Università degli studi di Milano

FACOLTA' DI MEDICINA E CHIRURGIA

Dipartimento di Scienze Cliniche e di Comunità

Laboratorio di Statistica Medica, Biometria ed Epidemiologia "G. A. Maccacaro"

Corso di Dottorato di Ricerca in
Epidemiologia, Ambiente e Sanità Pubblica

Estimates of cancer population attributable fractions for multiple risk factors from a network of Italian case-control studies

Dottorando:	Tutor:
Matteo Di Maso	Chiar.mo Prof.
	Monica Ferraroni

Anno III

XXX Ciclo – A.A. 2016-2017

Index

Abstract

Introduction *Attributable fraction* (AF), proposed by Levin, quantifies the reduction in the disease prevalence that could be achieved by eliminating the exposure (or risk factor) of interest from the population. Disease etiology involves multiple risk factors that may act simultaneously in the occurrence of disease and the optimal approach to quantify the individual and the joint effects of different risk factors on the disease burden is one of the goals in epidemiological research. *Adjusted AFs* quantify the effect of one risk factor after controlling of other factors (i.e., risk factors that may act together to cause disease, adjustment variables or confounders). Adjusted AFs may add up more than the joint AF (i.e., the AF for eliminating all risk factors from the population) and in some situation may add up to more than 1, leading to the conclusion that adjusted AFs should not be used to the purpose of partitioning the joint effect into individual contributions. Eide and Gefeller proposed a way to accomplish this task. *Sequential AFs* quantify the additional effect of one risk factor on the disease risk after the preceding risk factors have already been removed in a specified order from the population. However, sequential AFs depend on the order in which risk factors are removed from the population. *Average AFs* overcome this shortcoming by averaging sequential AFs for a risk factor over all orders by which risk factors can be removed from the population. Average AFs quantify the additional effect of one risk factor on the disease risk after the preceding factors selected randomly have already been removed from the population.

Objective This work aims to illustrate the main methodologies to estimate AFs and corresponding confidence intervals in presence of multiple risk factors with a focus on case-control study design. Moreover, we provide AF estimates for the major risk factors using Italian case-control data on oral cavity and breast cancers.

Modification of case-control study design In the original notation**,** sequential and average AFs could not be used in case-control study design, since the ratio of controls to cases in the sample is fixed a priori and the resulting AF estimates will be biased. Ferguson et al. proposed a prevalence-based weighting approach to correct the imbalance between controls and cases. The method consists in weighting the likelihood function of the model used to estimate sequential and average AFs for the disease prevalence.

Variance estimation The main approaches for estimating AF confidence intervals (CIs) are based on asymptotic approximation (Delta method) and simulations (Monte Carlo method). Ferguson proposed a method based on Monte Carlo simulations for constructing average AF variance. They also proposed the “averisk” R package for calculating average AFs and corresponding CIs in both prospective and case-control studies. In this work, we proposed a modification of the Ferguson’s method to account for sequential AF variability on the total variability.

Variances comparison We compared our and Ferguson’s methods to estimate average AF variance using simulated data. We generate two classes of simulated dataset. Each class included four scenarios according to different correlation structure: from independence (scenario 1) to strong correlation among risk factors (scenario 4). The two classes differed in the prevalence and strength of the association between risk factors. In particular, the first class had a high prevalence and modest relative risks, whereas the second class had a low prevalence and huge relative risks.

For both classes of simulated data, standard deviation increment (i.e., the relative difference between our and Ferguson’s methods) became gradually larger increasing the number of independent risk factors (from two to ten). Conversely, standard deviation increment decreased incrementing the number of correlated risk factors. Although in some situations (i.e., for correlated risk factors) the contribution of our method could have a substantial relative impact on total AF variability (up to 88%), the absolute standard deviation differences between two methods were very small (less than 0.15) indicating a limited contribution of our method than the Feguson’s one.

Application to real data We estimated average AFs using a case-control study conducted in Italy on 946 oral cavity cases and 2492 controls. Risk factors considered for AF estimation were smoking, alcohol drinking, red meat intake, vegetables intake, fruit intake, and family history of oral cavity cancer. The final model included also terms for sex, age, study centre, years of education, BMI, and non-alcohol energy intake to account for possible confounding effect. We set a 81×10^{-5} prevalence of oral cavity cancer according to statistics from the consortium of Italian Cancer Registry (AIRTUM) to adjust average AFs for case-control data structure. Eighty-eight percent (95% CI: 78%; 98%) of oral cavity cases were attributable to the considered risk factors. In particular, the average AF for smoking was

0.34 (95% CI: 0.27; 0.41), indicating that 34% of oral cavity cases would not has occurred if smoking was randomly removed from the population over all possible risk factor removal orders. For the remaining risk factors, average AFs were 0.27 (95% CI: 0.17; 0.37) for alcohol drinking, 0.11 (95% CI: 0.06; 0.17) for low vegetables intake, 0.08 (95% CI: 0.02; 0.15) for low fruit intake, 0.06 (95% CI: 0.01; 0.12) for high red meat intake, and 0.009 (95% CI: -0.001; 0.02) for family history.

We analyzed a further case-control study on 2569 breast cancer cases and 2588 controls. We set a 2019×10^{-5} prevalence of breast cancer to adjust average AFs for case-control data structure. The final model included alcohol drinking, parity, breastfeeding, use of oral contraceptives (OCs), and family history of breast cancer as risk factors; study centre, age, years of education, smoking, age at menarche and use of hormonal replacement therapy (HRT) as adjusting factors. The joint AF was 0.49 (95% CI: 0.35; 0.63) indicating that approximately half of the breast cancer cases would not has occurred if all risk factors were simultaneously eliminated from the population. In particular, average AFs were 0.27 (95% CI: 0.16; 0.39) for parity, 0.12 (95% CI: 0.06; 0.18) for alcohol drinking, 0.04 (95% CI: -0.02; 0.10) for breastfeeding (No or <4 months), 0.04 (95% CI: 0.03; 0.06) for family history of breast cancer, and 0.01 (95% CI: -0.01; 0.03) for OCs users.

Conclusions Sequential and average AFs are useful tools to apportion exposure-specific contributions in a population exposed to multiple risk factors. Sequential and average AFs share some mathematical properties such as component-additivity, symmetry, marginal rationality, and internal marginal rationality. Average AFs, however, do not represent the actual amount of disease ascribable for each risk factors because they assume that risk factors are removed from the population in a random order. Nevertheless, average AFs could be useful parameters to estimate the average burden of disease for each risk factors across all possible removal orders.

In this work, we proposed an alternative approach to estimate the average AF confidence interval accounting for sequential AF variability on the total AF one. We compared the performance between our and Fergusons' methods to estimate AF variance. Although our method could have a relative impact on total AF variability, the absolute standard deviation differences suggest a limited contribution of our method. However, this topic should be further analyzed.

Chapter 1

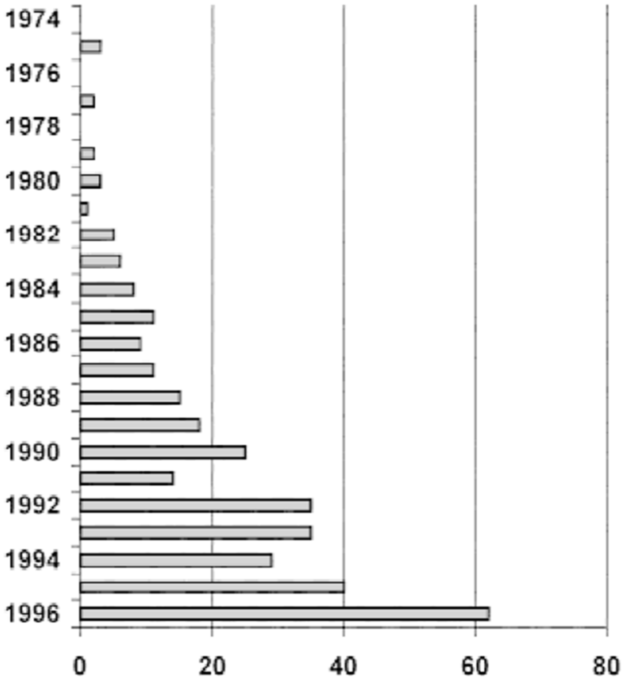
Introduction

1.1 Background

One of the research goals in epidemiology concerns quantifying the impact of exposures (or risk factors) on the outcome, such as mortality or a certain disease, at the population level. In epidemiology, relative risks (RRs) or odds ratios (ORs) measure the strength of the association between risk factors and an outcome. These parameters, however, do not define the impact of risk factors in the population, as their prevalence are not taken into account. The parameter that includes both the strength of association between risk factors and disease and the risk factor prevalence in the population is the *attributable fraction* (AF). The AF measures the proportion of diseased subjects (cases) ascribable to one or more exposures, or, in other words, the percentage reduction in the probability of disease that could hypothetically be achieved by completely eliminating the risk factors from the population. Over the years, the concept of AF has grown in importance because it can be used to plan public health actions for the reduction of the disease. Uter and Pfahlberg [1] reviewed the epidemiological literature related to the use of the AF from 1966 to 1996. No publication was observed in the first ten years, but a considerable increase in numbers was found in the last ten years (Figure 1.1). The mounting literature dealing with AF comes from the wish to quantify the number of the observed cases that can be ascribed to risk factors. Indeed, the AF answers questions like: “how much do risk factors contribute to the burden of morbidity (or mortality) in a population?”, or “how much of disease load could be eliminated if risk factors were eliminated from the population?”.

Initially, AF was formulated for a single dichotomous risk factor [2] and was later extended to polytomous or continuous ones [3-5]. These expressions, however, ignored the presence of other factors (i.e., confounders, effect modifications, or other exposures that may act simultaneously in the disease onset), assuming the AF as a one-dimensional parameter. The AFs, termed also *unadjusted or crude attributable fractions*, as originally defined, lead to biased estimates. Walter [6] first discussed the biases in the AF estimates when the distribution and the effect of other factors are considered.

Figure 1.1. Temporal distribution (from 1966 to 1996) of publications employing the attributable fraction (AF)§.



§Source: Uter W, Pfahlberg A. *The concept of attributable risk in epidemiological practice.* Biometrical J. 1999; 41(8):985-93.

In practice, disease outcomes have multiple contributing determinants that may act together to produce a given instance of disease [7] and may have a remarkable impact on the estimation of AF. Adjustment procedures to estimate AF were the subject of intensive research activities by several authors [4, 6, 8-11]. Stratification and modeling approaches are the main adjustment strategies. Thus, the *adjusted attributable fraction* quantifies the effect of removing one or more risk factors after controlling for other risk factors and possible confounders. Individual adjusted AF, however, do not sum up to their joint AF (i.e., the AF for eliminating all risk factors from the population) and frequently may sum up to more than 1, indicating an unrealistic scenario in which the cases attributable to the risk factors are more than the burden of cases in the population. This depends on both interactions and correlations between risk factors that lead to overlapping contributions to the occurrence of the disease.

In a multifactorial setting, the interest is to apportion the joint effect attributed to a collection of risk factors into individual contributions. The task of dividing the joint effect into exposure-specific components requires methods that do not impose any hierarchy among risk factors. Clearly, adjusted AFs violate this requirement and partitioning approaches have been developed in order to accomplish this. A sequential approach consists in stepwise removal of one risk factor at a time in a pre-specified order. The *sequential attributable fraction* is the AF for eliminating a risk factor in a particular order from the risk system. It quantifies the additional effect of one risk factor on disease risk after the preceding risk factors in a specified sequence have already been taken into account. Thus, the sequential AF depends on the ordering in the set of exposures within the sequence. Indeed, the sequential AF for a specific risk factor may differ even for the same set of risk factors according to different removal orders. One way to avoid this ambiguity is to average sequential AFs over all removal orders, leading to the *average attributable fraction*, also termed *partial attributable fraction* [12]. The average AF quantifies the additional effect due to the elimination of a risk factor that can be expected after the effect of other risk factors (randomly selected) have already been taken into account. The concept of averaging sequential AFs is rooted in the principle suggested by Kruskal [13] for determining relative importance for independent variables in a multiple regression setting. Also Cox Jr. [14] gave a justification for using the average AF in a way based on the analogy between game-theoretical reasoning in profit allocation among several players that acting together in a coalition and the epidemiological task of apportioning disease risk among multiple risk factors.

1.2 Aim and description of the study

This works aims to illustrate the main methodologies to estimate the attributable fractions and corresponding confidence intervals in presence of multiple risk factors with a focus on case-control study design. Moreover, we provide AF estimates using Italian case-control data on oral cavity and breast cancers.

The rest of this work is organized as follows. In the chapter 2, we will review the concept of AF including adjusted strategies to control for confounders or effect modifications and their shortcomings in the presence of multiple risk factors. Moreover, sequential and average AFs will be illustrated and their mathematical properties as well. Then, we will describe an approach to estimate sequential and average AFs accounting for case-control data structure based on incorporating disease prevalence in the regression model (chapter 3). We will illustrate the delta [15] and Monte Carlo methods [16] to estimate AF confidence intervals. Moreover, we will propose a modification of the method based on Monte Carlo simulations to estimate average AF confidence intervals (chapter 4). Performance between our and existing methods will be compared through simulated datasets (chapter 5). Using case-control data, we will provide average AF estimates for the major risk factors for oral cavity and breast cancers (chapter 6). Chapter 7 will discuss the results, some methodological issues, and future developments.

Chapter 2

Review of the literature

2.1 Definition of population attributable fraction (AF)

The classical definition of attributable fraction (AF) was proposed by Levin [2]. Levin's interest was in quantifying the proportion of lung cancer cases in the population that could be ascribed to smoking and he introduced an "index S which is the indicative maximum proportion of lung cancer attributable to smoking". This parameter was formulated as follows:

$$AF = \frac{(RR-1)\cdot P(E)}{1+(RR-1)\cdot P(E)},$$

where $P(E)$ is the probability of risk factor E in the population (or prevalence) and $RR = P(D\,|\,E)\big/P(D\,|\,\bar{E})$ is the relative risk. In the Levin's formula, the denominator represents the disease probability in the overall population, whereas the numerator represents the difference between the disease probability in the overall population and the probability that would be observed if the whole population were unexposed. Indeed, the AF can also be formulated in terms of these probabilities (MacMahon and Pugh's formula) [17]:

$$AF = \frac{P(D) - P(D\,|\,\bar{E})}{P(D)},$$

where $P(D)$ denotes the probability of disease in the overall population and $P(D\,|\,\bar{E})$ denotes the probability of disease in the unexposed subjects. Leviton [18] proved the algebraic equivalence of these formulas (see Appendix 1).

Miettinen [3] proposed the attributable fraction in exposed (AFE) restricting the attention to the group of exposed subjects. Levin's and MacMahon and Pugh's formulas, respectively, can be written as:

$$AFE = 1 - RR^{-1},$$

and

$$AFE = \frac{P(D|E) - P(D|\bar{E})}{P(D|E)}.$$

The AFE quantifies the proportion of disease attributable to exposure solely in the group of exposed. The AF is related to the AFE by the equation:

$$AF = AFE \cdot P(E|D)$$

where $P(E|D)$ is the probability of exposure among subjects who developed the disease. Miettinen's formula is also called case-based version of the attributable fraction because it is based on the distribution of exposure in cases.

For the AF, Walter [4, 19, 20] developed asymptotic distributions for the maximum likelihood estimator in cross-sectional and cohort studies providing approximate standard errors. As AFE is a transformation of relative risk, standard errors are easily obtaining by transforming it from the relative risk scale. AF may also be estimated from case-control data of a rare disease where cases are representative of the diseased population [21]. Several authors, however, proposed alternative approaches for estimating AF for case-control data of common disease (i.e., high disease prevalence) [10, 16, 22-24].

