivariate delta method, given the variance-covariance matrix of the estimators , which can be estimated as the inverse of the negative of the observed information matrix of the logarithms of (2.9), (2.10) or (2.11).

2.3 Simulation Study

We conducted a simulation study to assess the finite sample performance of our method under the main study/internal validation study design and with five different parameter settings.

We chose our underlying parameter values based on the estimates from the Health Professionals Follow-Up Study (HPFS) data described in Section 2.4. We chose n_V to be 125, which was similar to the validation study size of 127 in the HPFS. However, for feasibility reasons, we reduced n_M by an order of magnitude to 5000. The baseline log odds of disease was increased from -6.28 to -3.14 so as to obtain a sufficient number of cases per simulated dataset, and increased the log OR estimates slightly to 0.3364 and 0.2878, which correspond to odds ratios of 1.40 and 1.333 respectively. In our simulations, we used a simple model for the misclassification process. Let i and j index strata of unique combinations of $\mathbf{X}$ and $\mathbf{Z}$ respectively, where $\mathbf{X} = (X_1, X_2)$ and $\mathbf{Z} = (Z_1, Z_2)$. For $i = 1, 2, 3, 4$, we set the probability of $j = i$ to be 70%. The probability that j would represent a stra-

tum different from i was 10% for each of the three other strata. We also investigated the potential impact of doubling the validation study size, as well as main study size, on the mean relative bias, mean squared error (MSE) and coverage probabililty (CP) of our point and interval estimates. The relative bias is defined as the percentage difference between the simulated pPAR estimates and the underlying pPAR value, calculated from the underlying parameter values for the exposure prevalences and relative risks.

For each simulated dataset, we estimated the pPAR for red meat, while treating alcohol intake as a non-modifiable exposure. In addition to the corrected pPAR estimates for the two different internal validation study sizes, we also calculated the pPAR that would have been estimated if *complete* data were observed for all participants in the study ($n_V = 5000, n_M = 0$), as well as the *uncorrected* pPAR which is obtained from regressing Y against Z in the main study. The complete data scenario is equivalent to that which occurs when the true exposure values are observed for all participants in the study.

The CP, MSE and relative bias in the pPAR are reported in Table 2.1. These results show that with a validation study of size 125, we can substantially reduce the bias of our point estimates and improve the coverage probability of our interval estimates for the pPAR, even in the face of substantial exposure misclassification.

We also repeated the simulations for different sets of parameter values for $\beta_1, \beta_2 = 0.1, 0.5$, to observe the coverage probabilities and bias properties would change when the parameters are varied. The results of these simulations are given in also given in Table 2.1. In all four of these new scenarios, our observations were similar to those from the scenario where $(\beta_1, \beta_2) = (0.336, 0.288)$. The MSE and relative bias for the corrected estimates generally decreased when the validation study size was increased, and the uncorrected estimates had extremely high mean relative bias. The coverage probabilities for the corrected interval estimates were also closer to the ideal value of 95% when the validation study size increased. The coverage probabilities for the uncorrected interval estimates were reasonably close to 95% when β_1 = 0.1, regardless of the value of β_2, but when $\beta_1 = 0.5$, the coverage probabilities were less than 50%, just like when $(\beta_1, \beta_2) = (0.336, 0.288)$.

Table 2.1: Coverage probability (CP), mean squared error (MSE) and relative/percentage bias (% Bias) in the pPAR, for $\beta_1 = \beta_2 = 0.5 = log(1.65)$ and $\beta_1 = \beta_2 = 0.1 = log(1.1)$ *Complete*: hypothetical scenario where X is given in the main study. *UC (Uncorrected)*: pPAR are obtained from naive estimates using Z in the main study. *IVS (Corrected)*: pPAR is derived from the likelihood based method for misclassification correction

	Complete	*UC*	*IVS*	*IVS*
n_V	5000	0	125	250
n_M	0	5000	4875	4750
$(\beta_1, \beta_2) = (0.336, 0.288)$				
% Bias	0.53	-41	5.4	1.59
MSE	0.005	0.022	0.082	0.046
CP (%)	95.7	41.7	92.5	94.4
$(\beta_1, \beta_2) = 0.5$				
% Bias	0.12	-44.9	1.70	0.26
MSE	0.004	0.018	0.019	0.014
CP (%)	96.2	41.7	96.8	95.7
$(\beta_1, \beta_2) = 0.1$				
% Bias	0.46	-43	4.1	0.72
MSE	0.006	0.006	0.022	0.019
CP (%)	95.5	94.5	96.2	95.4
$(\beta_1, \beta_2) = (0.1, 0.5)$				
% Bias	0.40	-31	0.89	-1.042
MSE	0.006	0.005	0.023	0.018
CP (%)	95.6	94.9	96.4	95.9
$(\beta_1, \beta_2) = (0.5, 0.1)$				
% Bias	-0.14	-47	1.08	-0.19
MSE	0.004	0.021	0.018	0.015
CP (%)	95.5	48.3	95.6	95.3

2.4 Application

We applied our likelihood-based method to the Health Professionals Follow-Up Study (HPFS) of risk factors for colorectal cancer (CRC) (Rimm et al., 1991). The HPFS began in 1986 when 51,529 male health professionals were enlisted to participate in the study. These participants filled in semi-quantitative questionnaires inquiring about topics such as dietary intake and diseases. The accuracy of the responses in the questionnaires was assessed using dietary records from a sub-sample of 127 participants (Hu et al., 1999). The information from these dietary records, together with the information provided by these 127 participants in the main questionnaires, formed our internal validation study data set. The information from the remaining 51402 participants formed the main study data set.

We aimed to estimate the corrected pPAR for CRC, treating high red meat intake as the modifiable exposure, while adjusting for the effect of high alcohol intake, the non-modified risk factor. High red meat intake (HRM) was defined as 2 or more servings of beef, pork or lamb each week, and high alcohol intake was defined as 7 or more servings of alcohol each week. We tested for an interaction effect between HRM and high alcohol intake, and did not find strong evidence of any interaction when comparing both models using a likelihood ratio test. We also included age as an additional risk factor in our model for the disease-exposure relationship, to obtain adjusted estimates for the relative risks for HRM and high alcohol intake. Let X_1 and Z_1 represent the true and surrogate values for HRM, while X_2 and Z_2 represent the true and surrogate values for high alcohol intake. For the reclassification model, we used a polytomous regression, following Equation (2.8), to model the conditional joint distribution of (X_1, X_2) given (Z_1, Z_2). Let i and j index unique combinations of $\mathbf{X}$ and $\mathbf{Z}$ respectively. Then for $i = 2, 3, 4, ..., 2^K$, and $j = 1, ..., 2^K$, $\log \frac{f_3(\boldsymbol{X}=i|\boldsymbol{Z}=j)}{f_3(\boldsymbol{X}=1|\boldsymbol{Z}=j)} = \gamma_{i-1,j}$

In Table 2.2, the first row shows the number of participants in each stratum of unique combinations of the surrogate exposures, among individuals who provided main study data only. The second row shows the number of participants in each stratum, among individuals who provided validation data. The last 4 rows shows the observed proportions

Table 2.2: Number/proportion of individuals with each exposure combination in main study (MS) and validation study (VS) of the Health Professionals Follow-Up Study (HPFS).
$\mathbf{Z} = (Z_1, Z_2)$, and $\mathbf{X} = (X_1, X_2)$, where X_1 and Z_1 represent the true and surrogate values for high red meat intake, while X_2 and Z_2 represent the true and surrogate values for high alcohol intake.

	$\mathbf{Z} = (0,0)$	$\mathbf{Z} = (1,0)$	$\mathbf{Z} = (0,1)$	$\mathbf{Z} = (1,1)$
No. in MS	6399	28889	1749	14365
No. in VS	18	65	5	39
$\mathbf{X} = (0,0)$	8	23	0	1
$\mathbf{X} = (1,0)$	7	34	0	4
$\mathbf{X} = (0,1)$	1	1	5	9
$\mathbf{X} = (1,1)$	2	7	0	25
$P(\mathbf{Z}\|\mathbf{X})$ in VS	$P(\mathbf{Z} = (0,0)\|\mathbf{X})$	$P(\mathbf{Z} = (1,0)\|\mathbf{X})$	$P(\mathbf{Z} = (0,1)\|\mathbf{X})$	$P(\mathbf{Z} = (1,1)\|\mathbf{X})$
$\mathbf{X} = (0,0)$	0.250	0.719	0	0.031
$\mathbf{X} = (1,0)$	0.156	0.755	0	0.089
$\mathbf{X} = (0,1)$	0.063	0.063	0.312	0.562
$\mathbf{X} = (1,1)$	0.059	0.206	0	0.735

of $f_{\boldsymbol{Z}|\boldsymbol{X}}(Z_1, Z_2|X_1, X_2)$ for all possible values of X_1, X_2, Z_1, Z_2 in the validation study.
We can see from Table 2.2 that there appears to be a substantial level of misclassification in the data. For example, 75% out of 32 individuals in the stratum with $X_1 = X_2 = 0$ had classified themselves into other strata, and likewise, 69% out of 16 individuals in the stratum with $X_1 = 0, X_2 = 1$ had incorrectly classified themselves in one or both risk exposures. We tested whether misclassification in HRM was dependent on high alcohol intake and vice versa, and found no evidence of any such dependencies. However, even though the misclassification model for $f_2(\mathbf{Z}|\mathbf{X})$ could be decomposed into $\prod_{k=1}^{2} f_{2,k}(Z_k|X_k)$ it does not guarantee that the reclassification model for $f_3(\mathbf{X}|\mathbf{Z})$ can be decomposed into $\prod_{k=1}^{2} f_{3,k}(X_k|Z_k)$, because another necessary condition for the latter to be true is that the exposures are independently distributed of one another, which is clearly not the case here. The naive, uncorrected estimates for the log odds ratios (log OR) for each of the exposures, HRM, high alcohol and age, were obtained from a logistic regression model regressing Y against Z. The uncorrected relative risk (RR) estimates, $\hat{RR}_k^{(U)}$, are showed in Table 2.3. These estimates were derived from the uncorrected log OR estimates, $\hat{\beta}_k^{(U)}$, and the uncorrected estimate for the baseline log odds of disease, $\hat{\beta}_0^{(U)}$, which is not shown in the