2.2 Generalization of AF: adjusted AF

The AF and the AFE are univariate parameters that describe the fraction of cases attributable to the elimination of one risk factor. Thus, they are considered crude, unadjusted, or marginal parameters. These unadjusted parameters are in general biased. Walter [6] worked out the conditions under which adjusted AF estimates that account for the distribution and the effect of other factors (i.e., risk factors that may act together to cause disease, adjustment variables or confounders) differed from unadjusted AF estimates. He showed that, if E_1 and E_2 are two dichotomous risk factors and if the interest is in the estimating AF for risk factor E_1, the crude and adjusted AFs will coincide if and only if at least one of the following conditions is true:

i. E_1 and E_2 are independently distributed in the population, that is:
$P(E_1=0,E_2=0)\cdot P(E_1=1,E_2=1)=P(E_1=0,E_2=1)\cdot P(E_1=1,E_2=0)$, where level 0 denotes the unexposed group and level 1 denotes the exposed group;

ii. exposure to E_2 alone does not increase disease risk, that is:

$$P(D\,|\,E_1=0,E_2=1)=P(D\,|\,E_1=0,E_2=0).$$

Considering a polytomous risk factor or more than one adjustment variable, the previous conditions can be extended to a set of sufficient conditions [11].

Stratification is one of the approaches proposed for quantifying attributable shares of the disease probabilities accounting for other risk factors and possible confounders or effect modifications [4, 6, 8, 9, 25]. In particular, three adjustment strategies were introduced:

i. the weighted-sum over all adjustment strata of the stratum-specific AFs (type I);
ii. the use of adjusted relative risk to obtain adjusted AF, exploiting the functional relationship between relative risk and AF (type II);
iii. the extension to the AF parameters of the factorization idea [26] used to adjust relative risk (type III).

Consider a simple 2×2 table, where rows report unexposed and exposed subjects and columns report diseased and non-diseased subjects.

Table 2.1. Two by two table summarizing data of a disease D and a risk factor E for the kth stratum of the adjustment factors.

Exposure level	**Disease**		**Total**
	$\boldsymbol{D}$	$\boldsymbol{\bar{D}}$	
$\boldsymbol{E}$	a_k	b_k	a_k+b_k
$\boldsymbol{\bar{E}}$	c_k	d_k	c_k+d_k
Total	a_k+c_k	b_k+d_k	N_k

The adjustment procedure leads to a $2\times2\times K$ table, where the third dimension is represented by a stratum variable C with K levels, which could be one observed variable or a combination of two or more variables. Formally, the observed data on a disease and a risk factor for the kth stratum of the adjustment factors is reported in table 2.1. The quantities N, a, b, c and d denote the sum of the stratum-specific values over all K strata, respectively.

Type I adjustment strategy

The type I adjustment strategy has the following expression:

$$_{adj}AF_E = \sum_{k=1}^{K} w_k \cdot AF_{E,k} \text{ ,}$$

where $AF_{E,k}$ denotes the AF for the risk factor E in the kth stratum and w_k, $k = 1, 2, \ldots, K$, is a set of weights with $\sum_{k=1}^{K} w_k = 1$. Any set of weights can theoretically be used, but only the following sets have been proposed:

$$w_k = \frac{a_k + c_k}{a + c} \qquad k = 1, 2, \ldots, K$$

and

$$w_k = \frac{\left[Var\left(AF_{E,k}\right)\right]^{-1}}{\sum_{k=1}^{K}\left[Var\left(AF_{E,k}\right)\right]^{-1}} \qquad k = 1, 2, \ldots, K \text{ .}$$

In the first expression, the weights are the proportions of cases in the strata among all cases in the sample [8]. This is called “case-load weighting”. The second expression, termed “precision-weighting”, employs the inverse stratum-specific variance of the AF over the sum of inverse variances over all strata in order to obtain increased precision.

Type II adjustment strategy

Miettinen discussed first the relationship between AF and relative risk [3]. Exploiting this relationship, the type II adjustment strategy consists in plugging-in an estimate of the common relative risk (or odds ratio in case-control studies) in the AF expression in order to obtain adjusted AF estimates. Several choices are available for the adjusted relative risk estimator. Kleinbaum et al [27] suggested using the Mantel-Heanszel estimator for the relative risk [28] in cross sectional and cohort studies (or the Mantel-Heanszel estimator for the odds ratio in case-control studies). The general formula of the Mantel-Heanszel estimator for the relative risk is:

$$RR = \frac{\sum_{k=1}^{K} \frac{a_k \cdot (c_k + d_k)}{N_k}}{\sum_{k=1}^{K} \frac{c_k \cdot (a_k + b_k)}{N_k}},$$

or the alternative formula:

$$RR = \frac{\sum_{k=1}^{K} \frac{a_k \cdot (c_k + d_k)}{b_k + d_k}}{\sum_{k=1}^{K} \frac{c_k \cdot (a_k + b_k)}{b_k + d_k}}.$$

Other authors proposed different estimators, such as the “precision weighting” of the stratum-specific log relative risks [9, 25], or weighted least-square estimators of a common relative risk [29]. Moreover, the type I adjustment strategy using the case-load weighting set equals to the type II adjustment strategy with the Mantel-Heanszel estimator [21]. For more details, see A.2.

Type III adjustment strategy

Walter [4] proposed an approach by adapting Miettinen’s factorization idea [26] to adjust relative risk to the context of the AF. This approach assumes that the crude AF (${}_{crude}AF$) is the sum of two parts: the component due to the effect of confounding (${}_{conf}AF$) and the adjusted component (${}_{adj}AF$). The ${}_{conf}AF$ is expressed by:

$${}_{conf}AF = \frac{(a \cdot d) - c \cdot \sum_{k=1}^{K} \frac{a_k \cdot d_k}{c_k}}{(a+c) \cdot (c+d)}.$$

Walter derived the previous formula by calculating the hypothetical number, $\tilde{b}$, of all exposed non-diseased subjects that would have occurred if in all strata the risk factor had no effect on the disease probabilities (i.e., $RR_k = 1$ for all $k = 1, 2, \ldots, K$) and all other cell entries were unchanged. Thus, he computed a value $\tilde{b}_k = a_k \cdot d_k / c_k$ for each stratum of C and summed it over all strata of C to yield $\tilde{b}$. Taking this hypothetical cell entry $\tilde{b}$ and the observed cell entries a,c and d, he estimated the AF which would be observed if there were no exposure effect on each stratum of C. Then, he yielded the adjusted AF according to the following formula:

$$_{adj}AF = {}_{crude}AF - {}_{conf}AF = \frac{c \cdot \sum_{k=1}^{K} \left(\frac{a_k \cdot d_k}{c_k} - b_k \right)}{(a+c) \cdot (c+d)} .$$

When there is no interaction between factors ("homogeneity model"), the type II adjustment strategy is preferable as it produces unbiased estimation even for sparse data [30, 31]. When interaction exists ("interaction model"), there is no common relative risk to be estimated and the type II approach becomes inconsistent [9, 32]. Moreover, the type I adjustment strategy assumes that stratum-specific AFs are constant across all strata, which is achieved in some situations of the interaction model and it is incompatible with the homogeneity model. Indeed, as a consequence of the homogeneity model at least two stratum-specific exposure prevalence $P(E \mid C = k)$ and $P(E \mid C = h)$ with $k \neq h$, have to be different, which results in non-identical stratum-specific AFs (i.e., $AF_{E,k} \neq AF_{E,h}$) [9, 11, 25]. The adjusted AF using type III strategy was criticized by Ejigou who pointed out the heuristic argumentation proposed by Walter to determine $_{conf}AF$. Simulations have shown that the adjusted AF estimated with type III strategy, while intuitively appealing, was inconsistent exhibiting very severe bias [33].

Modeling approach

Another class of adjustment strategies based on a modeling approach exploits the generality and flexibility of a regression model. The type I adjustment strategy does not impose any structure on the relative risk and its variation with risk factors and strata of adjustment factors. It simply requires estimating separate relative risks for each stratum of the adjustment factors. The type II strategy requires estimating a common adjusted relative risk, which corresponds to a regression model with only the main effects for risk factors and possible confounders. The regression model, instead, allows for taking into account adjustment factors as well as interaction terms (e.g., an interaction between two risk factors or an interaction between a risk factor and an adjustment factor). Some authors proposed the use of regression models for estimating adjusted AF [4, 34, 35]. Model-based adjusted AFs have been developed for case-control [10], cross-sectional [36], and cohort studies [11]. Bruzzi and colleagues [10] applied logistic regression models with respect to case-control studies. This approach is valid for cohort and cross-sectional studies. The method consists in a weighted sum of relative risk (or odds ratios in case-control studies) to estimate the adjusted AFs. For each stratum of adjustment factors, relative risk estimates are combined with the stratum-specific proportion of diseased subjects. The Bruzzi's formula is:

$$ {}_{adj}AF_E = 1 - \sum_{k=1}^{K} \sum_{q=0}^{Q} \frac{\rho_{q,k}}{RR_{q|k}} . $$

The first sum is taken over all strata of adjustment factors, and the second sum is taken over all exposure levels, assuming that the risk factor of interest E presents $Q+1$ levels, $q = 0,1,\ldots,Q$ (usually one unexposed level and Q exposed levels). The quantity $\rho_{q,k}$ represents the proportion of cases with respect to the qth exposure level and kth adjustment stratum, while $RR_{q|k}$ represents the relative risk for the qth exposure level given the kth adjustment stratum, that is:

$$ RR_{q|k} = \frac{P\left(D \mid E = q, C = k\right)}{P\left(D \mid E = 0, C = k\right)}, $$

where $E = 0$ indicates the unexposed level.

This model-based approach includes the crude and other adjusted approaches as special cases but offers additional options. The unadjusted approach corresponds to univariate models or models without confounders. The type II adjustment strategy corresponds to the model with risk factors of interest and adjustment factors, without any interaction term. The type I approach corresponds to the model with all interactions terms (saturated model). The model-based approach also allows for intermediate models, for instance, models with interaction between risk factors and only one adjustment factor. Benichou gave a comprehensive discussion including examples of the generality and flexibility of model-based adjustment strategy [11, 21].

2.3 Multifactorial setting

Questions addressing the joint effect of multiple risk factors or the effect of a single risk factor relative to the combination of others factors cannot be overcome by simply sum up the adjusted AFs. When several exposures act simultaneously in developing of disease, the sum of the individual AFs usually exceeds the joint AF (i.e., the proportion of disease that can be attributed by eliminating all risk factors from the population). In other cases, the sum of adjusted AFs might be more than 1 leading to the conclusion that adjusted AFs cannot be used for the purpose of partitioning the joint fraction into individual contributions [25, 37-39]. This is a consequence of both interaction between risk factors defined on an additive level scale for relative risk, and non-independence of the risk factors that leads to overlapping contributions to the occurrence of the disease.

In a multifactorial setting, it is particularly interesting to assess single risk factors with respect to their specific contributions to the joint effect of all risk factors and compare them to each other in order to identify those risk factors have a pronounced impact on the disease load in the population. Thus, another conceptual approach to the estimation of the AF is required.

2.4 Sequential AF

A method for calculating the contributions of individual risk factors to the disease risk in a multifactorial contest was developed by Eide and Gefeller. They introduced the sequential AF, “a proposal of the optimal preventive strategy for the elimination of risk factors with respect to the greatest impact for a given number of risk factors to be eliminated” [12].

Let $\mathbf{E} = \{E_1, E_2, \ldots, E_L\}$ be a set of L risk factors of interest and $\mathbf{C} = \{C_1, C_2, \ldots, C_M\}$ be another set of M adjusting factors. All combinations of values for the adjustment factors define a total of K strata, $k = 1, 2, \ldots, K$. Furthermore, all combinations of values for the risk factors define $Q+1$ exposure levels, $q = 0, 1, \ldots, Q$. The unexposed subjects, or subjects exposed to the lowest exposure level (previously denoted by $\bar{E}$ or by $E = 0$) is indicated by $q = 0$. The $Q+1$ exposure levels are generated by the

L risk factors according to $Q+1=\prod_{l=1}^{L} Q_l +1$. The interest lies in the potential reduction of disease load when eliminating the L risk factors, one at time, in a given sequence. For instance, a possible order for eliminating risk factors is as follows: starting with risk factor E_1, then risk factor E_2, and so on, until all L risk factors are eliminated from the population. A way to implement this task is to calculate the adjusted AF for risk factor E_1 considering all risk factors of interest and adjusting factors. This results in an adjusted AF, denoted by ${}_{adj}AF_1$ derived from $K \cdot \prod_{l=2}^{L}(Q_l+1)$ adjustment strata and Q_1+1 exposure levels. Thereafter, define ${}_{adj}AF_{12}$ as the adjusted AF calculated for the combined effect of first and second risk factor (generating $(Q_1+1)\cdot(Q_2+1)$ exposure levels), and the remaining risk factors and the adjusting factors forming the $K \cdot \prod_{l=3}^{L}(Q_l+1)$ strata. This stepwise procedure of calculating adjusted AF for different sets of risk factors can be continued until all L risk factors are incorporated among the exposure levels. The last term in this sequence ${}_{adj}AF_{12\ldots L}$ corresponds to the joint AF, i.e., the total population impact of all L risk factors controlled for the effect of the M adjustment factors.

Any difference ${}_{adj}AF_l - {}_{adj}AF_{l'}$ quantifies the additional effect of considering the $(l'+1)\text{st}, (l'+1)\text{nd}, \ldots, l\text{th}$ risk factors after having previously taken into account the effect of the first l' risk factors in the specified sequence. These differences are called sequential AFs. For instance, consider the following sequential AF: ${}_{sequential}AF_{2|1} = {}_{adj}AF_{12} - {}_{adj}AF_1$, which represents the proportion of disease that could be avoided by eliminating the risk factor E_2 after the risk factor E_1 has already removed from the population.

The sequential AF of a specific risk factor may differ even for the same set of L risk factors depending on the sequence considered during the stepwise process of elimination. The L risk factors lead to $L!$ different sequences of removal orders. For instance, consider the simple set of $L=2$ risk factors, $\mathbf{E}=\{E_1, E_2\}$, leading to $2!=2$ possible removal orders. For each risk factor, the corresponding sequential AFs will differ according to the removal order (i.e., if the risk factor is the first or the second to be eliminated). A way to choose a strategy of eliminating risk factors among all possible removal orders is to eliminate first the exposure among all L risk factors which gives the highest attributable fraction (e.g., ${}_{adj}AF_l$), and next, to remove the exposure among the remaining $L-1$ which, combined with the first one, leads to the highest attributable fraction (e.g., ${}_{adj}AF_{ll'}$), and so forth, until all risk factors are removed.

2.5 Average AF

The sequential AFs do not yield a unique value for a particular risk factor and in particular, there will be $L!$ sequential AFs for each risk factor. Thus, the consistent quantification problem of the contribution of one risk factor to the disease load in a population exposed to multiple risk factors remains unsolved. The average AF could mitigate this problem by averaging the differing sequential AFs over all possible orders by which a risk factor could be eliminated from the population. Clearly, the average AFs do not depend on the removal orders anymore. The average AF [12] for the lth risk factor is formally defined as:

$$ {}_{average}AF_l = \frac{\sum_{\gamma \in G_L} {}_{sequential}AF_{l|\gamma}}{L!}, $$

where γ is the generic permutation of the removal orders in which a risk factor can be eliminated and G_L is the set of all possible permutations.

For a given risk factor, the average AF can be interpreted as the expected proportion of preventable cases by the additional elimination of the risk factor considered after having already removed a random collection (independently from the order) of other risk factors. Finally, among the $L!$ different removal orders, some yield an identical value of the sequential AF for the same risk factor (see section 3.2 for a more details).

2.6 Properties of sequential and average AFs

Sequential and average AFs share some mathematical properties of particular interest, as outlined below.

Component-additivity

Sequential and average AFs subdivide the total burden of disease due to all risk factors into individual contributions. As result, sequential and average AFs satisfy the nice property that individual contributors sum up to the joint AF.

Symmetry

When assessing those fractions of disease burden in the population that can be attributed to risk factors of interest, it should be mandatory that the method used for dividing the joint AF is not influenced by the number of risk factors considered or by any ordering among them. A method that satisfy this condition is called symmetric. It is clear that sequential AFs are not symmetric, whereas average AFs are symmetric [40].

Marginal rationality

Let $\mathbf{E} = \{E_1, \ldots, E_L\}$ be a set of L risk factors and let a and b be two separate subpopulations. For instance, subpopulations might be defined with respect to the categories of an external variable like sex or race. The marginal rationality is defined as [41]:

$$AF_{12\ldots L}^{a} - AF_{12\ldots L\setminus\{l\}}^{a} \geq AF_{12\ldots L}^{b} - AF_{12\ldots L\setminus\{l\}}^{b} \text{ for all } L.$$

where $AF_{12,\ldots L}$ and $AF_{12,\ldots L\setminus\{l\}}$ are the joint AF and the AF due to the L-1 risk factors (excluding the lth), respectively.

Then for all permutations G_L of the L risk factors the following expressions are valid:

$$_{sequential}AF_{l,G_L}^{a} \geq {}_{sequential}AF_{l,G_L}^{b},$$

and

$$_{average}AF_l^a \geq {}_{average}AF_l^b \; .$$

The sequential and average AFs fulfill this condition. The marginal rationality means that whenever the effect of eliminating the lth risk factor in subpopulation a is higher than the effect in subpopulation b, this ranking among subpopulations is valid independently from the choice of risk factors that have been eliminated before. Thus, the lth risk factor has a more pronounced impact on disease load in subpopulation a than subpopulation b. Marginal rationality ensures a consistent comparison of the population impact of one risk factor with respect to separate subpopulations.

Internal marginal rationality

While the marginal rationality ensures a reasonable ranking of the same risk factor when its population impact in different strata of population is compared, the internal marginal rationality deals with the comparison of different risk factors concerning their respective impact on the disease load in one population [41]. Suppose that the following expression is valid for the lth and the l'th risk factors:

$$AF_{12\ldots L\backslash\{l\}} \geq AF_{12\ldots L\backslash\{l'\}} \text{ for all } L.$$

The partitioning procedure used to create average AF has the property of internal marginal rationality, which implies:

$$_{average}AF_l \geq {}_{average}AF_{l'} \, .$$

Internal marginal rationality means that whenever the lth risk factor contributes more to the joint AF than the l'th risk factor, this ranking among the risk factors is valid independently from the choice of risk factors that have been eliminated before. While average AF fulfill this condition, the sequential AF do not. Internal marginal rationality thus ensures a consistent comparison of the respective population impact of different risk factors to each other.

2.7 Sequential and average AFs in a group of multi-exposed subjects

The principle of sequential and average AFs outlined for a population that comprises unexposed and exposed subjects was extended by Eide and Heuch to a subpopulation of exposed subjects, i.e., the group of multi-exposed subjects [42].

Two dichotomous risk factors

Recalling the formula of AF in exposed (AFE) showed in section 2.1 for a single dichotomous risk factor E:

$$AFE = \frac{P(D \mid E) - P(D \mid \bar{E})}{P(D \mid E)},$$

and considering the simple case of two dichotomous risk factors, E_1 and E_2, various AF in the exposed (AFEs) can be defined corresponding to various choices of exposed group:

$$_{crude}AFE_1 = \frac{P(D \mid E_1 \cap \bar{E}_2) - P(D \mid \bar{E}_1 \cap \bar{E}_2)}{P(D \mid E_1 \cap \bar{E}_2)},$$

$$_{crude}AFE_2 = \frac{P(D \mid \bar{E}_1 \cap E_2) - P(D \mid \bar{E}_1 \cap \bar{E}_2)}{P(D \mid \bar{E}_1 \cap E_2)},$$

$$AFE_{12} = \frac{P(D \mid E_1 \cap E_2) - P(D \mid \bar{E}_1 \cap \bar{E}_2)}{P(D \mid E_1 \cap E_2)}.$$

The problem is how to calculate the contribution of each risk factor to the joint AF in those simultaneously exposed to E_1 and E_2, i.e., AFE_{12}. According to the procedure developed by Eide and Gefeller [12], the starting point is to determine adjusted AFEs to both risk factors and then the sequential AFs in these double-exposed subjects.

The adjusted AFs are defined as:

$$_{adj}AFE_1 = \frac{P(D \mid E_1 \cap E_2) - P(D \mid \bar{E}_1 \cap E_2)}{P(D \mid E_1 \cap E_2)},$$

and

$$_{adj}AFE_2 = \frac{P\left(D \mid E_1 \cap E_2\right) - P\left(D \mid E_1 \cap \bar{E}_2\right)}{P\left(D \mid E_1 \cap E_2\right)}.$$

The first expression is the AF of disease due to risk factor E_1 adjusted for risk factor E_2 in those people who are exposed to both risk factors. Likewise, the second expression is the AF of disease due to E_2 adjusted for E_1 in those simultaneously exposed to E_1 and E_2. The sequential AFs are defined as:

$$_{sequential}AFE_{1|\varnothing} = {}_{adj}AFE_1,$$

where the subscript $1|\varnothing$ indicates that risk factors E_1 is removed first (given no other risk factor have been already removed) from the subpopulation of exposed (for details see section 4.2);

$$\begin{aligned} _{sequential}AFE_{2|1} &= AFE_{12} - {}_{adj}AFE_1 = AFE_{12} - {}_{sequential}AFE_{1|\varnothing} \\ &= \frac{P\left(D \mid \bar{E}_1 \cap E_2\right) - P\left(D \mid \bar{E}_1 \cap \bar{E}_2\right)}{P\left(D \mid E_1 \cap E_2\right)} \end{aligned}$$

where the subscript $2|1$ indicates that E_2 is eliminated after having already removed E_1 from the subpopulation of exposed.

Similarly, for the reverse order:

$$_{sequential}AFE_{2|\varnothing} = {}_{adj}AFE_2,$$

$$\begin{aligned} _{sequential}AFE_{1|2} &= AFE_{12} - {}_{adj}AFE_2 = AFE_{12} - {}_{sequential}AFE_{2|\varnothing} \\ &= \frac{P\left(D \mid E_1 \cap \bar{E}_2\right) - P\left(D \mid \bar{E}_1 \cap \bar{E}_2\right)}{P\left(D \mid E_1 \cap E_2\right)}. \end{aligned}$$

Finally, the average AFs is given by:

$$_{average}AFE_1 = \frac{_{sequential}AFE_{1|\varnothing} + {}_{sequential}AFE_{1|2}}{2},$$

$$_{average}AFE_2 = \frac{_{sequential}AFE_{2|\varnothing} + {}_{sequential}AFE_{2|1}}{2}.$$

By construction, sequential and average AFEs, i.e., the fraction of diseased subjects belonging to the subpopulation exposed to both risk factors, sum to the joint AFE, i.e., the fraction of diseased subjects who are exposed to both risk factors (component-additivity).

Multiple risk factors

The method described above for two dichotomous risk factors can be easily generalized to the situation with any number of risk factors having any categories.

Consider a set of L risk factors with $Q+1$ exposure levels, $q=0,1,\ldots,Q$. Each risk factor has Q_l+1 categories that generate the $Q+1$ exposure levels, i.e., $Q+1=\prod_{l=1}^{L}\left(Q_l+1\right)$. The unexposed group is indicated by $e_0=\{E_1=0\}\cap\{E_2=0\}\cap\ldots\cap\{E_L=0\}$, whereas the group of subjects exposed to the highest level for all risk factors is indicated by $e_Q=\{E_1=Q_1\}\cap\{E_2=Q_2\}\cap\ldots\cap\{E_L=Q_L\}$. Thus, each exposure level is defined by a unique set of values for the L risk factors: $e_q=\{E_1=q_1\}\cap\{E_2=q_2\}\cap\ldots\cap\{E_L=q_L\}$. Cox Jr. [43] defined this as the risk profile for the qth exposure level: $(q_1,q_2,\ldots q_L)$.

The joint AF in the exposure level q is defined as:

$$AFE_{e_q}=\frac{P\left(D\,|\,e_q\right)-P\left(D\,|\,e_0\right)}{P\left(D\,|\,e_q\right)} \text{ for } q=0,1,\ldots,Q.$$

The task is to describe the contribution of the lth risk factor to joint AFE in e_q in accordance with the average AFs principle, i.e., to define the average AF in e_q for the lth risk factor, $AFE_{e_q,l}$. As usual, the L risk factors can be ordered in $L!$ ways. For the γth order, the AFE in e_q for the $\tilde{l}$ first risk factors adjusted for the $L-\tilde{l}$ last risk factors is:

$$AFE_{e_q,l\,|\,\gamma} = \frac{P\left(D|\bigcap_{l=1}^{L}\{E_{l\,|\,\gamma}=e_{l\,|\,\gamma}\}\right) - P\left(D|\bigcap_{l=1}^{\tilde{l}}\{E_{l\,|\,\gamma}=0\}\bigcap_{l=\tilde{l}+1}^{L}\{E_{l\,|\,\gamma}=e_{l\,|\,\gamma}\}\right)}{P\left(D|\bigcap_{l=1}^{L}\{E_{l\,|\,\gamma}=e_{l\,|\,\gamma}\}\right)} =$$

$$=1-\frac{P\left(D|\bigcap_{l=1}^{\tilde{l}}\{E_{l\,|\,\gamma}=0\}\bigcap_{l=\tilde{l}+1}^{L}\{E_{l\,|\,\gamma}=e_{l\,|\,\gamma}\}\right)}{P\left(D|\bigcap_{l=1}^{L}\{E_{l\,|\,\gamma}=e_{l\,|\,\gamma}\}\right)}.$$

Here, $\left(q_{\gamma_1}, q_{\gamma_2}, \ldots, q_{\gamma_L}\right)$ are the permuted values corresponding to the γth order of the risk profile's $(q_1, q_2, \ldots, q_L)$ defining e_q. The sequential AFE in e_q for the lth risk factor in the γth order is:

$$_{sequential}AFE_{e_q,l\,|\,\gamma} = AFE_{e_q,l\,|\,\gamma} - AFE_{e_q,(l-1)\,|\,\gamma},$$

with $l=1,2,\ldots,L$ and $\gamma=1,2,\ldots,L!$

Finally, the average AFE in e_q for the lth risk factor is the average of the sequential AFs in e_q over all orders:

$$_{average}AFE_{e_q,l} = \frac{1}{L!}\sum_{\gamma\in G_L} {}_{sequential}AFE_{e_q,l\,|\,\gamma}.$$

By construction, the average AFEs in e_q sum over l to the joint AFE in e_q.

2.8 Bridge with game-theory

The epidemiological problem of apportioning the joint AF among multiple exposures could be formalized in a way that is equivalent to the following economics problem. Assume that several players of different companies intend to divide the profit among them by acting together in a coalition. In their work, von Neumann and Morgenstern gave a solution to this economic problem [44]. They approached the problem of analyzing the complex structures of strategic and economic behavior by reducing cooperative L-

person games to a numerical description in terms of characteristic functions. These functions are defined as:

$$\upsilon:\left\{\mathbf{C}\,|\,\mathbf{C}\subset\left\{P_1,P_2,\dots P_L\right\}\right\}\to\mathbb{R}\ ,$$

where $P_1,P_2,\dots,P_L$ denote the players acting in the coalition $\mathbf{C}$.

Table 2.2 gives a summary of the correspondences between the game-theoretic and epidemiologic formalisms. The term $\upsilon(\mathbf{C})$ can be interpreted as the minimal (or expected) profit of the coalition $\mathbf{C}$. It corresponds to the joint risk $r(\mathbf{E})$ attributable to all risk factors that are included in a risk system. The space Γ of all L–person games corresponds to the system of all risk functions Θ (table 2.2).

Cox Jr. [14, 43] adapted the principles of game theory of profit allocation to epidemiological task of assessing the population impact of each risk factor in the context of several risk factors that jointly affect a disease.

Table 2.2. Correspondence between game-theoretic and epidemiologic formalisms.

Game theory	**Epidemiology**
Players: $P_1,P_2\dots,P_L$	Risk factors: $E_1,E_2\dots,E_L$
Coalition: $\mathbf{C}\subset\left\{P_1,P_2,\dots,P_L\right\}$	Set of risk factors: $\mathbf{E}\subset\left\{E_1,E_2\dots,E_L\right\}$
Cooperative L -person game: $\upsilon:\left\{\mathbf{C}\,|\,\mathbf{C}\subset\left\{P_1,P_2\dots,P_L\right\}\right\}\to\mathbb{R}$	Risk function: $r:\left\{\mathbf{E}\,|\,\mathbf{E}\subset\left\{E_1,E_2\dots,E_L\right\}\right\}\to\mathbb{R}$
Method of payoff allocation: $\Phi:\Gamma\to\mathbb{R}^n$ satisfying $\sum_{l=1}^{L}\Phi(\upsilon)=\upsilon\left(P_1,P_2\dots,P_L\right)$	Risk allocation function: $\Phi:\Theta\to\mathbb{R}^n$ satisfying $\sum_{l=1}^{L}\Phi(r)=r\left(E_1,E_2\dots,E_L\right)$
Worth or profit of the coalition: $\upsilon(\mathbf{C})$	Joint risk of all risk factors included in the set $\mathbf{E}$: $r(\mathbf{E})$

In the game theory the two fundamental procedures for dividing up the total profit that several players gain by acting together in coalition are the Shapley-solution [45], also termed *standard solution*, and the solution termed *proportional division*. In the game theory, these approaches are considered to be the two fundamental models of fair allocation of profits. In epidemiology, the standard solution corresponds to the average AF suggested by Eide and Gefeller [12], and the proportional division corresponds to the solution of McElduff et al. [46] for AFE, which was extended to AF by Llorca and Delgado-Rodrìguez [47].

Chapter 3

Estimation based on case-control design

3.1 AF expression using predicted cases

As described in previous chapters, our interest lies in estimating the AF for one risk factor, for example E_1, in a population exposed to L risk factors, as the proportional decrease in disease prevalence if E_1 was eliminated, but the distribution of other risk factors and confounders remained unchanged. The unadjusted AF cannot be interpreted in this way; instead, it summarizes the effects of both other relevant risk factors and confounders potentially overstating the effect of E_1. Obviously, the adjusted AF that takes in account for other risk factors and potential confounders is necessary. One of the adjustment strategies for estimating AFs adopts a modeling approach. It consists in fitting a regression model for disease occurrence with all variables of interest (risk factors and/or confounders). The model is used to predict the total number of cases that would have been observed in the sample under the scenario that no individual had the risk factor of interest, but with the levels of all other factors (other relevant risk factors and/or confounders) left unchanged.

Adjusted AF

In the general formula, the adjusted AF can be expressed in terms of predicted cases by:

$$_{adj}AF = \frac{\left(N_{cases} - \hat{N}\right)}{N_{cases}},$$

where $\hat{N}$ denotes the predicted cases from the model that would have been observed if no sample subject had been exposed to the risk factor (or risk factors) of interest and N_{cases} denotes the observed cases in the sample.