Table 2.3: Relative risk estimates based on data in the HPFS. Superscripts indicate whether estimates are *Uncorrected* or *Corrected*

k	*Exposure*	$\hat{RR}_k^{(U)}$ $(95\%\ C.I.)$	$\hat{RR}_k^{(C)}$ $(95\%\ C.I.)$
1	High Red Meat	1.06 (0.94, 1.20)	1.36 (0.99, 1.87)
2	High Alcohol	1.23 (1.12, 1.35)	1.32 (1.13, 1.53)
3	Age (+1 year)	1.05 (1.04, 1.05)	1.05 (1.04, 1.05)

table. In the table, we have that $k = 1, 2, 3$ corresponds to HRM, high alcohol intake and age, respectively.

The corrected relative risk estimates, $\hat{RR}_k^{(C)}$, were derived from the corrected log odds ratio estimates, $\hat{\beta}_k^{(C)}$, and the corrected estimate for the baseline log odds of disease, $\hat{\beta}_0^{(C)}$, where the $\hat{\beta}_k^{(C)}$ and $\hat{\beta}_0^{(C)}$ were obtained from maximizing the likelihood in Equation (2.9). While there was only an 7% difference in the uncorrected and corrected RR estimates for the effect of high alcohol intake, and the RR estimates associated with age were very similar, we observed that the corrected estimate for the RR associated with high red meat intake was 28% greater than its uncorrected counterpart.

In order to obtain corrected estimates for the pPAR, which is our primary quantity of interest, it is also necessary to estimate the corrected joint exposure prevalences. These are given in Table 2.4, where the $\hat{\pi}_{st}$ are the corrected joint prevalence estimates defined as $f_{\boldsymbol{X}}(X_1 = s, X_2 = t)$, for $s, t = 0, 1$, with values of 1 representing high red meat or alcohol intake respectively, and values of 0 representing low red meat or alcohol intake respectively, and where the $\hat{\zeta}_{st}$ are the uncorrected joint prevalence estimates defined as $f_{\boldsymbol{Z}}(Z_1 = s, Z_2 = t)$, estimated using the surrogate exposures. The estimate for the corrected pPAR is 0.196, with 95% C.I. (-0.044, 0.437), while the uncorrected pPAR estimate is 0.052, with 95% C.I. (-0.047, 0.151). The estimate for the corrected pPAR is about 280% greater than the uncorrected $\widehat{pPAR}$, due to the differences in the corrected and uncorrected estimates for the relative risks and the joint prevalences of the exposures. We also observe that the corrected pPAR estimate falls outside the estimated 95% confidence interval for the uncorrected estimate.

Table 2.4: Joint exposure prevalences (surrogate and true) estimated from data in the HPFS, with $\hat{\zeta}_{st} = P(Z_1 = s, Z_2 = t)$, $\hat{\pi}_{st} = P(X_1 = s, X_2 = t)$, where X_1 and Z_1 represent the true and surrogate values for high red meat intake, while X_2 and Z_2 represent the true and surrogate values for high alcohol intake.

Surrogate	*Estimates (95% CI)*	*True exposure*	*Estimates (95% CI)*
$\hat{\zeta}_{00}$	0.125 (0.122,0.127)	$\hat{\pi}_{00}$	0.262 (0.189, 0.336)
$\hat{\zeta}_{10}$	0.562 (0.558,0.566)	$\hat{\pi}_{10}$	0.370 (0.291, 0.449)
$\hat{\zeta}_{01}$	0.034 (0.033,0.036)	$\hat{\pi}_{01}$	0.115 (0.068, 0.161)
$\hat{\zeta}_{11}$	0.280 (0.276,0.283)	$\hat{\pi}_{11}$	0.253 (0.189, 0.317)

2.5 Discussion

Misclassification continues to be a problem in large studies, and results in biased estimates of exposure prevalences and relative risks that in turn lead to biased public health quantities of interest such as the pPAR. In this paper, we described a likelihood based method for obtaining corrected exposure prevalence and relative risk estimates in the presence of exposure misclassification, provided that validation study data is available. This allows us to obtain corrected point and interval estimates for the pPAR.

We applied the proposed method to data from the Health Professionals Follow-up Study, and observed that the joint exposure prevalences of the surrogate exposures are very different from the corrected estimates of the true joint exposure prevalences. We believe this is due to high level of misclassification in the exposures considered. The relatively large difference in the corrected and uncorrected estimates for the relative risk, representing the HRM-CRC association, can also be attributed to the high level of misclassification in the data. Consequently, when we calculated the pPAR estimates, we found that the uncorrected and corrected estimates also differed from each other, with the corrected point estimate falling outside the uncorrected interval estimate.

Through conducting a simulation study covering multiple study size combinations, we saw that even though the corrected pPAR estimates have greater MSE compared to the incorrected estimates, the corrected estimates have an average relative bias of less than 5%. In contrast, the uncorrected estimates have an average relative bias of about 40% toward the null. We concluded from the simulation study results that even a small validation

study of size 125 can go a long way in reducing the bias of pPAR estimates, compared with the uncorrected estimates that use surrogate exposure data. We also observed that increasing the validation study size to 250 can also result in a decrease in the MSE and relative bias of the point estimates and an improvement in the coverage probability of the interval estimates. Future work will develop methods to obtain corrected point and interval estimates for the pPAR when one or more risk factors are continuous.

A Bayesian Approach to Correcting for Risk Factor Misclassification

Benedict Wong, Donna Spiegelman and Lorenzo Trippa

Department of Biostatistics

Harvard T.H. Chan School of Public Health

Chapter 3 Abstract

Estimation of the population attributable risk (PAR) has become an important goal in public health research, because it describes the proportion of disease cases that could be prevented if an exposure were entirely eliminated from a target population as a result of some intervention. In epidemiological studies, categorical covariates are often misclassified. We present methods for obtaining point and interval estimates of the PAR in the presence of misclassification, using a Bayesian approach to estimate the parameters of the misclassification process, as well as the prevalences and relative risks of the exposures, under two different study designs. We apply this method to estimate the PAR in the Health Professionals Follow-Up Study of risk factors for colorectal cancer.

3.1 Introduction

The population attributable risk (PAR) is often defined as the fraction of disease cases that would be prevented if an exposure were to be entirely eliminated from the population of interest. It has attracted much interest in cancer research, as well as in epidemiology and public health, as it allows us to evaluate the impact of public health interventions which remove the harmful exposure. In the presence of confounding due to existing risk factors for the disease under study whose distribution is not affected by the interventions, the effect of the interventions can be evaluated using the partial population attributable risk (pPAR) (Bruzzi et al., 1985). In epidemiologic studies, it is common for categorical variables to be misclassified, resulting in biased estimates for the exposure prevalences as well as the relative risks. This in turn affects the accuracy of the pPAR estimates, which are functions of the exposure prevalences and the relative risk estimates.

Bayesian methods for analyzing the odds ratio estimates in the presence of misclassification have been proposed. Spiegelhalter et al. used Markov chain Monte Carlo (MCMC) methods to estimate the odds ratio of a disease in a case-control study when the binary exposure is subject to misclassification (Spiegelhalter et al., 1996). They do this by estimating the prevalence of the true exposure among the cases and controls separately, and obtain odds ratio estimates using a simple function of those prevalences.

In this paper, we propose a two stage Bayesian method for estimating the pPAR, where we estimate the overall prevalence of the exposures of interest in the population in the first stage, and we estimate the odds ratios directly in the second stage. Prescott and Garthwaite proposed a Bayesian method for odds ratio estimation in case-control studies with a single misclassified exposure (Prescott and Garthwaite, 2002). Their method works well only in the single-exposure setting, and require that an internal validation study is conducted. In contrast, our method is easily extended to multiple exposures, and allows for estimation of the adjusted odds ratios in both the internal validation study and external validation study designs.

In Section 3.2, we derive expressions for the true and uncorrected PARs, and we define the other quantities of interest. We also describe our proposed Bayesian method and its two stages. We apply the proposed methods to data from the U.S. Health Professionals Follow-Up Study for risk factors of colorectal cancer (Rimm et al., 1991). The results of a simulation study are presented in Section 3.3, and Section 3.4 concludes this paper.

3.2 Notation

In this section, we define the quantities of interest in the absence and presence of misclassification. For a single exposure, the PAR is defined as the proportion of disease cases that would have been avoided if the exposure were completely eliminated from the population. It is given by

$$par = 1 - \frac{I_0}{(1-p)\cdot I_0 + p \cdot I_1} = 1 - \frac{1}{(1-p) + p \cdot rr}, \tag{3.1}$$

where p is the prevalence of the exposure of interest in the population, I_1 is the probability of disease among the exposed, I_0 is the probability of disease among the non-exposed, and rr is the relative risk for the disease given the exposure, defined by $\frac{I_1}{I_0}$. Hence, the PAR is the expected percentage reduction in disease cases, resulting from a public health intervention that removes this exposure from the study population, and any other population to which the results can reasonably be generalized.