Sequential AF

The previous notation can be easily extended to sequential AFs. Consider two risk factors E_1 and E_2. Suppose that the interest is the proportion of disease attributable to eliminating the risk factor E_1 after the risk factor E_2 has already been removed from the population. The sequential AF for E_1, after the removal of E_2 is represented by the difference:

$$_{sequential}AF_{1|2} = AF_{12} - {}_{adj}AF_2 = AF_{12} - {}_{sequential}AF_{2|\varnothing},$$

where AF_{12} is the fraction of disease ascribable to both E_1 and E_2 (i.e., the joint AF for both risk factors) and $_{adj}AF_2$ is the fraction of disease ascribable to E_2 adjusting for E_1 (i.e., the adjusted AF for E_2). The expression above can be expressed as:

$$_{sequential}AF_{1|2} = AF_{12} - {}_{adj}AF_2 = \frac{\left(N_{cases} - \hat{N}_{12}\right)}{N_{cases}} - \frac{\left(N_{cases} - \hat{N}_2\right)}{N_{cases}} = \frac{\left(\hat{N}_2 - \hat{N}_{12}\right)}{N_{cases}},$$

where $\hat{N}_2$ and $\hat{N}_{12}$ represent, respectively, the predicted number of cases that would have been observed when E_2 and both risk factors were eliminated from the population. Clearly, since the predicted probabilities are calculated from a regression model that include both risk factors E_1 and E_2, $_{adj}AF_2$ is the attributable fraction for risk factor E_2 adjusted for risk factor E_1.

This formula can be extended to any number of risk factors. Suppose that there are L risk factors of interest, $E_1, E_2, \ldots, E_L$; then the sequential AF for the risk factor l, with $2 \le l \le L$, after the removal of risk factors $1, 2, \ldots, (l-1)$ is:

$$_{sequential}AF_{l|12\ldots(l-1)} = AF_{12\ldots l} - {}_{adj}AF_{12\ldots(l-1)} = \frac{\left(N_{cases} - \hat{N}_{12\ldots l}\right)}{N_{cases}} - \frac{\left(N_{cases} - \hat{N}_{12\ldots(l-1)}\right)}{N_{cases}} = \\ = \frac{\left(\hat{N}_{12\ldots(l-1)} - \hat{N}_{12\ldots l}\right)}{N_{cases}},$$

where $AF_{12\ldots l}$ and $_{adj}AF_{12\ldots(l-1)}$ denote the joint and the adjusted AFs for risk factors $E_1, E_2, \ldots, E_l$, and $E_1, E_2, \ldots, E_{(l-1)}$, respectively. Since the regression model includes all risk factors of interest $E_1, E_2, \ldots, E_L$, and adjustment factors, the generic ${}_{adj}AF_l$ is the attributable fraction for the risk factor lth, adjusted for the remaining risk factors $E_1, E_2, \dots, E_L \setminus \{E_l\}$ and possible confounders.

Average AF

Let γ be a permutation and invertible function over the integers $1, 2, \dots, L$, $\gamma : \{1, 2, \dots, L\} \to \{1, 2, \dots, L\}$. For example, a permutation γ with $\gamma(1) = 5$ represents a removal order where risk factor E_1 is the 5th factor to be removed from the risk system. For L risk factors of interest, the set of all possible permutation functions is denoted by G_L. The average AF for risk factor l can be represented as:

$$ {}_{average}AF_l = \frac{\left(\sum_{\gamma \in G_L : \gamma(l)=1} {}_{sequential}AF_{l|\varnothing} + \sum_{\gamma \in G_L : \gamma(l) \neq 1} {}_{sequential}AF_{l|\gamma^{-1}(1)\, \gamma^{-1}(2)\, \dots\, \gamma^{-1}(\gamma(l)-1)} \right)}{L!}, $$

where ${}_{sequential}AF_{l|\gamma^{-1}(1)\dots\gamma^{-1}(\gamma(l)-1)}$ is the sequential AF from removing risk factor l having already removed the $l-1$ risk factors in the order indicated by γ. For example, consider a situation with $L = 4$ risk factors. Moreover, consider the sequential AF corresponding to removing E_2, after having removed first E_4, then E_3 and then E_1, that is: ${}_{sequential}AF_{2|431}$. Then, since E_4 is the first risk factor to be removed, $\gamma(4) = 1$. Similarly, $\gamma(3) = 2$, $\gamma(1) = 3$ and $\gamma(2) = 4$. Applying the inverse permutation, it follows that $\gamma^{-1}(1) = 4$, $\gamma^{-1}(2) = 3$, $\gamma^{-1}(3) = 1$, that are the three risk factors E_4, E_3 and E_1 that are removed before removing risk factor E_2, $\gamma^{-1}(4) = 2$.

3.2 Exact computation for the average AF

As outlined in section 2.5, there are some identical sequential AFs for a risk factor in the determination of the average AF. For example, consider the sequential AF: ${}_{sequential}AF_{l|321}$ corresponding to a permutation γ with

$\gamma^{-1}(1)=3$, $\gamma^{-1}(2)=2$, $\gamma^{-1}(3)=1$, and $\gamma^{-1}(4)=l$. It represents the proportional decrease in disease prevalence from removing the lth risk factor from the population, given that the risk factors E_3, E_2, and E_1 were already removed in that order. This calculation is the same regardless the order in which E_3, E_2, and E_1 were removed. In particular, the sequential AFs for the risk factor l are identical:

$$_{sequential}AF_{l|321} = {}_{sequential}AF_{l|312} = {}_{sequential}AF_{l|123} = {}_{sequential}AF_{l|132} = \\ = {}_{sequential}AF_{l|213} = {}_{sequential}AF_{l|231}.$$

Similarly, if there are 10 risk factors, the sequential AF $_{sequential}AF_{l|312}$ does not depend on which of the $(10-1-3)!=6!$ orders the remaining risk factors are deleted from the population, after deleting l. In general, let $C_r^{L,-l}$ be the set of all subsets of size r (i.e., unordered choices of r integers) from the set $E_1, E_2, \ldots, E_L \setminus \{E_l\}$. For $s \in C_r^{L,-l}$, let $_{sequential}AF_{l|s}$ be the sequential AF for the risk factor l, given that risk factors corresponding to the subset s were already removed from the population. Denoting the cardinality of the subset s, as $|s|=r$, the average AF formula can be re-expressed as:

$$_{average}AF_l = \sum_{r=0}^{L-1} \frac{\left((L-1-r)! \cdot r! \cdot \sum_{s \in C_r^{L,-l}} {}_{sequential}AF_{l|s} \right)}{L!},$$

where $_{sequential}AF_{l|s=\varnothing}$ indicates that risk factors l is removed first from the population. In this formula, the number of summands is 2^L as opposed to $L!$ summands included in the previous average AF's formula and thus can result in a substantial computational saving for large L.

3.3 Modification for case-control study design

Sequential and average AFs can best be developed in prospective longitudinal cohort studies, as such studies allow optimal measurements of risk factors, outcome, and direct estimation of relative risk (i.e., unbiased predicted probabilities). Cohort studies may require a large sample size or long follow-up duration for rare diseases. In contrast, case-control studies are more efficient, as they require fewer subjects and can be performed in a shorter timeframe than cohort studies. Thus, case-control sampling is widely used study design. For instance, van der Laan [24] searched for case-control analysis on PubMed resulting in a list of approximately 56,000 articles. Their use is not confined to public health applications; case-control studies are also frequently performed in econometric applications [48, 49]. However, if not explicitly nested within a cohort study, case-control studies are generally deemed less suitable for developing AFs due to their inability to allow the calculation of unbiased predicted probabilities. Case-control studies differ from cohort studies in that they sampled diseased (cases) and non-diseased (controls) subjects rather than exposed and unexposed ones. Thus, since the ratio of controls to cases in the sample is fixed a priori, the resulting predicted probabilities from the model will be biased, as will the AFs.

Ferguson et al. [16] proposed an approach for estimating sequential AF accounting for case-control study design by incorporating disease prevalence in the model used to estimate predicted cases. In particular, the method requires to find the coefficient vector $\hat{\boldsymbol{\beta}}$ to maximize:

$$\log L\left(\boldsymbol{\beta} \mid \mathbf{E},\mathbf{C}\right) = \sum_{i=1}^{N} w_i \cdot \log_i L\left(\boldsymbol{\beta} \mid \mathbf{E},\mathbf{C}\right),$$

where $\mathbf{E} = \{E_1, E_2, \ldots E_L\}$ is the set of risk factors, $\mathbf{C} = \{C_1, C_2, \ldots, C_M\}$ is the set of confounders, $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots \beta_{L+M}\}$ is the set of coefficient, $\log_i L\left(\boldsymbol{\beta} \mid \mathbf{E},\mathbf{C}\right)$ is the individual log-likelihood contribution for the ith observation, and $\mathbf{w} = \{w_1, w_2, \ldots w_N\}$ is the set of weights. Suppose that the ratio of controls to cases in the sample is $v:1$ and the prevalence of disease in the population is p. If each case is given a weight of 1, then each control should be given a weight of $(1-p)/v \cdot p$. In the case that $(1-p)/v \cdot p$ is an integer, the estimated probabilities of disease from the weighted model are identical to those estimates that would be found from an unweighted model with an altered design matrix where the row for each control is repeated $(1-p)/v \cdot p$ times, and the row for each case only once. The formula for the sequential AF also need to be adjusted to account for the imbalance between cases and controls, as follows:

$${}_{sequential}\widehat{AF}_{l|12\ldots(l-1)} = \frac{\left(\sum_{i=1}^{N} w_i \cdot \hat{p}_{i,12\ldots(l-1)} - \sum_{i=1}^{N} w_i \cdot \hat{p}_{i,12\ldots l}\right)}{N_{cases}},$$

where $\hat{p}_{i,12\ldots(l-1)}$ is the predicted probability that the ith individual is a case, assuming the values of the risk factors $E_1, E_2, \ldots, E_{(l-1)}$ and $\hat{p}_{i,12\ldots l}$ is the predicted probability that the ith individual is a case, assuming the values of the risk factors $E_1, E_2, \ldots, E_{(l-1)}, E_l$.

The average AF for risk factor l, can be estimated by substituting the sequential AFs adjusted for the case-control structure, in the expression described previously:

$${}_{average}\widehat{AF} = \frac{\left(\sum_{\gamma\in G_L:\gamma(l)=1} {}_{sequential}\widehat{AF}_{l|\varnothing} + \sum_{\gamma\in G_L:\gamma(l)\neq 1} {}_{sequential}\widehat{AF}_{l|\gamma^{-1}(1)\;\gamma^{-1}(2)\;\ldots\;\gamma^{-1}(\gamma(l)-1)}\right)}{L!}.$$

Haaf and Steyerberg [50] discussed a number of aspects of weighting controls to calculate absolute risk in non-nested case-control studies. They focused on the issue regarding the selection process of the cases and controls. A major limitation of case-control studies is the difficulty to ensure that cases and controls are a representative sample of the same source of population [28]. By weighting the controls, the assumption that there are no factor influencing the selection of controls other than those considered in the weighting formula (i.e., the prevalence of disease) should be carefully considered. However, this method can be easily applied. It does not require approximating relative risks with odds ratios and is a valid approach for common disease (i.e., disease with high prevalence). Van der Laan discussed the approach of applying weights to maximum likelihood estimator in case-control studies in a more general context [24].

3.4 AF estimate comparisons

As an example, we estimated crude, adjusted, sequential and average fractions attributable to tobacco smoking and alcohol drinking for oral cavity cancer. We used data from an Italian multicentre case-control study on 946 cases and 2492 controls [51]. This analysis was conducted on 942 cases and 2490 controls with complete information about smoking and alcohol to ensure comparability across results. Overall, 324 subjects were not exposed to both risk factors (i.e., subjects who never smoked and drank during their lifetime), 892 subjects were never smokers and ever (current or former) drinkers, 185 subjects were ever smokers and never drinkers, and 2031 subjects were jointly exposed to smoking and alcohol. We set a 81×10^{-5} prevalence of oral cavity cancer [52] to adjust sequential and average AF estimates for the study design.

AF estimates were dependent on the method applied for their estimation (table 3.1). For example, the crude fraction of oral cavity cases attributable to smoking was 66%. The AF for smoking was 65% adjusting for alcohol consumption. The effect of removing tobacco smoking as first risk factor from the population yielded a sequential AF of 65%; when smoking was removed after alcohol yielded a sequential AF of 33%. The average AF for smoking was 49%.

Table 3.1. AFs for tobacco smoking and alcohol drinking for oral cavity cancer according to different methods.

Risk factors	Method				
	Crude AF	Adjusted AF§ (Bruzzi's formula)	Sequential AF†	Sequential AF‡	Average AF
Smoking	0.66	0.65	0.65	0.33	0.49
Alcohol	0.61	0.48	0.16	0.48	0.32
Sum	1.27	1.13	0.81	0.81	0.81

§Mutually adjusted; †Smoking was removed first and then alcohol; ‡Alcohol was removed first and then smoking.

For alcohol consumption, the crude and adjusted AFs were 61% and 48%, respectively. The sequential AFs were 48% and 16% according to the removal order considered and the average AF was 32%.

Note that, both crude and adjusted AFs did not sum up to the joint AF (81%) and their sum exceeded 100% (table 3.1). Moreover, if the risk factors are independent (i) and if there is no interaction among the risk factors in the logistic model (ii) and if the disease is rare (iii), so that:

$$P\left(D \mid E_1, E_2, \ldots, E_L\right) = \mu + \sum_{l=1}^{L} \beta_l E_l \ ,$$

then

$$\left(1-AF\right)=\sum_{\mathbf{E}}P\left(\mathbf{E}\right)\cdot\exp\left(-\sum_{l=1}^{L}\beta_l E_l\right)=\sum_{\mathbf{E}}\prod_{l=1}^{L}P\left(\mathbf{E}\right)\cdot\exp\left(\beta_l E_l\right)=$$
$$=\prod_{l=1}^{L}\sum_{E_l}P\left(E_l\right)\cdot\exp\left(\beta_l E_l\right)=\prod_{l=1}^{L}\left(1-AF_l\right)$$

In the example, applying this formula to adjusted AFs, we get a joint AF of 82%, which is very close to the joint AF in table 3.1, suggesting that the assumptions were not violated. Although there is some correlation between risk factors (ρ=0.25), leading us to expect a difference in joint AFs, the high exposure prevalence (65% of the sampled subjects were smokers and 85% were drinkers) could explain the converge in values. Finally, the sequential and average AFs summed up to the joint AF according to the component-additivity property.

Chapter 4

Variance estimation

4.1 Introduction

Several authors dealt with the problem of estimating confidence intervals for AFs. Walter derived variance estimates for the AF using data from case-control studies with dichotomous risk factors [19]; later Denman and Schlesselman extended this work to risk factors with several levels [53]. Whittemore examined the role of confounders and proposed the type I stratification approach for estimating adjusted AFs together with variance estimates [8]. The method applies whether controls are selected at random or with frequency matching [54]. Using the delta method, Benichou and Gail [15, 54] gave variance estimates for AFs in case-control studies. Greenland provided for both cohort and case-control studies the maximum likelihood estimator for AF based on logistic regression model and corresponding variance estimators [22]. Kooperberg and Petiti [55] used Bruzzi's formula for AF to obtain a bootstrap confidence interval for case-control data. Graubard and Fears [56] considered AF estimates for unmatched, frequency matched, and individual matched case-control, cross-sectional, and cohort studies. They used influence function method to obtain Taylor deviates for estimating variances. Natarajan and Rimm [57] proposed a simple method to calculate AF confidence intervals using the Bonferroni inequality. In 2006, Ferguson and colleagues [16] introduced the "averisk" R package [58] for calculating average AFs and corresponding confidence intervals in prospective and case-controls studies. They derived confidence intervals using Monte Carlo simulations.

4.2 Delta Method

Benichou and Gail [15] proposed an approach for estimating the confidence intervals for AFs from logistic models based on case-control data using the implicit-function theorem [59] and the delta method [60]. The relative risk obtained from a logistic regression model may be correlated with the risk factor prevalence, and the usual likelihood equations for the logistic model do not yield estimates of the covariances between these two quantities that are needed to compute AF variance. The authors used the delta method for implicitly defined random variables to estimate all required terms for constructing confidence intervals for AFs.

Implicit function theorem

Let S be an open subset of the $(p+m)$-dimensional space with elements $(x_1, x_2, \ldots, x_p; y_1, y_2, \ldots, y_m) \equiv (\mathbf{X}, \mathbf{Y})$. Let p real functions g_i, be continuous in S that have continuous first partial derivatives in S and satisfy $g_i(x_1, x_2, \ldots, x_p; y_1, y_2 \ldots, y_m) = 0$ for all $i = 1, 2, \ldots, p$ at some point (x_0, y_0) in S. Define $\mathbf{J}$ as the $p \times p$ matrix with elements $(\partial g_i / \partial x_j)$ for $i, j = 1, 2, \ldots, p$. If the determinant $|\mathbf{J}| \neq 0$ at (x_0, y_0), there exists an open rectangular region in S satisfying:

$$R_1 : |x_1 - x_{10}| < a_1, |x_2 - x_{20}| < a_2, \ldots, |x_p - x_{p0}| < a_p,$$

and

$$R_2 : |y_1 - y_{10}| < b_1, |y_2 - y_{20}| < b_2, \ldots, |y_m - y_{m0}| < b_m.$$

Thus, there exists a set of p real functions f_i, that map each element $\mathbf{Y}$ in the region R_2 to a single element $\mathbf{X}$ in the region R_1 such that:

$$\mathbf{X} = \{x_1, x_2, \ldots, x_p\} = \{f_1(\mathbf{Y}), f_2(\mathbf{Y}), \ldots, f_p(\mathbf{Y})\},$$

and

$$g_i(\mathbf{X}, \mathbf{Y}) = 0 \text{ for } i = 1, 2, \ldots, p.$$

Moreover, the functions f_i are continuous and have continuous first partial derivatives $(\partial f_i / \partial y_j)$, which are the matrix product $-\mathbf{J}^{-1}\mathbf{H}$, where the $p \times m$ matrix $\mathbf{H}$ has elements $(\partial g_i / \partial y_j)$.

The implicit function theorem asserts the existence, in a neighborhood of (x_0, y_0), of the explicit functions f_i needed to apply the delta method theorem.

Delta method theorem

Let $\mathbf{Y}_n$ be a random vector that converges in probability to μ_y:

$$\mathbf{Y}_n \xrightarrow{p} \mu_y ,$$

and $\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right)$ converge to a multivariate normal distribution with mean 0 and variance $\mathbf{\Sigma}$:

$$\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right) \xrightarrow{p} N\left(0, \mathbf{\Sigma}\right).$$

Let the p real functions as $x_{ij} = f_i\left(\mathbf{Y}_n\right)$ with $i = 1, 2, \ldots, p$, have m continuous first partial derivatives at μ_y, where at least one of these derivatives is non-zero. Then, the $1 \times p$ vector with elements $\sqrt{n}\left(x_{ij} - f_i\left(\mu_y\right)\right)$ converges to a multivariate normal distribution with mean 0 and covariance $\mathbf{M\Sigma M}'$, where $\mathbf{M}$ is the $p \times m$ matrix with elements $\left(\partial f_i / \partial y_{jn}\right)$.

In their work, Benichou and Gail [15] proposed the following corollary based on the implicitly defined random variables.

Benichou and Gail's corollary

Let $\mathbf{Y}_n$ be a random vector as defined in the delta method theorem. Suppose that there exists a unique $1 \times p$ vector $\mu_x = \mathbf{X}_n$ satisfying the system of p equations $g_i\left(\mathbf{X}_n, \mu_y\right) = 0$. The functions g_i are continuous with continuous first partial derivatives in an open set containing $\left(\mu_x, \mu_y\right)$. Let $\mathbf{J}$ be a $p \times p$ matrix defined as in the implicit function theorem with non-zero determinant $|\mathbf{J}|$ at $\left(\mu_x, \mu_y\right)$. Moreover, each row of the $p \times m$ matrix $\mathbf{J}^{-1}\mathbf{H}$ defined in the implicit function theorem contains at least one non-zero element. Then, as n increases, to each $\mathbf{Y}_n$ there corresponds a unique solution $\mathbf{X}_n$ to the system of equations $g_i\left(\mathbf{X}_n, \mathbf{Y}_n\right) = 0$ with $i = 1, 2, \ldots, p$, and $\sqrt{n}\left(\mathbf{X}_n - \mu_x\right)$ converges to a p-variate normal distribution with mean 0 and covariance matrix $\mathbf{J}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}'\left(\mathbf{J}^{-1}\right)'$, where $\mathbf{H}$ and $\mathbf{J}$ are evaluated at $\left(\mu_x, \mu_y\right)$.

Proof. Let $\left(\mu_x, \mu_y\right) = \left(\mathbf{X}_0, \mathbf{Y}_0\right)$ as in implicit-function theorem. When $\mathbf{Y}_n \in R_2$, then $\mathbf{X}_n = \left\{ f_1\left(\mathbf{Y}_n\right), f_2\left(\mathbf{Y}_2\right), \ldots, f_p\left(\mathbf{Y}_n\right)\right\}$.

Since $\mu_n \equiv \left\{ f_1\left(\mu_y\right), f_2\left(\mu_y\right), \ldots, f_p\left(\mu_y\right)\right\}$ it follows that:

$$\sqrt{n}\left(\mathbf{X}_n - \mu_x\right) = \sqrt{n}\left(\left\{ f_1\left(\mathbf{Y}_n\right), f_2\left(\mathbf{Y}_n\right), \ldots, f_p\left(\mathbf{Y}_n\right)\right\} - \left\{ f_1\left(\mu_y\right), f_2\left(\mu_y\right), \ldots, f_p\left(\mu_y\right)\right\}\right) \cdot$$
$$\cdot I\left(\mathbf{Y}_n \in R_2\right) + \sqrt{n}\left(\mathbf{X}_n - \mu_x\right) \cdot I\left(\mathbf{Y}_n \notin R_2\right) \equiv \mathbf{Z}_{1n} + \mathbf{Z}_{2n} ,$$

where $I(\cdot) = 1$ is an indicator variable. Since $\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right)$ converges in distribution to a normal distribution with mean 0, then:

$$\mathbf{Y}_n \xrightarrow{p} \mu_y .$$

Thus, $P\left(\mathbf{Y}_n \in R_2\right) \to 1$ and $I\left(\mathbf{Y}_n \in R_2\right) \xrightarrow{p} 1$, which implies that $\mathbf{Z}_{2n} \xrightarrow{p} 0$. Moreover, the delta method theorem implies that $Z_{1n} / I\left(\mathbf{Y}_n \in R_2\right)$ has a limiting normal distribution with mean 0 and covariance matrix $\mathbf{J}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}'\left(\mathbf{J}^{-1}\right)'$.

Remarks If $\hat{\mathbf{\Sigma}}$ is a consistent estimate of $\mathbf{\Sigma}$ evaluated at $\mathbf{Y}_n$, then the asymptotic covariance $\mathbf{J}^{-1}\mathbf{H}\mathbf{\Sigma}\mathbf{H}'\left(\mathbf{J}^{-1}\right)'$ may be consistently estimated by substituting $\hat{\mathbf{\Sigma}}$ for $\mathbf{\Sigma}$ and by substituting $\mathbf{J}^{-1}\mathbf{H}$ evaluated at $\left(\mathbf{Y}_n, \mathbf{X}_n\right)$ for $\mathbf{J}^{-1}\mathbf{H}$ evaluated at $\left(\mu_y, \mu_x\right)$. This follows from the continuity of $\mathbf{J}$ and $\mathbf{H}$ in an open space containing $\left(\mu_y, \mu_x\right)$.

If two set of random variables, $\mathbf{Y}_n$ of dimension $1 \times m_1$ and $\mathbf{W}_n$ of dimension $1 \times m_1$, are separately defined by implicit functions $g_i\left(\mathbf{X}_n, \mathbf{Y}_n\right) = 0$ for $i = 1, 2, \ldots, m_1$ and $g_i\left(\mathbf{W}_n, \mathbf{Y}_n\right) = 0$ for $i = m_1 + 1, m_1 + 2, \ldots, m_1 + m_2$, then under the conditions of the corollary, $\sqrt{n}\left(\mathbf{X}_n - \mu_x\right)$ and $\sqrt{n}\left(\mathbf{W}_n - \mu_w\right)$ converge to a multivariate normal distribution with mean 0, and the $m_1 \times m_2$ partition of the covariace matrix is $\mathbf{J}_x^{-1}\mathbf{H}_x\mathbf{\Sigma}\mathbf{H}'_w\left(\mathbf{J}_w^{-1}\right)'$, where . $\left(\mathbf{J}_x, \mathbf{H}_x\right)$ and $\left(\mathbf{J}_w, \mathbf{H}_w\right)$ are evaluated at $\left(\mu_x, \mu_y\right)$ and $\left(\mu_w, \mu_y\right)$, respectively.

In 1989, Benichou and Gail proposed the variance estimation for the AFs using the results of these theorems as follows.

Univariate case

Let $\mathbf{Y}_n$ be an underlying random variable, and $\mathbf{X}_n$ be a derived random variable. Assume that $\mathbf{Y}_n$ converges in probability to μ_y:

$$\mathbf{Y}_n \xrightarrow{p} \mu_y ,$$

and $\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right)$ is asymptotically distributed as a normal random variable with mean 0 and finite variance σ^2:

$$\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right) \sim N\left(0, \sigma^2\right).$$

Let the totally differentiable function $\mathbf{G}\left(\mathbf{X}_n, \mathbf{Y}_n\right) = 0$ define $\mathbf{X}_n$ uniquely for each $\mathbf{Y}_n$. The derived random variable $\mathbf{X}_n$ converges in probability to μ_x:

$$\mathbf{X}_n \xrightarrow{p} \mu_x ,$$

which satisfies $\mathbf{G}\left(\mathbf{X}_n, \mathbf{Y}_n\right) = 0$. Because $\mathbf{G}$ is constant, then

$$0 = d\mathbf{G} = \left(\frac{\partial \mathbf{G}}{\partial \mathbf{X}_n}\right)\left(\mathbf{X}_n - \mu_x\right) + \left(\frac{\partial \mathbf{G}}{\partial \mathbf{Y}_n}\right)\left(\mathbf{Y}_n - \mu_y\right) + R_n .$$

Ignoring the remainder R_n and assuming $\left(\partial \mathbf{G} / \partial \mathbf{X}_n\right) \neq 0$, $\left(\mathbf{X}_n - \mu_x\right)$ can be expressed as:

$$\left(\mathbf{X}_n - \mu_x\right) = -\left(\frac{\partial \mathbf{G}}{\partial \mathbf{X}_n}\right)^{-1}\left(\frac{\partial \mathbf{G}}{\partial \mathbf{Y}_n}\right)\left(\mathbf{Y}_n - \mu_y\right) = -\mathbf{J}^{-1}\left(\frac{\partial \mathbf{G}}{\partial \mathbf{Y}_n}\right)\left(\mathbf{Y}_n - \mu_y\right),$$

consequently:

$$\sqrt{n}\left(\mathbf{X}_n-\mu_x\right)\sim N\left(0,\mathbf{J}^{-1}\left(\frac{\partial\mathbf{G}}{\partial\mathbf{Y}_n}\right)\sigma^2\left(\frac{\partial\mathbf{G}}{\partial\mathbf{Y}_n}\right)\mathbf{J}^{-1}\right),$$

i.e., $\sqrt{n}\left(\mathbf{X}_n-\mu_x\right)$ is asymptotically normal with mean 0 and variance $\mathbf{J}^{-1}\left(\partial\mathbf{G}/\partial\mathbf{Y}_n\right)\sigma^2\left(\partial\mathbf{G}/\partial\mathbf{Y}_n\right)\mathbf{J}^{-1}$.

Multivariate extension

Suppose that $\mathbf{G}=\left\{g_1,g_2,\ldots,g_p\right\}$ is a vector of p functions of $\mathbf{X}_n$ and $\mathbf{Y}_n$. Denote $\mathbf{J}$ the $p\times p$ matrix with elements $\left(\partial g_i/\partial x_{jn}\right)$, where $\mathbf{X}_n=\left\{x_{1n},x_{2n},\ldots,x_{pn}\right\}$ and let $\mathbf{H}$ denote the $p\times m$ matrix with elements $\left(\partial g_i/\partial y_{mn}\right)$, where $\mathbf{Y}_n=\left\{y_{1n},y_{2n},\ldots,y_{mn}\right\}$. Then, the random vector $\mathbf{X}_n=\left\{x_{1n},x_{2n},\ldots,x_{pn}\right\}$ may be normalized so that:

$$\sqrt{n}\left(\mathbf{X}_n-\mu_x\right)\sim N\left(0,\mathbf{J}^{-1}\mathbf{H}\boldsymbol{\Sigma}\mathbf{H}'\left(\mathbf{J}^{-1}\right)'\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\sqrt{n}\mathbf{Y}_n$. In the special case $g_i=x_{in}-h_i\left(\mathbf{Y}_n\right)$ the Jacobian matrix $\mathbf{J}$ is the $p\times p$ identity matrix, and the results of the classical delta method follow.

4.3 Confidence interval for AF using the delta method: an example

As an example, consider the hypothetical case-control study in the work of Benichou and Gail [15]. The study includes one risk factor E with $Q+1=4$ levels ($q=0,1,2,3$), for a given disease (table 4.1). Denote n_{1q} and n_{0q} the number of cases and controls at exposure level q, respectively. Moreover, denote by n_{+q} the total number of individual in the qth exposure level. Denote n_{1+} and n_{0+} the total number of cases and controls, respectively.

Table 4.1. Hypothetical case-control data for a risk factor E with four levels.

Subjects	Exposure level				Total
	$q=0$	$q=1$	$q=2$	$q=3$	
Controls	$n_{00}=132$	$n_{01}=36$	$n_{02}=22$	$n_{03}=10$	$n_{0+}=200$
Cases	$n_{10}=45$	$n_{11}=33$	$n_{12}=55$	$n_{13}=67$	$n_{1+}=200$
Total	$n_{+0}=177$	$n_{+1}=69$	$n_{+2}=77$	$n_{+3}=77$	

As outlined in the previous section, the interest is the estimation of the covariance between relative risk and exposure levels, i.e., between $\widehat{RR}_q$ and $\hat{\rho}_{1q}$ with $q=0,1,2,3$. The log-likelihood for the logistic model is:

$$\log L(\boldsymbol{\beta}\,|\,E)=\sum_{q=0}^{3}\left[n_{1q}\cdot(\beta_0+q\beta_E)-n_{+q}\cdot\log\left\{1+\exp(\beta_0+q\beta_E)\right\}\right].$$

Logistic regression model yields the estimates $\hat{\beta}_E=9.94\times10^{-1}$ and $\widehat{Var}\left(\hat{\beta}_E\right)=1.19\times10^{-2}$. The variance is obtained from the information matrix evaluated at $\rho_{iq}=n_{iq}/n_{i+}$, with $i=1,2$ and $q=0,1,2,3$. Furthermore, from the Bruzzi's formula the AF for the risk factor E can be estimated by:

$${}_{adj}AF_E=1-\sum_{q=0}^{3}\frac{\hat{\rho}_q}{\widehat{RR}_q}=1-\sum_{q=0}^{3}\hat{\rho}_{iq}\cdot\exp\left(-q\hat{\beta}_E\right)=0.695\,.$$

To apply the delta method for implicitly defined random variables, denote the vectors $\hat{\boldsymbol{\rho}}=\{\hat{\rho}_{00},\hat{\rho}_{01},\hat{\rho}_{02},\hat{\rho}_{03},\hat{\rho}_{10},\hat{\rho}_{11},\hat{\rho}_{12},\hat{\rho}_{13}\}$ and $\hat{\boldsymbol{\beta}}=\left\{\hat{\beta}_0,\hat{\beta}_E\right\}$ by $\mathbf{Y}_n$ and $\mathbf{X}_n$ respectively. The maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_E$ are implicit functions of $\mathbf{Y}_n$ as defined by the score equations:

$$g_1(\mathbf{X}_n,\mathbf{Y}_n)=\sum_{q=0}^{3}\left[n_{1+}+\hat{\rho}_{1q}-\left(n_{1+}\cdot\rho_{1q}+n_{0+}\cdot\hat{\rho}_{0q}\right)\cdot\hat{\psi}_q\right]=0\,,$$

and

$$g_2(\mathbf{X}_n,\mathbf{Y}_n)=\sum_{q=0}^{3}\left[n_{1+}+\hat{\rho}_{1q}-\left(n_{1+}\cdot\hat{\rho}_{1q}+n_{0+}\cdot\hat{\rho}_{0q}\right)\cdot\hat{\psi}_q\right]=0\,,$$

where $\hat{\psi}_q = \exp\left(\hat{\beta}_0 + q\hat{\beta}_E\right) \cdot \left\{1 + \exp\left(\hat{\beta}_0 + q\hat{\beta}_E\right)\right\}^{-1}$.