3.2.1 The partial PAR

If there exists other risk factors for the outcome under study that is either non-modifiable or is not the target of the public health intervention, the above formula needs to be revised. Let X_1 be the value of the modifiable exposure and X_2 be the value of the non-modified exposure. When X_1 and X_2 are binary variables, taking values 0 or 1, then, under the assumption of no interaction between the modifiable and non-modified exposures, the partial PAR (pPAR) is

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_1rr_2}, \tag{3.2}$$

where $p_{ij} = P(X_1 = i, X_2 = j)$, for $i, j = 0, 1$, $rr_1 = \frac{P(Y=1|X_1=1)}{P(Y=1|X_1=0)}$, where Y is the value of the disease and is binary, and $rr_2 = \frac{P(Y=1|X_2=1)}{P(Y=1|X_2=0)}$.

In practice, the assumption of no interaction may not hold, so the above equation can be written as

$$par_P = 1 - \frac{(p_{00} + p_{10}) + (p_{01} + p_{11})rr_2}{p_{00} + p_{10}rr_1 + p_{01}rr_2 + p_{11}rr_3}, \tag{3.3}$$

where, for $i \in \{1, 2, 3\}$, $rr_i = \frac{P(Y=1|X_1+2\cdot X_2=i)}{P(Y=1|X_1=X_2=0)}$.

In the presence of misclassification, we observe values for the surrogate exposures rather than the true exposures. An example of a surrogate exposure in epidemiologic research is self-reported dietary intake, which is a commonly used, inaccurate substitute for actual dietary intake, the true exposure of interest. The uncorrected pPAR is given by

$$PAR_P = 1 - \frac{(P_{00} + P_{10}) + (P_{01} + P_{11})RR_2}{P_{00} + P_{10}RR_1 + P_{01}RR_2 + P_{11}RR_3}, \tag{3.4}$$

where uppercase letters indicate that the parameter values are associated with the surrogate exposures and not the true exposures.

We can extend the expressions in (2.3) and (2.4) for the scenarios where there may be multiple binary exposures and/or categorical exposures with more than two risk levels. Suppose there are $S + 1$ possible unique combinations from the set of modifiable expo-

sures, and $T+1$ possible unique combinations from the set of non-modified exposures. Under the assumption of no interaction between the modifiable exposures and the non-modified exposures, the pPAR can be expressed as:

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

where s denotes a stratum of unique combinations of levels of all modifiable exposures that are removed, t denotes a stratum of unique combinations of levels of all non-modified exposures, p_{st} indicates the proportion of the population with respective risk levels s and t, with 0 indexing the lowest risk strata, rr_{1s} is the relative risk of disease for stratum s relative to the lowest risk stratum, $s = 0, ..., S$ and rr_{2t} is the relative risk in stratum t relative to the lowest risk stratum, $t = 0, ..., T$. Note that $rr_{10} = rr_{20} = 1$ and $p_{.t} = \sum_{s=0}^{S} p_{st}$ for all t. Each of the relative risk estimands has 2 subscripts, the first of which indicates whether it corresponds to the modifiable or non-modified exposures, and the second subscript indicates to which stratum it belongs. The expression for the uncorrected pPAR can be analogously defined.

3.2.2 The model, prior and posterior probability distributions

For $k = 1, ..., K$, let Z_k be the value of the surrogate exposure corresponding to X_k. When misclassification is assumed to be non-differential, define the marginal prevalence, sensitivity, specificity, and relative risk respectively as

$$\pi_k = P(X_k = 1),$$

$$\theta_k = P(Z_k = 1 | X_k = 1),$$

$$\phi_k = P(Z_k = 0 | X_k = 0),$$

and

$$rr_k = \frac{P(Y = 1 | X_k = 1)}{P(Y = 1 | X_k = 0)}.$$

The quantities of interest are the π_k and the rr_k because the pPAR is a function of these quantities. However, these parameters are not identifiable, even in large cohort studies, because only the Y's and the Z_k's are observed, but not the X_k's, due to exposure misclassification.

Nevertheless, it is possible to correct for bias arising from misclassification if validation data is available. One way this can be achieved is by obtaining gold standard measurements of the exposures of interest, from a subset of individuals in the main study. This is known as the main study/internal validation study design (MS/IVS), as validation data is obtained from within the main study. It is sometimes not feasible to obtain internal validation study data, due to reasons related to cost or otherwise. When this is the case, external validation data can often be obtained from another study, under the assumption that the levels of misclassification in the main study can be reasonably assumed to be similar to those underlying the external validation study data.

Let n_M be the number of subjects that provided only main study data, and let n_V be the number of subjects in the validation study, which could be internal or external. Furthermore, define $n_{V,k,x,z}$ to be the number of validation study subjects with $X_k = x$ and $Z_k = z$, for $x = 0, 1$ and $z = 0, 1$. Then, $n_{V,k,1,0} + n_{V,k,1,1}$ is an observation from the binomial distribution, Binomial (n_V, π_k). Likewise, $n_{V,k,1,1}$ and $n_{V,k,0,0}$ are observations from the respective binomial distributions, Binomial $(n_{V,k,1,1} + n_{V,k,1,0}, \theta_k)$ and Binomial $(n_{V,k,0,0} + n_{V,k,0,1}, \phi_k)$.

We use the Beta(1, 1) distribution as a prior distribution for π_k, θ_k, and ϕ_k, for all $k = 1, ..., K$. We chose the Beta distribution because the data in $\mathbf{X}$ and $\mathbf{Z}$ follows a binomial distribution and the Beta distribution is the conjugate prior distribution for the binomial distribution. We chose both parameters of the Beta distribution to be equal to 1 because that gives us a uniform prior distribution over the interval [0,1], which is the set of values that π_k, θ_k, and ϕ_k can take, for all $k = 1, ..., K$. An alternative prior is the Haldane improper prior, the Beta(0,0) distribution, however this prior leads to improper posterior distributions when the observed outcomes are extreme (Kerman, 2011). In the context of our method, the observed outcomes are extreme when one or more of $\{n_{V,k,1,1}, n_{V,k,1,0}, n_{V,k,0,1}, n_{V,k,0,0}\}$ are equal to 0.

For each of the log odds ratios, we use a uniform prior , $f(\beta_k) \propto 1, -\infty < \beta_k < \infty$ for all $k = 0, 1, ..., K$. We start by assuming that the parameters are mutually independent of one another, but we will also discuss more realistic scenarios, such as those where the exposures are not independent of one another. Under this assumption, the joint prior distribution of $(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is given by $f(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) = f(\beta_0) \prod_{k=1}^{K} f(\beta_k) f(\pi_k) f(\theta_k) f(\phi_k)$, where $\boldsymbol{\beta} = (\beta_0, \tilde{\boldsymbol{\beta}}) = (\beta_0, \beta_1, ...\beta_K)$, and $\boldsymbol{\pi} = (\pi_1, ...\pi_K)$, $\boldsymbol{\theta} = (\theta_1, ...\theta_K)$, $\boldsymbol{\phi} = (\phi_1, ...\phi_K)$. The joint posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is therefore given by

$$f(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi} | Y, Z, X) \propto f(Y, Z, X | \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) \tag{3.5}$$

using the fact that $f(\boldsymbol{\beta}) f(\boldsymbol{\pi}) f(\boldsymbol{\theta}) f(\boldsymbol{\phi}) \propto 1$.

3.2.3 The posterior distribution for the MS/IVS design

We can break down the study data into smaller components by defining $Z = (Z_M \; Z_V)^T$, where Z_M is the $n_M \times K$ submatrix of Z containing the rows of surrogate exposure values from subjects who only contributed main study data, and Z_V is the $n_V \times K$ submatrix containing surrogate exposure values from subjects who contributed validation study data. Likewise, $Y = (Y_M \; Y_V)$, in an internal validation study design, where Y_M is the n_M vector of outcome values from subjects who contributed only main study data, and Y_V is the n_V vector of outcome values from subjects who contributed validation study data. Note that $X = X_V$, because the true exposure values are only observed in subjects who contributed validation study data and X_V has the same dimensions as Z_V.
Now we can rewrite Equation (3.5) as

$$f(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi} | Y_M, Y_V, Z_M, Z_V, X_V) \propto f(Y_M, Y_V, Z_M, Z_V, X_V | \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$$

$$\propto f(Y_M | Z_M; \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) f(Y_V | X_V; \boldsymbol{\beta}) f(Z_M | \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) f(Z_V | X_V, \boldsymbol{\theta}, \boldsymbol{\phi}) f(X_V | \boldsymbol{\pi}) \tag{3.6}$$

Consider a scenario where we sought to obtain the posterior distributions of $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ in the absence of Y_M and Y_V. We will compare this pseudo-likelihood approach to the full likelihood approach, where we include the information in Y_M and Y_V in sampling $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$. Under this pseudo-likelihood approach, the joint posterior distribution of

$(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ would be given by

$$f(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}|Z_M, Z_V, X_V) \propto f(Z_M, Z_V, X_V|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) f(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$$

$$\propto f(Z_M, Z_V, X_V|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto f(Z_M|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) f(Z_V|X_V, \boldsymbol{\theta}, \boldsymbol{\phi}) f(X_V|\boldsymbol{\pi}) \tag{3.7}$$