Let $E(\mathbf{Y}_n) \equiv \mu_y = (p_{00}, p_{01}, p_{02}, p_{03}, p_{10}, p_{11}, p_{12}, p_{13})$. The vector $\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right)$ converges to an 8-dimensional multivariate normal distribution with mean 0 and block-diagonal covariance matrix $\mathbf{\Sigma}$ because the estimates $\hat{\rho}_{0q}$ and $\hat{\rho}_{1q}$ are from independent multinomial samples. The upper left block of $\mathbf{\Sigma}$ is multinomial covariance matrix $\mathbf{\Sigma}_0$ of the 1×4 vector $\sqrt{n}(\hat{\rho}_{00}, \hat{\rho}_{01}, \hat{\rho}_{02}, \hat{\rho}_{03})$; the lower right block is the multinomial covariance matric $\mathbf{\Sigma}_1$ of the 1×4 vector $\sqrt{n}(\hat{\rho}_{10}, \hat{\rho}_{11}, \hat{\rho}_{12}, \hat{\rho}_{13})$. Let $I_i = \sum_{q=0}^{3} q^i \cdot \left(n_{1+} \cdot \hat{\rho}_{1q} + n_{0+} \cdot \hat{\rho}_{0q}\right) \cdot \hat{\psi}_q \cdot \left(1 - \hat{\psi}_q\right)$. The estimate of the Jacobian matrix $\mathbf{J}_x$ is:

$$-\mathbf{J}_x = \begin{bmatrix} \hat{I}_0 & \hat{I}_1 \\ \hat{I}_1 & \hat{I}_2 \end{bmatrix} = \begin{bmatrix} 75.20 & 74.76 \\ 74.76 & 158.30 \end{bmatrix},$$

at the point $(\mathbf{X}_n, \mathbf{Y}_n)$, and $\left|-\hat{\mathbf{J}}_x\right| \neq 0$. Moreover, the estimate of the 2×8 matrix $\mathbf{H}_x$ is:

$$\mathbf{H}_x = \begin{bmatrix} -50.82 & -95.86 & -142.70 & -174.10 & 149.20 & 104.10 & 57.35 & 25.90 \\ 0.00 & -95.86 & -285.30 & -522.30 & 0.00 & 104.10 & 114.70 & 77.70 \end{bmatrix}$$

at the point $(\mathbf{X}_n, \mathbf{Y}_n)$; each row of the 2×8 matrix $-\hat{\mathbf{J}}_x^{-1}\hat{\mathbf{H}}_x$ has non-zero elements at $(\mathbf{X}_n, \mathbf{Y}_n)$. Note that all the condition of the Benichou and Gail's corollary are satisfied at the point $(\mathbf{X}_n, \mathbf{Y}_n)$.

From corollary, $\sqrt{n}(\mathbf{X}_n - \mu_x)$ tends to normality with 0 mean and covariance matrix estimated by $-\hat{\mathbf{J}}_x^{-1}\hat{\mathbf{H}}_x\hat{\mathbf{\Sigma}}\hat{\mathbf{H}}_x'\left(-\hat{\mathbf{J}}_x^{-1}\right)'$, where $\hat{\mathbf{\Sigma}}$ is the estimate of $\mathbf{\Sigma}$ at $\mathbf{Y}_n$. In particular, an estimate of the covariance matrix of $\hat{\boldsymbol{\beta}} = \left\{\hat{\beta}_0, \hat{\beta}_E\right\}$ is obtained without relying on likelihood theory except for the functional form of the estimating score equations. Likewise, it is possible to compute the covariance between $\widehat{RR}_q$ and $\hat{\rho}_{1q}$, which are not obtainable from the logistic regression model. From the corollary, the quantities $\sqrt{n}(\mathbf{X}_n - \mu_x)$ and $\sqrt{n}\left(\mathbf{Y}_n - \mu_y\right)$ tend to normality with mean 0, and the corresponding 2×8 partition of the covariance matrix can be estimated by $-\hat{\mathbf{J}}_x^{-1}\hat{\mathbf{H}}_x\hat{\mathbf{\Sigma}}$. This covariance matrix is obtained by $\mathbf{W}_n = \mathbf{Y}_n$ (see remarks in the previous section). Thus the covariance between $\widehat{RR}_q$ and $\hat{\rho}_{1q}$ are:

$$\widehat{Cov}\left(\hat{\beta}_E, \hat{\rho}_{10}\right) = -1.986\times10^{-3};\ \widehat{Cov}\left(\hat{\beta}_E, \hat{\rho}_{11}\right) = 5.989\times10^{-6}$$

$$\widehat{Cov}\left(\hat{\beta}_E, \hat{\rho}_{12}\right) = 9.441\times10^{-4};\ \widehat{Cov}\left(\hat{\beta}_E, \hat{\rho}_{13}\right) = 1.036\times10^{-3}.$$

From the delta method:

$$\widehat{Var}\left(AF_E\right) = \frac{1}{n}\cdot\mathbf{B}\cdot\sum\mathbf{B}' + k^2\cdot\widehat{Var}\left(\beta_E\right) + k\mathbf{B}\hat{\mathbf{C}}',$$

where $k = -\sum_{q=0}^{3} q\rho_{1q}\exp\left(-q\beta_E\right)$, $\mathbf{B}$ is the 1×4 matrix with elements $\exp\left(-q\beta_E\right)$ and $\hat{\mathbf{C}}$ is the 1×4 matrix with elements $\widehat{Cov}\left(\hat{\beta}_E, \hat{\rho}_{1q}\right)$, with $q = 0,1,2,3$. By replacing parameters by their estimates, the variance of AF is: $\widehat{Var}\left(AF_E\right) = 1.768\times10^{-3}$. Thus, a $100\cdot(1-\alpha)\%$ confidence interval for the AF for the risk factors E may be compute by:

$$AF_E \pm Z_{(1-\alpha/2)}\cdot\sqrt{\widehat{Var}\left(AF_E\right)} = 0.695\pm1.96\cdot\sqrt{1.768\cdot10^{-3}} = \left[0.577; 0.742\right].$$

4.4 Monte Carlo confidence interval for average AF

The Monte Carlo confidence interval is a computer-based approach for constructing variance parameter via simulation in place of the theoretical analysis [61]. Ferguson et al. used a Monte Carlo approach to compute the confidence interval for the average AF [16]. The average AF can be regarded as a function, $f\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)$ of both the estimated regression model coefficients, $\hat{\boldsymbol{\beta}}$ that is used to generate the predicted probabilities and the observed design matrix, $\mathbf{X}$. Moreover, the sampling variance of $\hat{\boldsymbol{\beta}}$ is the most important factor in the determination of $Var\left({}_{average}AF_l\right)$:

$$Var\left({}_{average}AF_l\right) = Var_{\mathbf{X},\hat{\boldsymbol{\beta}}}\left\{f\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)\right\} \approx Var_{\hat{\boldsymbol{\beta}}}\left\{f\left(\mathbf{X},\hat{\boldsymbol{\beta}}\right)\mid\mathbf{X}\right\}.$$

The method consists of simulating B vectors, $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_B$ from the estimated asymptotic sampling distribution of $\hat{\boldsymbol{\beta}}$. The simulated values follow a multivariate normal distribution with mean $\hat{\boldsymbol{\beta}}$ and covariance $\hat{\mathbf{C}}$: i.e., $\hat{\boldsymbol{\beta}}_b \sim N\left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{C}}\right)$. For a risk factor of interest l, separate estimates ${}_{average}AF_l^1, {}_{average}AF_l^2, \ldots, {}_{average}AF_l^B$ for the average AF of that risk factor are produced from each sampled $\boldsymbol{\beta}_b$ vector ($b = 1, 2, \ldots, B$), by setting ${}_{average}AF_l^b = f\left(\mathbf{X}, \boldsymbol{\beta}_b\right)$. In practice, each Monte Carlo simulation yields an average attributable fraction, ${}_{average}AF_l^b$, where the estimated sequential attributable fractions, ${}_{sequential}AF_{l\,|\,\gamma}^b$, are calculated using the coefficient vector $\boldsymbol{\beta}_b$ rather than $\hat{\boldsymbol{\beta}}$, and the same design matrix $\mathbf{X}$. The variance of ${}_{average}AF_l$ can be estimated as:

$$\widehat{Var}\left({}_{average}AF_l\right) = \frac{\sum_{b=1}^{B}\left({}_{average}AF_l^b - {}_{average}\overline{AF_l}\right)^2}{\left(B-1\right)},$$

where ${}_{average}\overline{AF}_l = \dfrac{\sum_{b=1}^{B} {}_{average}AF_l^b}{B}$.

Now, a $100\cdot\left(1-\alpha\right)\%$ confidence interval is produced using:

$${}_{average}AF_l \pm t_{B-1;(1-\alpha/2)} \cdot \sqrt{\widehat{Var}\left({}_{average}AF_l\right)}.$$

The t-distribution quantile is used to reflect the fact that the variance is estimated, as opposed to a Delta Method using a normal distribution quantile.

This method is valid both for cohort and case-control study designs. In case-control studies, the sample size of the weighted likelihood is artificially high to account for the imbalance between cases and controls. Thus, the matrix $\hat{\mathbf{C}}$ is estimated using the covariance matrix from the unweighted model because it correctly respects the actual number of cases and controls in the sample. Efron and Tibshirani [61] compared the standard error estimates for various values of B. They suggested $B = 100$ replicates because "there is little improvement in the standard error estimates for values of $B > 100$".

4.5 Variance of average AF: law of the total variance

In probability theory, the law of the total variance (also termed variance decomposition formula) states that if X and Y are random variables on the same probability space, and the variance of X is finite, then

$$Var(X) = E\left[Var(X \mid Y)\right] + Var\left[E(X \mid Y)\right].$$

The first term is the expectation of the conditional variance of X given Y and it is also termed “unexplained component of the variance”, whereas the second term is the variance of the conditional expectation of X given Y, also termed the “explained component of the variance”. For a proof of the variance decomposition formula, see appendix A.3.

Recalling the approach proposed by Ferguson to estimate average AF variance, a generic Monte Carlo simulation, b, yield $L!$ sequential AFs and one average AF for each risk factor. Now, B simulations generate a total of $B \cdot L!$ sequential AFs and B average AFs for the lth risk factor, as reported in table 4.2.

Table 4.2. Sequential and average AFs generated by B simulations and $L!$ removal orders for a generic risk factor l.

Simulation	Removal order					
	1	$\cdots$	l	$\cdots$	$L!$	
1	${}_{seq}AF^{1}_{l\mid\gamma(l)=1}$	$\cdots$	${}_{seq}AF^{1}_{l\mid\gamma(l)=l}$	$\cdots$	${}_{seq}AF^{1}_{l\mid\gamma(l)=L!}$	${}_{ave}AF^{1}_{l}$
$\vdots$	$\vdots$	$\ddots$			$\vdots$	$\vdots$
b	${}_{seq}AF^{b}_{l\mid\gamma(l)=1}$		${}_{seq}AF^{b}_{l\mid\gamma(l)=l}$		${}_{seq}AF^{b}_{l\mid\gamma(l)=L!}$	${}_{ave}AF^{b}_{l}$
$\vdots$	$\vdots$			$\ddots$	$\vdots$	$\vdots$
B	${}_{seq}AF^{B}_{l\mid\gamma(l)=1}$	$\cdots$	${}_{seq}AF^{B}_{l\mid\gamma(l)=l}$	$\cdots$	${}_{seq}AF^{B}_{l\mid\gamma(l)=L!}$	${}_{ave}AF^{B}_{l}$

For a generic risk factor l, we can assume that ${}_{average}AF_{l}$ and the vector of simulated parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_B$ are realizations of two random variables belonging to the same probability space, and that the variance of ${}_{average}AF_{l}$ is finite. Thus, according to the law of the total variance, the variance of ${}_{average}AF_{l}$ can be expressed as:

$$
\begin{aligned}
\widehat{Var}\left({}_{average}AF_l\right) &= E\left[Var\left({}_{average}AF_l \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_B\right)\right] + \\
&\quad + Var\left[E\left({}_{average}AF_l \mid \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_B\right)\right] = \\
&= \frac{\sum_{b=1}^{B} \sum_{\gamma \in G_L} \left({}_{sequential}AF_l^b - {}_{average}AF_l^b\right)^2}{\left(L! - 1\right) \cdot B} + \frac{\sum_{b=1}^{B} \left({}_{average}AF_l^b - {}_{average}\overline{AF}_l\right)^2}{\left(B - 1\right)}.
\end{aligned}
$$

The formula proposed by Ferguson et al. to estimate average AF variance corresponds to the second summand: i.e., the variation of average AFs from their total mean (${}_{average}\overline{AF}_l$) across the B simulations. In the formula above, we introduce an additional variability source given by the variation of sequential AF for a fixed b across the $L!$ permutations. Then, we compute the average of such within-simulation variances (internal component of variance).

In the next chapter, we will compare the performance between our and Ferguson's methods using different simulated datasets.

Chapter 5

Variance estimate comparisons

5.1 Variance behavior: an example

The behavior of the AF variability may vary according to the number of the risk factors and the correlation among them. Let E_1 and E_2 be two independent risk factors with a 0.5 prevalence and a relative risk equals to 2, as reported in table 5.1. For risk factor E_1, sequential AFs are 0.33 when E_1 is the first risk factor removed and 0.22 when E_1 is eliminated after having already removed E_2 from the population. The corresponding average AF is 0.275.

Table 5.1. Prevalence (P) and relative risk (RR) for two independent risk factors.

Risk factor			
E_1	E_2	P	RR
0	0	0.25	1.0
0	1	0.25	2.0
1	0	0.25	2.0
1	1	0.25	4.0

Let E_3 be a third independent risk factor with the same prevalence and RR as before (table 5.2). Now, when E_1 is the first risk factor removed, the corresponding sequential AF is 0.28. This calculation is the same no matter the order in which risk factors E_2 and E_3 are removed after E_1. For the other removal orders, sequential AFs are 0.24 when E_1 is eliminated after E_2 or after E_3, and 0.16 when E_1 is the last risk factor eliminated regardless of the order in which E_2 and E_3 have been already removed. The average AF is 0.227. Note that increasing the number of independent risk factors increases the sequential AF variability. Sequential AFs ranges from 0.22 to 0.33

considering only E_1 and E_2, whereas they ranges from 0.16 to 0.28 when E_3 is also considered.