Under an assumption of independence between the exposures as well as the assumption of independent misclassification of exposures, the last 2 terms in Equations (3.6) and (3.7) are

$$f(Z_V|X_V, \boldsymbol{\theta}, \boldsymbol{\phi}) f(X_V|\pi) = \prod_{k=1}^{K} f(Z_{V,k}|X_{V,k}, \theta_k, \phi_k) f(X_{V,k}|\pi_k) \tag{3.8}$$

$= \prod_{k=1}^{K} (\theta_k)^{n_{V,k,1,1}} (1-\theta_k)^{n_{V,k,1,0}} (\phi_k)^{n_{V,k,0,0}} (1-\phi_k)^{n_{V,k,0,1}} (\pi_k)^{n_{V,k,1,1}+n_{V,k,1,0}} (1-\pi_k)^{n_{V,k,0,1}+n_{V,k,0,0}}$

$$= \prod_{k=1}^{K} (\pi_k\theta_k)^{n_{V,k,1,1}} (\pi_k[1-\theta_k])^{n_{V,k,1,0}} ([1-\pi_k]\phi_k)^{n_{V,k,0,0}} ([1-\pi_k][1-\phi_k])^{n_{V,k,0,1}}$$

and the first term on the right hand side in Equation (3.7) is $f(Z_M|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \prod_{k=1}^{K} f(Z_{M,k}|\pi_k, \theta_k, \phi_k)$

$= \prod_{k=1}^{K} (\pi_k\theta_k + [1-\pi_k][1-\phi_k])^{n_{M,k,1}} (\pi_k[1-\theta_k] + [1-\pi_k]\phi_k)^{n_{M,k,0}}$ where $n_{M,k,z}$ is the number of main study subjects with $Z_k = z$. The three terms on the right hand side of Equation (3.7) are exactly equal to the last three terms in Equation (3.6). The second term in Equation (3.6) is given by

$$\prod_{i=1}^{n_V} \frac{e^{Y_{V,i}(\beta_0 + \boldsymbol{X_{V,i}}\tilde{\boldsymbol{\beta}})}}{1 + e^{(\beta_0 + \boldsymbol{X_{V,i}}\tilde{\boldsymbol{\beta}})}},$$

where $\boldsymbol{X_{V,i}}$ is the i-th row of the X_V submatrix, and the first term in (3.6) is

$$f(Y_M|Z_M; \boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \sum_x f(Y_M|x; \boldsymbol{\beta}) f(x|Z_M; \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}) \tag{3.9}$$

Now that we have the kernel for the posterior distribution, we can proceed to draw samples for $(\boldsymbol{\beta}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ by implementing a Metropolis-within-Gibbs sampling algorithm. We earlier noted that the three terms in Equation (3.7) are exactly equal to the last three terms in Equation (3.6). Therefore, to reduce computational burden, we can adopt a pseudo-likelihood approach by first drawing samples for $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ from the posterior distribution

defined in Equation (3.7), and then drawing samples for $\boldsymbol{\beta}$ from the conditional posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}, Y_M, Y_V, Z_M, Z_V, X_V)$, given the drawn values for $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$ at that iteration. We refer to these as the two stages of our Bayesian approach. Each of these 2 stages will be carried out in each iteration.

3.2.4 The first stage

In this stage, we sample from the joint posterior distribution of $(\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})$.
As it is reasonable to incorporate information from the validation study into our MCMC algorithm, for all $k = 1, ..., K$, we can draw J proposals for π_k, θ_k, and ϕ_k from the following distributions in each of J Metropolis-Hastings steps:

$$\pi_k^{(j)} \sim \text{Beta } (n_{V,k,1,1} + n_{V,k,1,0} + 1, n_{V,k,0,1} + n_{V,k,0,0} + 1)$$

$$\theta_k^{(j)} \sim \text{Beta } (n_{V,k,1,1} + 1, n_{V,k,1,0} + 1)$$

$$\phi_k^{(j)} \sim \text{Beta } (n_{V,k,0,0} + 1, n_{V,k,0,1} + 1)$$

for $j = 1, ..., J$.
Other proposal distributions may be considered, such as uniform distributions over the [0,1] interval. However, the acceptance rate for samples drawn from the uniform distribution are likely to be low. On the other hand, if the distributions of π, θ, and ϕ in the validation study are close to those in the main study, then the acceptance rate for our MCMC samples will be higher than when we use the Uniform [0,1] distribution. Moreover, the chosen distributions are computationally convenient choices because when these distributions are used to generate proposals, the acceptance probabilities reduce to functions of the counts in the main study, after the terms containing the counts in the validation study cancel each other out from the numerator and the denominator, which we will see in Equation (3.11).
Consider drawing a sample for π_k at the $(r+1)$-th step, which we will call $\pi_k^{(r+1)}$, then the acceptance probability for the proposal, π_k^*, would be given by

$$a_r(\pi_k^*) = min\Big(1, \frac{f(\pi_k^*|\theta_k^{(r)}, \phi_k^{(r)}, X, Z)}{f(\pi_k^{(r)}|\theta_k^{(r)}, \phi_k^{(r)}, X, Z)} \frac{g(\pi_k^{(r)})}{g(\pi_k^*)}\Big) \tag{3.10}$$

where $g(.)$ is the density function of the proposal distribution for π_k:

$$g(\pi_k) \propto (\pi_k)^{n_{V,k,1,1}+n_{V,k,1,0}}(1-\pi_k)^{n_{V,k,0,1}+n_{V,k,0,0}}$$

Note that$f(\pi_k|\theta_k,\phi_k,X_V,Z_V,Z_M) \propto f(Z_M|\pi_k,\theta_k,\phi_k)f(X_V|\pi_k) \propto h(.)g(.)$, where $h(.) = (\pi_k\theta_k+[1-\pi_k][1-\phi_k])^{n_{M,k,1}}(\pi_k[1-\theta_k]+[1-\pi_k]\phi_k)^{n_{M,k,0}}$

Now, after cancelling the $g(.)$ terms that appear in both the numerator and the denominator, we are left with the $h(.)$ terms in the following ratio:

$$\frac{f(\pi_k^*|\theta_k^{(r)},\phi_k^{(r)},X,Z)}{f(\pi_k^{(r)}|\theta_k^{(r)},\phi_k^{(r)},,X,Z)}\frac{g(\pi_k^{(r)})}{g(\pi_k^*)} = \frac{(\pi_k^*\theta_k^{(r)}+[1-\pi_k^*][1-\phi_k^{(r)}])^{n_{M,k,1}}(\pi_k^*[1-\theta_k^{(r)}]+[1-\pi_k^*]\phi_k^{(r)})^{n_{M,k,0}}}{(\pi_k^{(r)}\theta_k^{(r)}+[1-\pi_k^{(r)}][1-\phi_k^{(r)}])^{n_{M,k,1}}(\pi_k^{(r)}[1-\theta_k^{(r)}]+[1-\pi_k^{(r)}]\phi_k^{(r)})^{n_{M,k,0}}}, \tag{3.11}$$

proving that, after we have incorporated the information contained in X_V and Z_V into our choice of proposal distributions, the decision to accept or reject a proposal π_k^* depends only on new information contained in Z_M. Hence, if the fraction in Equation (3.10) is greater than 1, then $a_r(\pi_k^*)$ in Equation (3.10) is 1, and the proposal π_k^* will be accepted. However, if the fraction in Equation (3.10) is less than 1, then $a_r(\pi_k^*)$ will be equal to that fraction, and the proposal π_k^* will be accepted with probability $a_r(\pi_k^*)$

Using the same reasoning as above, we can show that the acceptance probabilities for proposals θ_k^* and ϕ_k^* depend only on new information contained in Z_M.

3.2.5 The second stage

In the second stage, we then draw samples for β from $f(\boldsymbol{\beta}|\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\phi},Y_M,Y_V,Z_M,X_V)$, the conditional posterior distribution of $\boldsymbol{\beta}$, given the data as well as the sampled values of $(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\phi})$ at each iteration.

$$f(\beta|\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\phi},Y_M,Y_V,Z_M,X_V) = \prod_{i=1}^{n_M}\Big[\sum_x f(Y_{Mi}|x;\boldsymbol{\beta})f(x|Z_{Mi};\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\phi})\Big]\cdot\prod_{i=n_M+1}^{n_M+n_V} f(Y_{Vi}|X_{Vi};\boldsymbol{\beta}) \tag{3.12}$$

We set the starting vector $\boldsymbol{\beta}^{(1)}$ to be $\hat{\boldsymbol{\beta}}_M$, where $\hat{\boldsymbol{\beta}}_M$ is the vector of logistic regression estimates obtained from regressing Y_M on Z_M in the naive model, which ignores misclassification in the main study. We use a Metropolis algorithm to obtain samples for the $\boldsymbol{\beta}$, such that for $r = 1,2,...,R-1$, the proposal vector for the next iterative step is drawn

from a Normal($\hat{\boldsymbol{\beta}}^{(r)}, I^{-1}_{\boldsymbol{\beta}_M\boldsymbol{\beta}_M}$) distribution, where $I^{-1}_{\boldsymbol{\beta}_M\boldsymbol{\beta}_M}$ is obtained from the earlier logistic regression. At the $(r+1)$-th step, when deciding whether to accept a proposal, $\boldsymbol{\beta}^*$, as our sample, $\boldsymbol{\beta}^{(r+1)}$, the acceptance probability is given by

$$a_r(\boldsymbol{\beta}^*_k) = min\Big(1, \frac{f(\boldsymbol{\beta}^*|\boldsymbol{\pi}^{(r+1)}, \boldsymbol{\theta}^{(r+1)}, \boldsymbol{\phi}^{(r+1)}, Y_M, Y_V, Z_M, X_V)}{f(\boldsymbol{\beta}^{(r)}|\boldsymbol{\pi}^{(r+1)}, \boldsymbol{\theta}^{(r+1)}, \boldsymbol{\phi}^{(r+1)}, Y_M, Y_V, Z_M, X_V)}\Big) \tag{3.13}$$

We performed a single MCMC run to obtain sets of samples for $\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}$, and $\boldsymbol{\beta}$. This gives us the joint posterior distribution of $\boldsymbol{\pi}$ and $\boldsymbol{\beta}$, which are the primary parameters of interests, as the pPAR only depends on them.

3.2.6 Sampling from the posterior distribution when the validation study is external

While an MS/IVS design is preferred over an MS/EVS design, there may be scenarios in which it is costly or otherwise unfeasible to conduct an internal validation study, in addition to the main study. Sometimes, a validation study from another study is readily available. As long as the transportability assumptions hold, such that the exposure prevalences and measurement error model in this external validation study are transportable to the current study, this external validation study can be used to obtain estimates that are corrected for measurement error .

The key difference between the external validation study and the internal validation study is the absence of Y_V in the former. This does not affect the first stage of our Bayesian approach, as Y_V is not used in the first stage. Furthermore, in the second stage, it does not affect our choice of proposal distribution, which is simply a multivariate Normal distribution centered at the previous sampled values. It does, however, require us to make a slight modification to the calculation of the acceptance probability in Equation (3.13) because the last term of the conditional probability, in Equation (3.12), is now omitted. The new conditional probability is given by

$$f(\beta|\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi}, Y_M, Y_V, Z_M, X_V) = \prod_{i=1}^{n_M} \Big[\sum_x f(Y_{Mi}|x; \boldsymbol{\beta}) f(x|Z_{Mi}; \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\phi})\Big] \tag{3.14}$$

Therefore, the sampled values of β under the MS/EVS design may differ from those un-

der the MS/IVS design. As a result, the sequence of samples, and hence the PAR estimates, may also differ.

3.3 Application

We applied our two-stage Bayesian method to the Health Professionals Follow-Up Study (HPFS) of risk factors for colorectal cancer. The HPFS began in 1986 when 51,529 male health professionals were enlisted to participate in the study (Rimm et al., 1991). These participants filled in semi-quantitative questionnaires inquiring about topics such dietary intake and diseases. The accuracy of the responses in the questionnaires was assessed using dietary records from a sub-sample of 127 participants.

We sought to obtain the corrected pPAR for high red meat intake, while treating alcohol intake and folate intake as the unmodified risk factors in this example. High red meat intake (HRM) was defined as 2 or more servings of beef, pork or lamb each week, and high alcohol intake was defined as 7 or more servings of alcohol each week. We included age in the model, as an error-free variable.

From the validation data, we obtained the values of $n_{V,k,x,z}$ and $n_{M,k,z}$, (x, z) in $\{0, 1\}$, for the risk factors being studied. We fitted the logistic regression model using the surrogate exposures, regressing Y_M on Z_M, and obtained the variance-covariance matrix $I^{-1}_{\beta_M\beta_M}$ that is to be used in our Metropolis step in the second stage. We then used the two-stage method to sample from the joint posterior distribution of $(\beta, \pi, \theta, \phi)$. We ran a chain of 20,000 iterations with a burn-in period of 1000 iterations, and we report the parameter estimates in Table 3.2. Convergence of the Markov chains was assessed by visually inspecting the trace plots, 2 of which are given in Figures 3.1 and 3.2, for evidence of mixing. The Geweke test confirmed that there was no difference between the posterior means of the first 10% and last 50% of the chain.

We observed that in this study, the effect of misclassification was greater on the binary variable corresponding to HRM, because the corrected estimate for the relative risk is 1.23, whereas the uncorrected estimate is 1.06. This is likely attributed to the very low specificity of red meat. We also observed that the relative risk estimate for high alcohol

Table 3.1: Counts from validation study and main study data in the Health Professionals Follow-Up Study (HPFS)

Exposure	$n_{V,k,1,1}$	$n_{V,k,1,0}$	$n_{V,k,0,0}$	$n_{V,k,0,1}$	$n_{M,k,1}$	$n_{M,k,0}$
Red Meat	70	9	14	34	43254	8148
Alcohol	39	11	72	5	16114	35288
Low Folate	96	18	11	2	38595	12934

Table 3.2: Parameter values estimated from data in the HPFS

Exposure	$\hat{\theta}_k$	$\hat{\phi}_k$	$\hat{\pi}_k$	$\hat{P}_k$	$\hat{\beta}_k$	$\hat{rr}$ $(95\% \; C.I.)$	$\hat{RR}$ $(95\% \; C.I.)$
Red Meat	0.88	0.30	0.62	0.84	0.21	1.23 (0.78, 1.96)	1.06 (0.94, 1.20)
Alcohol	0.74	0.94	0.40	0.31	0.29	1.33 (1.17, 1.53)	1.23 (1.12, 1.35)
Age (+5 years)	NA	NA	NA	NA	0.24	1.28 (1.25, 1.31)	1.27 (1.25, 1.30)

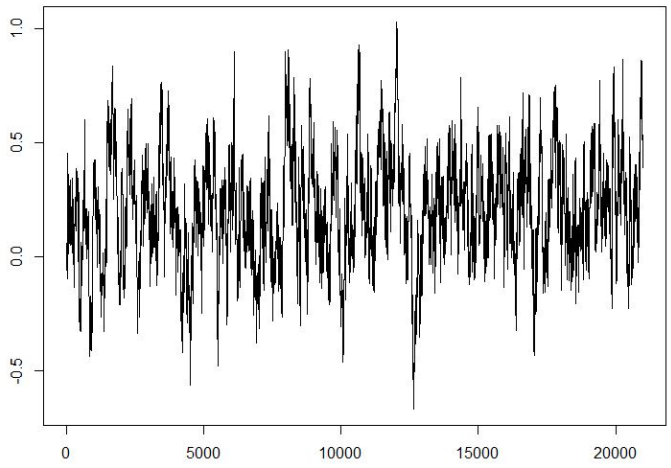

Figure 3.1: Trace plot for β_1

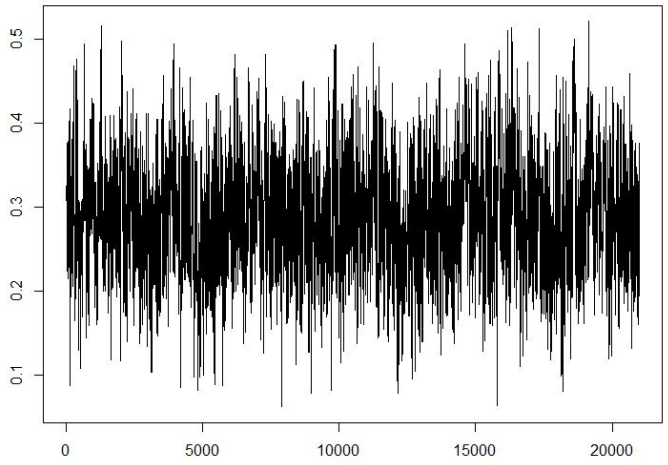

Figure 3.2: Trace plot for β_2

intake was attenuated due to the low sensitivity of this exposure. The relative risk estimate for a 5-year increase in age did not change much. In all cases, we used the odds ratio estimates as approximations for the relative risk estimates, given the low baseline risk of disease, with expit$(-6.229949) \approx 0.002$.

3.4 Discussion

In this paper, we proposed a two-stage Bayesian method for correcting for exposure misclassification in cohort studies, using standard MCMC methods that are well understood and relatively straightforward to implement. In the first stage of each iteration, we obtain samples from the posterior distribution of the exposure prevalences, sensitivities and specificities. In the second stage of each iteration, we obtain samples from the posterior distribution of the log odds ratios for the exposure-disease association, given the samples from the first stage. Our method incorporates information about the misclassification process from the validation study before analyzing the main study, and reduces the computational burden by choosing priors that are convenient to implement. Future work will develop MCMC methods to correct estimates for the PAR and pPAR when one or more mis-measured risk factors are continuous.

Summary

In chapter 1, we extended the work of Hsieh and Walter (Hsieh and Walter, 1988) to consideration of the pPAR, taking into account the multifactorial disease setting common in chronic disease epidemiology and public health research. We studied the behaviour of the pPAR, the semi-adjusted PAR, and the crude PAR, as well as their uncorrected estimands. Over a wide range of parameteres 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 non-differential misclassification.

Based on 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 semi-adjusted PAR, par_{semi}, fell within the interval for low bias nearly all of the time. Furthermore, we observed that θ_S and $rr_{S|T}$ 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 $rr_{S|T}$. The change in $bias_R$ with respect to a change in any of the six other parameters depended non-monotonically and non-linearly on the values of the other parameters, with $rr_{S|T}$ often having the greatest influence.