Table 5.2. Prevalence (P) and relative risk (RR) for three independent risk factors.

Risk factor				
$\boldsymbol{E_1}$	$\boldsymbol{E_2}$	$\boldsymbol{E_3}$	$\boldsymbol{P}$	$\boldsymbol{RR}$
0	0	0	0.125	1.0
0	1	0	0.125	2.0
0	0	1	0.125	2.0
0	1	1	0.125	4.0
1	0	1	0.125	2.0
1	1	1	0.125	4.0
1	0	1	0.125	4.0
1	1	1	0.125	6.0

Let E_1 and E_2 be two correlated risk factors with a 0.5 prevalence and a RR equal to 2 as shown in table 5.3. Sequential AFs for E_1 are 0.375 and 0.208 according as E_1 is eliminated before or after E_2, respectively. The average AF is 0.292.

Table 5.3. Prevalence (P) and relative risk (RR) for two correlated risk factors.

Risk factor			
$\boldsymbol{E_1}$	$\boldsymbol{E_2}$	$\boldsymbol{P}$	$\boldsymbol{RR}$
0	0	0.4	1.0
0	1	0.10	2.0
1	0	0.10	2.0
1	1	0.40	4.0

Again, let E_1, E_2 and E_3 be three correlated risk factors with the same prevalence and RR as before (table 5.4). Sequential AFs for E_1 are 0.3 when

E_1 is eliminated as first risk factor, 0.26 when E_1 is eliminated after having already removed E_2, 0.26 when E_1 is eliminated after E_3, and 0.16 when E_1 is eliminated as last risk factor.

Table 5.4. Prevalence (P) and relative risk (RR) for three independent risk factors.

Risk factor			***P***	***RR***
$\boldsymbol{E}_1$	$\boldsymbol{E}_2$	$\boldsymbol{E}_3$		
0	0	0	0.2375	1.0
0	1	0	0.05	2.0
0	0	1	0.05	2.0
0	1	1	0.125	4.0
1	0	1	0.05	2.0
1	1	1	0.125	4.0
1	0	1	0.125	4.0
1	1	1	0.2375	6.0

Unlike the case of independent risk factors, increasing the number of correlated risk factors leads to reduced sequential AF variability. Here, sequential AFs ranges from 0.208 to 0.375 and from 0.3 to 0.16 for two and three risk factors, respectively.

As outlined in the previous chapter, we propose a different formula for the average AF variance which it includes the variation of average AFs from their total mean across simulations (Ferguson's component) and the average of the variation of sequential AF for across simulations and permutations (internal component). We simulated two classes of datasets to compare the performance of average AF variance estimates between our and Ferguson's formulas.

5.2 Simulated data

Each class of simulated data included four datasets according to different correlation structures, from independence (scenario 1) to strong correlation (scenario 4). The two classes differed in the prevalence and strength of the association between risk factors: the first class had an high risk factor prevalence and modest relative risks; the second class had a low risk factor prevalence and huge relative risks. In particular, we generated scenario 1 for the first class simulating ten independent risk factors, $\mathbf{E}=\{E_1,E_2,\dots,E_{10}\}$, with a prevalence of 0.5, $P(E_1)=P(E_2)=\cdots=P(E_{10})=0.5$, and a relative risk ranging from 1.2 to 1.5. We simulated the binary outcome, $\mathbf{y}=\{y_1,y_2,\dots,y_n\}$, fixing a 0.05 log odds of $\mathbf{y}$ in the unexposed group. We generated the remaining three scenarios increasing the magnitude of the correlation and leaving the number of the risk factors, the prevalence, the relative risk, and the log odds of the outcome in the unexposed group unchanged. Therefore, we generated data considering weak ($\rho=0.4$), moderate ($\rho=0.6$), and strong ($\rho=0.8$) correlation among risk factors (table 5.5).

Table 5.5. Setting of parameters for two classes of simulating data.

Class	**Risk factors**	$P(E_l)$	RR
First			
Scenario 1 – Independence	$E_1,E_2,\dots,E_{10}$	0.5	1.2-1.5
Scenario 2 – Weak correlation ($\rho=0.4$)	$E_1,E_2,\dots,E_{10}$	0.5	1.2-1.5
Scenario 3 – Moderate correlation ($\rho=0.6$)	$E_1,E_2,\dots,E_{10}$	0.5	1.2-1.5
Scenario 4 – Strong correlation ($\rho=0.8$)	$E_1,E_2,\dots,E_{10}$	0.5	1.2-1.2
Second			
Scenario 1 – Independence	$E_1,E_2,\dots,E_{10}$	0.1	5.0-10.0
Scenario 2 – Weak correlation ($\rho=0.4$)	$E_1,E_2,\dots,E_{10}$	0.1	5.0-10.0
Scenario 3 – Moderate correlation ($\rho=0.6$)	$E_1,E_2,\dots,E_{10}$	0.1	5.0-10.0
Scenario 4 – Strong correlation ($\rho=0.8$)	$E_1,E_2,\dots,E_{10}$	0.1	5.0-10.0

The second class of simulated data is the same as the first with the prevalence set to $P(E_1)=P(E_2)=\cdots=P(E_{10})=0.1$ and the relative risk from 5.0 to 10.0 (table 5.5). We set $n=1{,}000$ observations for each dataset.

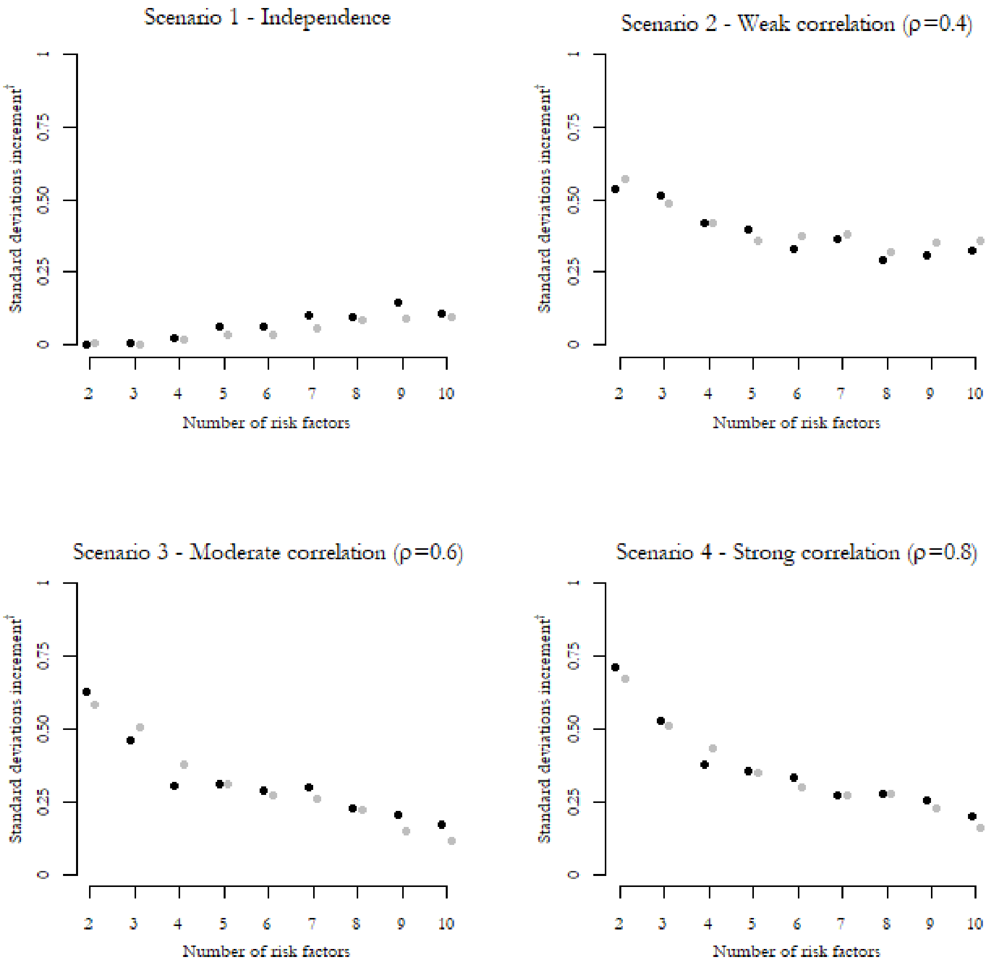
5.3 Method comparisons

We computed variability increment between our and Ferguson's methods to estimate average AF variance (figures 5.1 and 5.2), as follows:

$$1-\frac{SD_{Ferguson}\left({}_{average}AF_l\right)}{SD_{our}\left({}_{average}AF_l\right)},$$

where $SD_{\bullet}\left({}_{average}AF_l\right)$ denote average AF standard deviation for the lth risk factor.

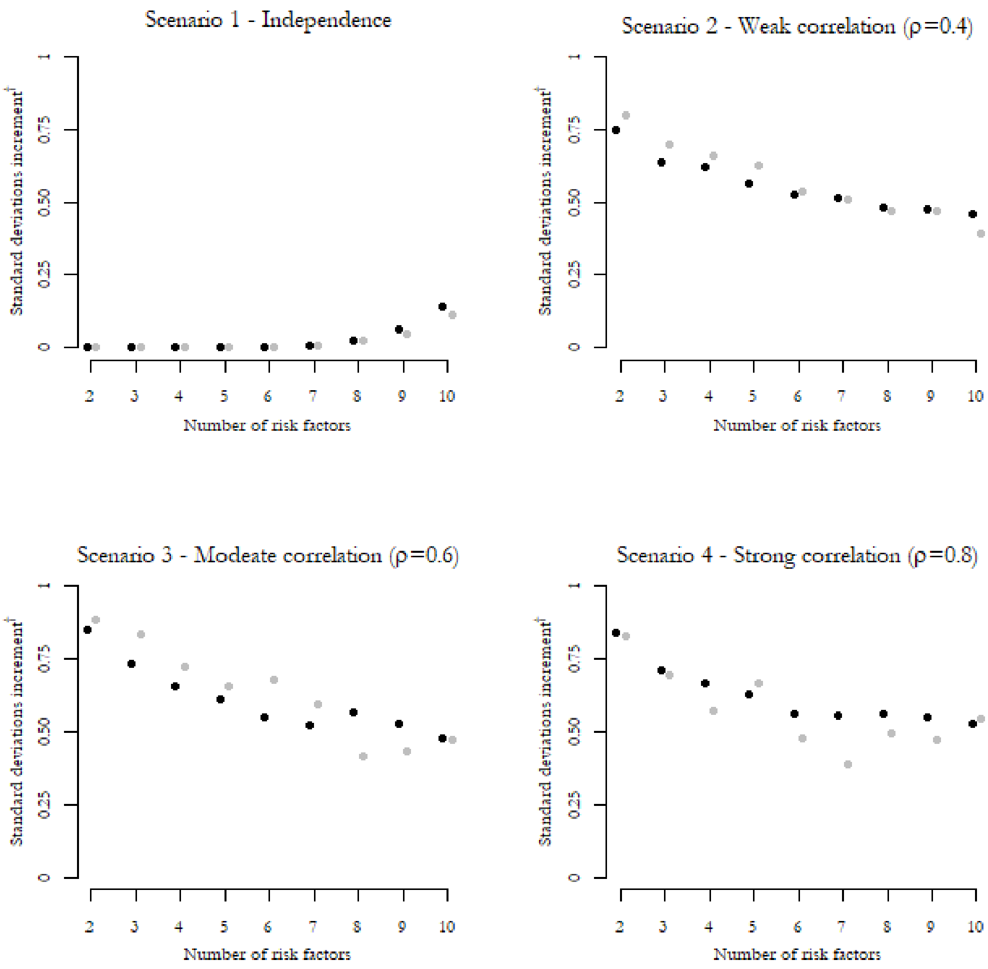
Figure 5.1. Standard deviation increment between our and Ferguson's methods to estimate average AF variance for the first class of simulated data[§].



[§]Data have been simulated for ten risk factors with a prevalence of 0.5, a RR ranging from 1.2 to 1.5 and different correlation structures; [†]Computed by $1-\left(SD_{Ferguson}/SD_{our}\right)$; Black circles indicate risk factor E_1; Gray circles indicate risk factor E_2.

In particular, for each class of simulated data, we reported standard deviation increment for risk factor E_1 (black circles) and risk factor E_2 (gray circles) according to an increasing number of risk factors considered (from two to ten) and different scenarios.

Figure 5.2. Standard deviation increment between our and Ferguson's methods to estimate average AF variance for the second class of simulated data[§].



[§]Data have been simulated for ten risk factors with a prevalence of 0.1, a RR ranging from 5.0 to 10.0 and different correlation structures; [†]Computed by $1-\left(SD_{Ferguson}/SD_{our}\right)$; Black circles indicate risk factor E_1; Gray circles indicate risk factor E_2.

Standard deviation increment became gradually larger increasing the number of independent risk factors for both E_1 and E_2 (figure 5.1; upper-left panel). For example, standard deviation increment for risk factor E_1 ranged from 0.3% (when only E_1 and E_2 were considered) to 29% (when all risk factors were considered). Risk factor E_2 showed a similar pattern. As described in the previous section, increasing the number of independent risk factors increased the sequential AF variability and consequently the total AF variability tended to be large. Conversely, standard deviation increment ranged from 71% to 9%, indicating that the contribution of the internal component decreased with increasing number of correlated risk factors (figure 5.1; upper-right and bottom panels).

The second class of simulated data showed similar trends in standard deviation increment as those observed for the first class (figure 5.2). Increasing the number of independent risk factors led to a larger variability. Standard deviation estimates were similar between two methods up to seven risk factors, whereas the total variability became slightly larger for eight or more independent risk factors (figure 5.2; upper-left panel). Standard deviation increment decreased from 88% to 26% increasing the number of correlated risk factors (figure 5.2; upper-right and bottom panels).

Although in some situations (i.e., for correlated risk factors) the contribution of the internal component could have a substantial relative impact on total AF variability, the absolute standard deviation differences between two methods were very small (<0.15) indicating a limited contribution of our method than the Feguson’s one.

Chapter 6

Application to real data

6.1 Epidemiology of oral cavity cancer

Oral cavity cancer is the 8th most frequent cancer in the world among males and the 14th among females, accounting for nearly 3% of all cancer worldwide [62]. The annual global incidence of oral cavity cancer is estimated at approximately 263,000 cases, and the corresponding number of deaths at 127,000 [63]. Incidence and mortality rates vary widely according to geographical areas. In particular, less developed and few developing countries (i.e., India, Pakistan, Bangladesh, Hong Kong, Singapore, and the Philippines) report higher incidence rates. Conversely, mortality rates are highest in the less developed and developing countries [64, 65]. As with most upper aerodigestive tract cancers, tobacco smoking and alcohol drinking are the major risk factors for oral cavity cancer [66]. Betel-quid and smokeless tobacco chewing play a role in oral carcinogenesis [67, 68]. Other risk factors include diet, hot mate consumption, human papillomavirus (HPV) infection, and oral hygiene [69-72].

Tobacco smoking

The risk of oral cavity cancer increases with the intensity (number of cigarettes, cigars or pipe smoked per day), duration of consumption and lifetime cumulative consumption of tobacco smoking. Moreover, several studies reported a dose-response relationship between intensity and oral cavity cancer risk [73]. Tobacco smoking contains a number of carcinogens known to cause cancers. These carcinogens are derived from various chemical classes such as polycyclic aromatic hydrocarbons, nitrosamines, aromatic amines, volatile hydrocarbons, nitro compounds, and other organic and inorganic compounds [74]. Nicotine is generally accepted as non-carcinogenic, but it may promote cancer by activating signaling pathways facilitating cancer cell growth, angiogenesis, migration, and invasion [75]. Moreover, nicotine can undergo chemical conversions into carcinogenic substances during the process of curing or smoking. The majority of nicotine can be metabolized to cotinine and aldehyde oxidase, and the remaining nicotine may be converted to other metabolites such as nicotine-N-oxide [76, 77]. Nicotine and nitrosamines are implicated in tumor promotion by activating signal transduction pathways that facilitate tumor progression [78].