The results and findings from chapter 1 give us an idea of when $bias_R$ might be positive, e.g. 2 necessary conditions are that $rr_{S|T} < 1$ and $\theta_T + \phi_T < 2$, but they do not tell us the full extent of the bias, and its impact on estimation and inference in the pPAR. Hence, methods for correcting this bias are needed.

In chapter 2, we described a likelihood based method for obtaining corrected exposure prevalence and relative risk estimates in the presence of exposure misclassification, provided that validation study data is available. This method allows us to obtain corrected point and interval estimates for the pPAR.

Through conducting a simulation study covering multiple study size combinations, we saw that even though the corrected pPAR estimates have greater MSE compared to the in-

corrected estimates, the corrected estimates have an average relative bias of less than 5%. In contrast, the uncorrected estimates have an average relative bias of about 40% toward the null. We concluded from the simulation study results that even a small validation study of size 125 can go a long way in reducing the bias of pPAR estimates, compared with the uncorrected estimates that use surrogate exposure data. We also observed that increasing the validation study size to 250 can also result in a decrease in the MSE and relative bias of the point estimates and an improvement in the coverage probability of the interval estimates.

In chapter 3, we proposed a two-stage Bayesian method for correcting for bias in the PAR and pPAR due to exposure misclassification in cohort studies, using standard MCMC methods that are well understood and relatively straightforward to implement. In the first stage of each iteration, we draw samples from the posterior distribution of the exposure prevalences, sensitivities and specificities. In the second stage of each iteration, we draw samples from the posterior distribution of the log odds ratios for the exposure-disease association, given the samples from the first stage. Samples for the pPAR can be calculated after each iteration, given the samples of the exposure prevalences drawn in the first stage and the samples of the log odds ratio drawn in the second stage. Point estimates for the pPAR are obtained from the mean of the samples and credible intervals can be obtained from the relevant quantiles. Our method incorporates information about the misclassification process from the validation study before analyzing the main study, and reduces the computational burden by choosing priors that are convenient to implement. Future work will develop MCMC methods to correct estimates for the PAR and pPAR when one or more mis-measured risk factors are continuous.

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 chapter 1, it is possible to dramatically underestimate the pPAR and also to overestimate it, depending on the extent of misclassification, the prevalences of the risk factors, their association with one another, and the relative risks. The methods proposed in chapters 2 and 3 enable us to obtain corrected point and interval estimates for the pPAR, in the presence of risk factor misclassification, under different study designs that incorporate a validation sub-study.

Given the importance of the PAR and pPAR in driving health policy decisions, it is important that we correct for misclassification to ensure that we neither underestimate nor overstimate the pPAR. We believe that the methods and software developed in chapters 2 and 3 will be useful to researchers, clinicians and policymakers in the future.

References

BENICHOU, J. (2001). A review of adjusted estimators of attributable risk. *Statistical methods in medical research* **10** 195–216.

BLAIR, A., STEWART, P., LUBIN, J. H. and FORASTIERE, F. (2007). Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures. *American journal of industrial medicine* **50** 199–207.

BRUZZI, P., GREEN, S. B., BYAR, D. P., BRINTON, L. A. and SCHAIRER, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *American journal of epidemiology* **122** 904–914.

CERHAN, J. R., POTTER, J. D., GILMORE, J. M., JANNEY, C. A., KUSHI, L. H., LAZOVICH, D., ANDERSON, K. E., SELLERS, T. A. and FOLSOM, A. R. (2004). Adherence to the aicr cancer prevention recommendations and subsequent morbidity and mortality in the iowa women's health study cohort. *Cancer Epidemiology Biomarkers & Prevention* **13** 1114–1120.

COLE, P. and MACMAHON, B. (1971). Attributable risk percent in case-control studies. *British journal of preventive & social medicine* **25** 242–244.

COPELAND, K. T., CHECKOWAY, H., MCMICHAEL, A. J. and HOLBROOK, R. H. (1977). Bias due to misclassification in the estimation of relative risk. *American journal of epidemiology* **105** 488–495.

DALEN, I., BUONACCORSI, J. P., SEXTON, J. A., LAAKE, P. and THORESEN, M. (2009). Correction for misclassification of a categorized exposure in binary regression using replication data. *Statistics in medicine* **28** 3386–3410.

FESKANICH, D., RIMM, E. B., GIOVANNUCCI, E. L., COLDITZ, G. A., STAMPFER, M. J., LITIN, L. B. and WILLETT, W. C. (1993). Reproducibility and validity of food intake measurements from a semiquantitative food frequency questionnaire. *Journal of the American Dietetic Association* **93** 790–796.

FLEGAL, K. M., GRAUBARD, B. I. and WILLIAMSON, D. F. (2004). Methods of calculating deaths attributable to obesity. *American Journal of Epidemiology* **160** 331–338.

GIOVANNUCCI, E., RIMM, E. B., ASCHERIO, A., STAMPFER, M. J., COLDITZ, G. A. and WILLETT, W. C. (1995). Alcohol, low-methionine-low-folate diets, and risk of colon cancer in men. *Journal of the National Cancer Institute* **87** 265–273.

GIOVANNUCCI, E., RIMM, E. B., STAMPFER, M. J., COLDITZ, G. A., ASCHERIO, A. and WILLETT, W. C. (1994). Intake of fat, meat, and fiber in relation to risk of colon cancer in men. *Cancer research* **54** 2390–2397.

GOVINDARAJULU, U. S., SPIEGELMAN, D., MILLER, K. L. and KRAFT, P. (2006). Quantifying bias due to allele misclassification in case-control studies of haplotypes. *Genetic epidemiology* **30** 590–601.

GREENLAND, S. (1980). The effect of misclassification in the presence of covariates. *American Journal of Epidemiology* **112** 564–569.

GREENLAND, S. and MORGENSTERN, H. (1983). Morgenstern corrects a conceptual error. *American Journal of Public Health* **73** 703–704.

HANLEY, J. (2001). A heuristic approach to the formulas for population attributable fraction. *Journal of epidemiology and community health* **55** 508–514.

HSIEH, C.-C. (1991). The effect of non-differential outcome misclassification on estimates of the attributable and prevented fraction. *Statistics in medicine* **10** 361–373.

HSIEH, C.-C. and WALTER, S. D. (1988). The effect of non-differential exposure misclassification on estimates of the attributable and prevented fraction. *Statistics in medicine* **7** 1073–1085.

HU, F. B., RIMM, E., SMITH-WARNER, S. A., FESKANICH, D., STAMPFER, M. J., ASCHERIO, A., SAMPSON, L. and WILLETT, W. C. (1999). Reproducibility and validity of dietary patterns assessed with a food-frequency questionnaire–. *The American journal of clinical nutrition* **69** 243–249.

HUNTER, D. J., KRAFT, P., JACOBS, K. B., COX, D. G., YEAGER, M., HANKINSON, S. E., WACHOLDER, S., WANG, Z., WELCH, R., HUTCHINSON, A. ET AL. (2007). A genome-wide association study identifies alleles in fgfr2 associated with risk of sporadic postmenopausal breast cancer. *Nature genetics* **39** 870–874.

JOHNSON, C. Y., FLANDERS, W. D., STRICKLAND, M. J., HONEIN, M. A. and HOWARDS, P. P. (2014). Potential sensitivity of bias analysis results to incorrect assumptions of non-differential or differential binary exposures misclassification. *Epidemiology (Cambridge, Mass.)* **25** 902.

KLEINBAUM, D. G. and KUPPER, L. L. (1982). *Epidemiologic research: principles and quantitative methods*. John Wiley & Sons.

LEE, I.-M., SHIROMA, E. J., LOBELO, F., PUSKA, P., BLAIR, S. N., KATZMARZYK, P. T., GROUP, L. P. A. S. W. ET AL. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet* **380** 219–229.

LEVIN, M. L. (1952). The occurrence of lung cancer in man. *Acta-Unio Internationalis Contra Cancrum* **9** 531–541.

MOCELLIN, S., VERDI, D. and NITTI, D. (2009). Dna repair gene polymorphisms and risk of cutaneous melanoma: a systematic review and meta-analysis. *Carcinogenesis* **30** 1735–1743.

MORGENSTERN, H. (1982). Uses of ecologic analysis in epidemiologic research. *American journal of public health* **72** 1336–1344.

NORTHRIDGE, M. E. (1995). Public health methods–attributable risk as a link between causality and public health action. *American journal of public health* **85** 1202–1204.

PLATZ, E. A., WILLETT, W. C., COLDITZ, G. A., RIMM, E. B., SPIEGELMAN, D. and GIOVANNUCCI, E. (2000). Proportion of colon cancer risk that might be preventable in a cohort of middle-aged us men. *Cancer Causes & Control* **11** 579–588.

RIMM, E. B., GIOVANNUCCI, E. L., WILLETT, W. C., COLDITZ, G. A., ASCHERIO, A., ROSNER, B. and STAMPFER, M. J. (1991). Prospective study of alcohol consumption and risk of coronary disease in men. *The Lancet* **338** 464–468.

ROCKHILL, B., NEWMAN, B. and WEINBERG, C. (1998). Use and misuse of population attributable fractions. *American journal of public health* **88** 15–19.

SPIEGELMAN, D., HERTZMARK, E. and WAND, H. (2007). Point and interval estimates of partial population attributable risks in cohort studies: examples and software. *Cancer Causes & Control* **18** 571–579.

SPIEGELMAN, D., ROSNER, B. and LOGAN, R. (2000). Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association* **95** 51–61.

TAMIMI, R. M., SPIEGELMAN, D., SMITH-WARNER, S. A., WANG, M., PAZARIS, M., WILLETT, W. C., ELIASSEN, A. H. and HUNTER, D. J. (2016). Population attributable risk of modifiable and nonmodifiable breast cancer risk factors in postmenopausal breast cancer. *American Journal of Epidemiology* .

VOGEL, C., BRENNER, H., PFAHLBERG, A. and GEFELLER, O. (2005). The effects of joint misclassification of exposure and disease on the attributable risk. *Statistics in medicine* **24** 1881–1896.

VOGEL, C., GEFELLER, O. ET AL. (2002). Implications of nondifferential misclassification on estimates of attributable risk. *Methods Inf Med* **41** 342–348.

WALTER, S., HSIEH, C. and LIU, Q. (2007). Effect of exposure misclassification on the mean squared error of population attributable risk and prevented fraction estimates. *Statistics in medicine* **26** 4833–4842.

WONG, B. H., PESKOE, S. B. and SPIEGELMAN, D. (2018). The effect of risk factor misclassification on the partial population attributable risk. *Statistics in medicine* .

Appendix A

Supplementary Materials for Chapter 1

A.1 Derivation of the cell probabilities:

For the two-factor case where $S = T = 1$, let X_1 be a modifiable binary exposue and X_2 be a non-modified risk factor that is also binary. We derive the joint prevalences, p_{st} = $P(X_1 = s, X_2 = t)$, for $(s, t) \in \{0, 1\}^2$, given the marginal probabilities $p_{S.}$ and $p_{.T}$, and $rr_{S|T}$, the latter being defined as $rr_{S|T} = rr_{X_1=S|X_2=T} = \frac{P(X_1=S|X_2=T)}{P(X_1=S|X_2=0)} = \frac{P(X_1=1|X_2=1)}{P(X_1=1|X_2=0)}$.

Let $a = p_{00}$, $b = p_{S0}$, $c = p_{0T}$, $e = p_{ST}$

$e = p_{ST} = P(X_1 = 1, X_2 = 1) = p_{S.} * \frac{(p_{.T} * rr_{S|T})}{p_{.T} * rr_{S|T} + 1 - p_{.T}}$

$c = p_{0T} = P(X_1 = 0, X_2 = 1) = p_{.T} - e$

$b = p_{S0} = P(X_1 = 1, X_2 = 0) = p_{S.} - e$

$a = p_{00} = P(X_1 = 0, X_2 = 0) = 1 - p_{S.} - c$

Some constraints exists, that each of a, b, c, e must be bounded between 0 and 1. Since b, e are guaranteed to be bounded by construction, the constraints apply to a, c. The two constraints reduce to:

$p_{S.} \leq p_{.T} + (1 - p_{.T})/rr_{S|T}$ when $rr_{S|T} < 1$, and

$p_{S.} \leq rr_{S|T} * p_{.T} + (1 - p_{.T})$ when $rr_{S|T} > 1$.

Now for $i = 0, 1$, define a_i, b_i, c_i, and e_i:

$a_1 = P(D = 1, X_1 = 0, X_2 = 0) = p_d * \frac{a}{a + b * rr_{1S} + c * rr_{2T} + e * rr_{1S} * rr_{2T}}, a_0 = a - a_1$

$b_1 = P(D = 1, X_1 = 1, X_2 = 0) = p_d * \frac{b * rr_{1S}}{a + b * rr_{1S} + c * rr_{2T} + e * rr_{1S} * rr_{2T}}, b_0 = b - b_1$

$c_1 = P(D = 1, X_1 = 0, X_2 = 1) = p_d * \frac{c * rr_{2T}}{a + b * rr_{1S} + c * rr_{2T} + e * rr_{1S} * rr_{2T}}, c_0 = c - c_1$

$e_1 = P(D = 1, X_1 = 1, X_2 = 1) = p_d * \frac{e*rr_{1S}*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}, e_0 = e - e_1$

We now demonstrate that, under the assumption of no interaction between risk factors, rr_{1S} is the relative risk of having the disease in the event that $X_1 = 1$, regardless of the value of X_2, by noting that:

$\frac{b_1}{b}\frac{a}{a_1} = \frac{p_d * \frac{b*rr_{1S}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}}{b} \frac{a}{p_d * \frac{a}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}} = rr_{1S}$, and

$\frac{e_1}{e}\frac{c}{c_1} = \frac{p_d * \frac{e*rr_{1S}*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}}{e} \frac{c}{p_d * \frac{c*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}} = \frac{rr_{1S}*rr_{2T}}{rr_{2T}} = rr_{1S}$

A.2 An expression for the crude relative risk when 2 binary risk factors

The crude relative risk is given by:

$rr_{1S}^{(c)} = \frac{b_1+e_1}{b_1+b_0+e_1+e_0}\frac{a_1+a_0+c_1+c_0}{a_1+c_1} = \frac{b_1+e_1}{a_1+c_1}\frac{a_1+a_0+c_1+c_0}{b_1+b_0+e_1+e_0} = \frac{(p_{S.}-e)*rr_{1S}+e*rr_{1S}*rr_{2T}}{(1-p_{S.}-p_{.T}+e)+(p_{.T}-e)*rr_{2T}}\frac{1-p_{S.}}{p_{S.}}$

A.3 Expression for the misclassified model

From the law of total probability:

$P(Y=y|\vec{Z}=\vec{z}) = \sum_{\vec{x}} P(Y=y|\vec{X}=\vec{x},\vec{Z}=\vec{z})P(\vec{X}=\vec{x}|\vec{Z}=\vec{z})$

Using independence of X_1 and X_2:

$P(Y=y|\vec{Z}=\vec{z}) = \sum_{x_1,x_2\in\{0,1\}^2} P(Y=y|\vec{X}=\vec{x},\vec{Z}=\vec{z})P(X_1=x_1|Z_1=z_1)P(X_2=x_2|Z_2=z_2)$

Using the logit-link expression for $P(Y|X)$ and Bayes Theorem for $P(X|Z)$

$$= \sum_{x_1,x_2\in\{0,1\}^2} \frac{e^{y(\beta_0+\beta_1x_1+\beta_2x_2)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2)}} \prod_{j=1}^{2} \frac{\theta_j^{x_jz_j}\phi_j^{(1-x_j)(1-z_j)}(1-\theta_j)^{x_j(1-z_j)}(1-\phi_j)^{z_j(1-x_j)}p_j^{x_j}(1-p_j)^{(1-x_j)}}{(p_j\theta_j+[1-\phi_j][1-p_j])^{z_j}(\phi_j[1-p_j]+[1-\theta_j]p_j)^{(1-z_j)}} \tag{A.1}$$

Writing out the four terms in the sum yields

$$P(Y=y|\vec{Z}=\vec{z}) = \frac{e^{y(\beta_0)}}{1+e^{(\beta_0)}} \frac{\phi_1^{(1-z_1)}(1-\phi_1)^{z_1}(1-p_1)\phi_2^{(1-z_2)}(1-\phi_2)^{z_2}(1-p_2)}{\prod_{j=1}^{2}(p_j\theta_j+[1-\phi_j][1-p_j])^{z_j}(\phi_j[1-p_j]+[1-\theta_j]p_j)^{(1-z_j)}} \tag{A.2}$$

$$+\frac{e^{y(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}} \frac{\theta_1^{z_1}(1-\theta_1)^{(1-z_1)}p_1\phi_2^{(1-z_2)}(1-\phi_2)^{z_2}(1-p_2)}{\prod_{j=1}^{2}(p_j\theta_j+[1-\phi_j][1-p_j])^{z_j}(\phi_j[1-p_j]+[1-\theta_j]p_j)^{(1-z_j)}}$$

$$+\frac{e^{y(\beta_0+\beta_2)}}{1+e^{(\beta_0+\beta_2)}} \frac{\phi_1^{(1-z_1)}(1-\phi_1)^{z_1}(1-p_1)\theta_2^{z_2}(1-\theta_2)^{(1-z_2)}p_2}{\prod_{j=1}^{2}(p_j\theta_j+[1-\phi_j][1-p_j])^{z_j}(\phi_j[1-p_j]+[1-\theta_j]p_j)^{(1-z_j)}}$$

$$+\frac{e^{y(\beta_0+\beta_1+\beta_2)}}{1+e^{(\beta_0+\beta_1+\beta_2)}} \frac{\theta_1^{z_1}(1-\theta_1)^{(1-z_1)}p_1\theta_2^{z_2}(1-\theta_2)^{(1-z_2)}p_2}{\prod_{j=1}^{2}(p_j\theta_j+[1-\phi_j][1-p_j])^{z_j}(\phi_j[1-p_j]+[1-\theta_j]p_j)^{(1-z_j)}}$$

thus showing that the model is not linear in Z_1, Z_2 on the logistic scale.

A.4 Proving Independence of bias from p_d

We now prove that the PAR values are independent of the disease prevalence, p_d. Let us look at the expression for the pPAR.

$par_P = 1 - \frac{(a+b)+(c+e)*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{3ST}}$

$\Rightarrow$

$1 - par_P = \frac{(a+b)+(c+e)*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{3ST}} = \frac{(a+b)+(c+e)*rr_{2T}}{a+b*(b_1/b)/(a_1/a)+c*(c_1/c)/(a_1/a)+e*(e_1/e)/(a_1/a)} = \frac{(a+b)+(c+e)*(c_1/c)/(a_1/a)}{a+a*b_1/a_1+a*c_1/a_1+a*e_1/a_1}$

$= \frac{(a+b)+(c+e)*(c_1/a_1)/(c/a)}{a(1+b_1/a_1+c_1/a_1+e_1/a_1)}$

Consider b_1/a_1:

$\frac{b_1}{a_1} = \frac{p_d * \frac{b*rr_{1S}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}}{p_d * \frac{a}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}}} = \frac{b*rr_{1S}}{a}$, so the term p_d drops out as it appears in both the numerator and denominator.

Likewise for c_1/a_1 and e_1/a_1 and hence p_d drops out of the pPAR expression.

Alternatively, consider that rr_{1S}, rr_{2T} are preset parameters and $rr_{3ST} = rr_{1S} * rr_{2T}$, while a, b, c, e are functions of p_S, p_T, and $rr_{S|T}$ hence the pPAR only depends on these 5 parameters.

Likewise the crude PAR and the semi-adjusted PAR only depend on these parameters too.

$1 - par_c = \frac{1}{1-p_S+p_S rr_{1S}^{(c)}}$

$rr_{1S}^{(c)} = \frac{(e_1+b_1)/(e+b)}{(c_1+a_1)/(c+a)} = \frac{(e_1+b_1)/(c_1+a_1)}{(e+b)/(c+a)} = \frac{(e_1/a_1+b_1/a_1)/(c_1/a_1+1)}{(e+b)/(c+a)} = \frac{(e*rr_{3ST}+b*rr_{1S})/(c*rr_{2T}+a)}{(e+b)/(c+a)}$

For the uncorrected PARs, A_i, B_i, C_i, E_i (for $i = 0, 1$) are dependent only on a_i, b_i, c_i, e_i and $\theta_S, \phi_s, \theta_t, \phi_t$, and hence independent of p_d too. (Note: $A = A_0 + A_1$ and so on. These are given in Section A7)

As an example we can consider PAR_P. Following the first line:

$1 - PAR_P = \frac{(A+B)+(C+E)*(C_1/A_1)/(C/A)}{A(1+B_1/A_1+C_1/A_1+E_1/A_1)} = f(A_i, B_i, C_i, E_i) = f(a_i, b_i, c_i, e_i, \theta_j, \phi_j) = f(rr_j, rr_{S|T}, p_S, p_T, \theta_j, \phi_j)$

A.5 Proof of Zero Bias when $\theta_S = \theta_T = \phi_T = 1$:

We show that the PAR_P is unbiased when $\theta_S, \theta_T, \phi_T = 1$ regardless of the value of ϕ_S. The following are the misclassified versions of the joint prevalences previously defined in Section A1.

For i = 0,1:

$A_i = a_i * \phi_S \Rightarrow A = a * \phi_S$

$B_i = b_i + a_i * (1 - \phi_S) \Rightarrow B = b + a * (1 - \phi_S)$

$C_i = c_i * \phi_S \Rightarrow C = c * \phi_S$

$E_i = e_i + c_i * (1 - \phi_S) \Rightarrow E = e + c * (1 - \phi_S)$

Consider RR_{1S}, RR_{2T}:

$RR_{1S} = \frac{B_1 * A}{A_1 * B} = \frac{[b_1 + a_1 * (1-\phi_S)] * [a * \phi_S]}{[a_1 * \phi_S][b + a * (1-\phi_S)]} = \frac{[b_1 + a_1 * (1-\phi_S)] * a}{a_1 [b + a * (1-\phi_S)]}$

$= \frac{[a * b_1 + a * a_1 * (1-\phi_S)]}{[a_1 * b + a_1 * a * (1-\phi_S)]} = \frac{[a * b_1 + a * a_1 * (1-\phi_S)]}{[a_1 * b + a_1 * a * (1-\phi_S)]} = \frac{[a * b_1 - a_1 * b + a_1 * b + a * a_1 * (1-\phi_S)]}{[a_1 * b + a_1 * a * (1-\phi_S)]} = 1 + \frac{[a * b_1 - a_1 * b]}{[a_1 * b + a_1 * a * (1-\phi_S)]} = 1 + \frac{[a_1 * b(rr_{1S} - 1)]}{[a_1 * b + a_1 * a * (1-\phi_S)]}$

$\Rightarrow (RR_{1S} - 1) = \frac{[a_1 * b(rr_{1S} - 1)]}{[a_1 * b + a_1 * a * (1-\phi_S)]}$ confirming that $rr_{1S} = 1 \Rightarrow RR_{1S} = 1$

$RR_{1S} a_1 [b + a * (1 - \phi_S)] = [b_1 + a_1 * (1 - \phi_S)] * a = b_1 * a + a a_1 * (1 - \phi_S)$

$RR_{1S} [b + a * (1 - \phi_S)] = \frac{b_1 * a + a * a_1 * (1-\phi_S)}{a_1} = \frac{b_1 * a}{a_1} + a * (1 - \phi_S) = b * rr_{1S} + a * (1 - \phi_S)$

Similarly, $RR_{3ST}[e + c(1 - \phi_S)] = \frac{a * e_1 + a * c_1 (1-\phi_S)}{a_1} = e * rr_{1S} * rr_{2T} + c * rr_{2T}(1 - \phi_S)$

$RR_{2T} = \frac{C_1 * A}{A_1 * C} = \frac{c_1 * \phi_S * a * \phi_S}{a_1 * \phi_S * c * \phi_S} = \frac{c_1 * a}{a_1 * c} = rr_{2T}$

$1 - par_{P,(X_1)} = \frac{(a+b) + (c+e) * rr_{2T}}{a + b * rr_{1S} + c * rr_{2T} + e * rr_{1S} * rr_{2T}}$

$1 - PAR_{P,(X_1)} = \frac{(A+B) + (C+E) * RR_{2T}}{A + B * RR_{1S} + C * RR_{2T} + E * RR_{3ST}} = \frac{(a+b) + (c+e) * rr_{2T}}{a * \phi_S + [b + a * (1-\phi_S)] * RR_{1S} + c * \phi_S * rr_{2T} + [e + c * (1-\phi_S)] * RR_{3ST}}$

$= \frac{(a+b) + (c+e) * rr_{2T}}{a * \phi_S + b rr_{1S} + a * (1-\phi_S) + c * \phi_S * rr_{2T} + e * rr_{1S} * rr_{2T} + c * rr_{2T}(1-\phi_S)} = \frac{(a+b) + (c+e) * rr_{2T}}{a + b * rr_{1S} + c * rr_{2T} + e * rr_{1S} * rr_{2T}} = 1 - par_{P,(X_1)}$ Confirming that there is no bias when the sufficient condition of $\theta_S, \theta_T, \phi_T = 1$ is met, regardless of the value of ϕ_S.

A.6 Proof that the pPAR in the two-factor bivariate case reduces to the univariate case under the assumption of independence between risk factors and multiplicity of the relative risks:

In the scenario where $rr_{ST} = rr_{1S} * rr_{2T}$ and $rr_{S|T} = 1$, we have:

$par_P = 1 - \frac{(1-p_{.T})+p_{.T}*rr_{2T}}{a+b*rr_{1S}+c*rr_{2T}+e*rr_{1S}*rr_{2T}} = 1 - \frac{(1-p_{.T})+p_{.T}*rr_{2T}}{(1-p_{S.})*(1-p_{.T})+p_{S.}*(1-p_{.T})*rr_{1S}+(1-p_{S.})*p_{.T}*rr_{2T}+p_{S.}*p_{.T}*rr_{1S}*rr_{2T}}$

$= 1 - \frac{(1-p_{.T})+p_{.T}*rr_{2T}}{(1-p_{S.})*(1-p_{.T})+(1-p_{.T})*p_{S.}*rr_{1S}+(1-p_{S.})*p_{.T}*rr_{2T}+p_{S.}*rr_{1S}*p_{.T}*rr_{2T}}$

$= 1 - \frac{(1-p_{.T})+p_{.T}*rr_{2T}}{\left((1-p_{S.})+*p_{S.}*rr_{1S}\right)(1-p_{.T})+\left((1-p_{S.})+p_{S.}*rr_{1S}\right)*p_{.T}*rr_{2T}} = 1 - \frac{1}{\left((1-p_{S.})+p_{S.}*rr_{1S}\right)}$

which is the expression for the univariate PAR.

The expression for PAR_P likewise reduces to $1 - \frac{1}{\left((1-P_{s.})+P_{s.}*RR_{1S}\right)}$ and the bias is given by

$bias_{A,rr_{S|T}=0} = PAR_P - par_P = \frac{1}{\left((1-P_{s.})+P_{s.}*RR_{1S}\right)} - \frac{1}{\left((1-p_{S.})+p_{S.}*rr_{1S}\right)}$.

$\frac{\partial bias_{A,rr_{S|T}=0}}{\partial \theta_S} = \frac{(rr_{1S}-(1-p_{S.})\phi_S p_{S.})}{(1-p_{S.}+p_{S.}*rr_{1S})(p_{S.}*\phi_S+p_{S.}*\theta_S-p_{S.}-\phi_S)^2} > 0$ and when $\theta_S = 1$, $bias_{A,rr_{S|T}=0} = 0$.

This indicates that the bias under these conditions increases with θ_S and is hence towards the null for all $\theta_S < 1$ when $rr_{S|T} = 1$. Calculations were done using Maple.

Appendix B

Published Manuscript for Chapter 1

The published form of Chapter 1, as it appears in *Statistics in Medicine*, and online at https://doi.org/10.1002/sim.7559, is enclosed.