Alcohol drinking

The risk of oral cavity cancer increases with the daily quantity, duration of consumption and lifetime cumulative consumption of alcohol drinking [73, 79]. However, a large pooled analysis in the International Head and Neck Cancer Epidemiology (INHANCE) consortium found a low and non-significant risk for oral cavity cancer in non-smokers for both frequency and duration of alcohol consumption [80]. The mechanism by which ethanol promotes oral cavity cancer remains unclear and several explanations have been suggested. First, alcohol may act as a solvent facilitating the transport of carcinogens through cellular membranes [81]. Second, ethanol may enhance liver metabolizing activity contributing to the activation of carcinogenic substances (e.g., polycyclic aromatic hydrocarbons or benzopyrene). Other mechanisms include nutritional deficiencies due to chronic alcohol abuse, which may impair some cellular functions such as mitochondrial function and DNA repair system.

Other factors

Other factors are associated with oral cavity carcinogenesis. Briefly, many studies conducted in Asia, where betel consumption is common, assessed the relationship between betel-quid and tobacco smokeless chewing and the risk of oral cavity cancer. Dietary factors (diet low in fruits and vegetables and high in red meat) are associated with oral cavity cancer. Mate consumption (an herbal tea), particularly widespread in South America, increases the risk of oral cavity cancer up to 2-fold. The role of HPV infection in the occurrence of upper aero digestive tract cancers, especially of pharynx and oral cavity, has been widely discussed in literature, confirming a positive association between HPV and the risk of oral cavity cancer. Finally, poor oral and dental hygiene was related to an increase risk of oral cavity cancer [73].

6.2 Case-control data on oral cavity cancer

Data came from a hospital-based case-control study on oral cavity cancer conducted between 1991 and 2009 in the provinces of Milan and Pordenone in Northern Italy and Latina and Rome in Central Italy [51]. Cases were 946 patients aged 18 years or older with incident histologically confirmed oral cavity cancer diagnosis admitted to major general hospitals. The control group included 2492 patients frequency-matched to the cases by sex and age (table 6.1). Controls were admitted to the same network of hospitals as the cases for a wide spectrum of acute, non-neoplastic conditions unrelated to tobacco and alcohol consumption, to known or likely risk factors for oral cavity cancer, or to other conditions associated with long-term diet modification. All patients enrolled in the study signed an informed consent, according to the recommendations of the Board of Ethics of study hospitals.

Trained interviewers administered a structured and validated questionnaire to cases and controls during their hospital stay [82-84]. The questionnaire collected information on socio-demographic characteristics, anthropometric measures, lifetime smoking and alcohol drinking habits, dietary habits related to two years before diagnosis/interview.

Anthropometric measures included self-reported height and weight one year prior to diagnosis/interview and at age 30 and 50 years.

Information on tobacco smoking included smoking status (never, former, or current smokers), daily number of cigarettes, cigar and grams of tobacco pipe smoked during lifetime, age at starting to smoke, duration of the habit, and, for former smokers, age at stopping. The intensity of smoking habit was measured as daily number of cigarettes smoked. Taking into account the different types of smoking, one gram of tobacco pipe was considered as corresponding to one cigarette and one cigar as corresponding to three cigarettes. Current smokers were people who had smoked at least 1 cigarette, cigar, or one gram of tobacco pipe within one year previous to the interview. Former smokers were people who had abstained from any type of tobacco smoking within one year previous to the interview.

Information on alcohol consumption included drinking status (never, former, and current drinkers), daily number of drinks consumed for the most common Italian and Swiss alcoholic beverages (i.e., wine, beer, and spirits which included amari, grappa, whisky, cognac, brandy, etc.), age at starting to drink, duration of alcohol consumption, and, for former drinkers, age at stopping.

Table 6.1. Distribution of 946 oral cavity cancer cases and 2492 controls according to socio-demographic characteristics and selected risk factors. Italy, 1991-2009.

Variable	Cases		Controls	
	n	(%)	n	(%)
Study centre				
Pordenone	494	(52.2)	1053	(42.3)
Milan	348	(36.8)	1001	(40.2)
Rome/Latina	104	(11.0)	438	(17.6)
Sex				
Men	756	(79.9)	1497	(60.1)
Women	190	(20.1)	995	(39.9)
Age (years)				
<55	328	(34.7)	940	(37.7)
55-64	341	(36.1)	778	(31.2)
≥65	277	(29.3)	774	(31.1)
Education (years)				
<7	558	(59.0)	1283	(51.5)
7-11	260	(27.5)	726	(29.1)
≥12	128	(13.5)	438	(19.4)
BMI (Kg/m^2)				
<25	545	(57.6)	1049	(42.1)
25-<30	322	(34.0)	1101	(44.2)
≥30	79	(8.4)	342	(13.7)
Non-alcohol energy intake (kcal/day)				
<1884.6	300	(31.7)	832	(33.4)
1884.6-<2395	289	(30.6)	831	(33.4)
≥2395	357	(37.7)	829	(33.3)
Smoking§				
Never	137	(14.5)	1079	(43.3)
Former	268	(28.3)	764	(30.7)
Current (cigarettes/day)				
<15	175	(18.5)	357	(14.3)
≥15	362	(38.3)	290	(11.6)
Alcohol drinking§				
Never	66	(7.0)	445	(17.9)
Ever (drinks/day)				
<2	197	(20.8)	1023	(41.1)
≥2	681	(72.0)	1019	(40.9)
Red meat intake (servings/week)				
<3	242	(25.6)	908	(36.4)
3-<4.5	271	(28.7)	779	(31.3)
≥4.5	433	(45.8)	805	(32.3)

Variable	Cases		Controls	
	n	(%)	n	(%)
Vegetables intake (servings / week)				
<10.75	456	(48.2)	838	(33.6)
10.75-<15.375	256	(27.1)	834	(33.5)
≥15.375	234	(24.7)	820	(32.9)
Fruit intake (servings / week)				
<14.17	457	(48.3)	840	(33.7)
14.17-<23.375	282	(29.8)	826	(33.2)
≥23.375	207	(21.9)	826	(33.2)
Family history				
No	905	(95.7)	2449	(98.3)
Yes	41	(4.3)	43	(1.7)

§The sum does not add up to the total because of missing values.

Taking into account the different alcohol concentrations, one drink corresponded approximately to 125 ml of wine, 330 ml of beer and 30 ml of spirit (i.e., about 12 gram of ethanol).

A food frequency questionnaire (FFQ) assessed patients' habitual diet in two years before diagnosis / interview. The FFQ included information on weekly intake of 78 foods or recipes according to the following 6 sections: (i) milk, hot beverages and sweeteners; (ii) bread, cereals and first courses; (iii) second courses (e.g., meat and other main dishes); (iv) side dishes (i.e., vegetables); (v) fruits; (vi) sweets, desserts and soft drinks. For 40 out 78 food items, the portion size was defined in "natural" units (e.g., 1 teaspoon of sugar, 1 egg, 1 apple, etc.), whereas for the remaining ones, it was defined as small, average, or large with the help of pictures. Seasonal variation in fruit and vegetable consumption was also considered to account for the fluctuations in food intake.

6.3 Attributable fraction for oral cavity cancer

We estimated average AF and Bruzzi's AF estimates for oral cavity cancer. We set a 81×10^{-5} prevalence of oral cavity cancer [52] to adjust average AF estimates for case-control design. The final model included smoking, alcohol drinking, red meat intake, vegetables intake, fruit intake, and family history of oral cavity cancer as risk factors; study centre, sex, age, years of education, BMI and non-alcohol energy intake as adjusting factors.

Table 6.2. Odds ratios (ORs) and corresponding 95% confidence intervals (CIs), Bruzzi's attributable fraction (AF) estimates, and average AF estimates for oral cavity cancer according to selected risk factors. Italy, 1991-2009.

Risk factor	OR (95% CI)[§]	Bruzzi's AF	Average AF (95% CI)
Smoking[a]			
Never	Ref		
Former	2.04 (1.58; 2.63)		
Current (cigarettes/day)			
<15	2.82 (2.14; 3.72)		
≥15	6.40 (4.89; 8.42)	0.60	0.34 (0.27; 0.41)
Alcohol drinking			
Never	Ref		
Ever (drinks/day)			
<2	1.15 (0.83; 1.60)		
≥2	2.70 (1.93; 3.81)	0.51	0.27 (0.17; 0.37)
Red meat intake (servings/week)			
<3	Ref.		
3-<4.5	1.08 (0.87; 1.36)		
≥4.5	1.42 (1.14; 1.78)	0.14	0.06 (0.01; 0.12)
Vegetables intake (servings/week)			
<10.75	1.74 (1.39; 2.18)		
10.75-<15.375	1.11 (0.89; 1.40)		
≥15.375	Ref	0.24	0.11 (0.06; 0.17)
Fruit intake (servings/week)			
<14.17	1.35 (1.07; 1.70)		
14.17-<23.375	1.22 (0.97; 1.55)		
≥23.375	Ref.	0.18	0.08 (0.02; 0.15)
Family history			
No	Ref.		
Yes	2.40 (1.44; 3.99)	0.02	0.009 (-0.001; 0.02)
Joint		0.88	0.88 (0.78; 0.98)

[§] Adjusted for study centre, age (<55; 55-64; ≥65 years), education (<7; 7-11; ≥12 years), BMI (<25; 25-<30; ≥30 Kg/m^2), tertiles of non-alcohol energy intake (kcal/day).

Cases had lower years of education and had more frequently a BMI <25 kg/m^2 than controls. Cases were more likely current smokers and drinkers and consumed more frequently red meat than controls. Moreover, cases consumed vegetables and fruit less frequently than controls. Finally, cases had more likely a first-degree relative with oral cavity cancer than controls (table 6.1).

Smoking and alcohol drinking were strongly associated with oral cavity cancer risk. In particular, former smokers had approximately a 2-fold higher oral cavity cancer risk than never smokers (OR=2.04 - 95% CI: 1.58; 2.63 - table 6.2). People who smoked up to 15 cigarettes per day had a higher oral cavity cancer risk with an OR of 2.82 (95% CI: 2.14; 3.72), whereas people who smoked more than 15 cigarettes per day had more than 6-fold higher risk (OR=6.40 - 95% CI: 4.89; 8.42). Likewise, heavy drinkers (≥2 drinks/day) had a higher oral cavity cancer risk (OR=2.70; 95% CI: 1.93; 3.81) than abstainers. High red meat consumption (≥4.5 servings/week) increased the risk of oral cavity cancer with an OR of 1.42 (95% CI: 1.14; 1.78). People who had a low intake of both vegetables and fruit had a higher risk of oral cavity cancer compared with people who had a high intake. In particular, the ORs for a low vegetables intake (<10.75 servings/week) and fruit intake (<14.17 servings/week) were 1.74 (95% CI: 1.39; 2.18) and 1.35 (95% CI: 1.07; 1.72), respectively.

Eighty-eight percent of oral cavity cases were attributable to risk factors considered (table 6.2). In particular, the average AF for smoking was 0.34 (95% CI: 0.27; 0.41), indicating that 34% of oral cavity cases would not have occurred if smoking were randomly removed from the population over all possible risk factor removal orders. The average AF for alcohol drinking was 0.27 (95% CI: 0.17; 0.37), whereas average AFs for high red meat intake, low vegetables intake and low fruit intake were 0.06 (95% CI: 0.01; 0.12), 0.11 (95% CI: 0.06; 0.17), and 0.08 (95% CI: 0.02; 0.15), respectively. The average AF for family history of oral cavity cancer was 0.009 (95% CI: -0.001; 0.02). According to the Bruzzi's method, AFs were 0.60 for smoking, 0.51 for alcohol drinking, 0.14 for high red meat intake, 0.24 for low vegetables intake, 0.18 for low fruit intake, and 0.02 for family history (table 6.2).

6.4 Epidemiology of breast cancer

Breast cancer is the second most common cancer in the world and the most common frequent cancer among women with an estimated 1.67 million new cases diagnosed in 2012 (25% of all cancers) [63]. It is the most common cancer in women both in more and less developed countries with slightly more cases in less developed (883,000 cases) than in more developed (794,000 cases) countries. Incidences rates vary nearly 4-fold across the world regions, with rates ranging from 27 per 100,000 new cases in Middle Africa and Eastern Asia to 92 per 100,000 in Northern America. Breast cancer ranks as the fifth cause of death from cancer overall (522,000 deaths), and while it is the most frequent cause of cancer death in women in less developed countries (324,000 deaths, 14.3% of total), it is the second cause of cancer death in more developed countries (198,000 deaths, 15.4% of total) after lung cancer. The range in mortality rates between world regions is less than that for incidence because of the more favorable survival of breast cancer in (high-incidence) developed regions, with rates ranging from 6 per 100,000 deaths in Eastern Asia to 20 per 100,000 in Western Africa [63]. Established risk factors for breast cancer include genetics, race/ethnicity, overweight, alcohol, diet, and reproductive factors.

Genetics

About 5% to 10% of breast cancers are thought to be hereditary, caused by abnormal genes passed from parent to child. Most inherited cases of breast cancer are associated with two abnormal genes: BRCA1 (breast cancer gene one) and BRCA2 (breast cancer gene two). The function of the BRCA genes is to repair cell damage and keep breast, ovarian, and other cells growing normally. Abnormalities or mutations in BRCA genes may increase breast, ovarian and other cancer risk [85].

Race/ethnicity

White women are slightly more likely to develop breast cancer than African, America, Hispanic, and Asian women. But African American women are more likely to develop more aggressive and more advanced-stage breast cancer that is diagnosed at a young age [86].

Overweight

Overweight and obesity are involved in breast carcinogenesis. Overweight and obese women have a higher breast cancer risk compared to women who maintain a healthy weight, especially after menopause. This higher risk is because fat cells make estrogen; extra fat cells mean more estrogen in the body, and estrogen can make hormone-receptor-positive breast cancer develop and grow [87].

Alcohol drinking

Alcohol intake is consistently associated with breast cancer risk. Alcohol can increase levels of estrogen and other hormones associated with hormone-receptor-positive breast cancer. Alcohol also may increase breast cancer risk by damaging DNA in cells [88].

Reproductive factors

Age at menarche

Age at menarche has been consistently associated with breast cancer risk. Several mechanism have been proposed. Menarche marks the onset of the mature hormonal milieu, that is cyclic hormonal changes that result in ovulation, menstruation, and cellular proliferation in the breast [89]. The earlier the age at menarche, the earlier a young woman starts experiencing increased steroid hormone levels. An earlier age at menarche also has been related to an earlier onset of regular ovulatory cycles. In addition, women with an earlier menarche may have higher circulating estrogen levels for a number of years afterward [90].

Age at first birth and parity

Overall, nulliparous women have a higher breast cancer risk than parous women [91]. Moreover, women who have not had a full-term pregnancy or have their first child after age 30 have a higher risk of breast cancer compared to women who gave birth before age 30. The biological mechanisms behind this have been studied extensively. The ductal system of the breast undergoes profound changes from birth through adulthood. After menarche but prior to a first pregnancy, the breast contains relatively undifferentiated ducts and associated alveolar buds. Differentiation of the glandular epithelial cells takes place gradually, culminating in terminally differentiated tissue. These changes occur largely after a first full-term pregnancy and, to a lesser extent, after subsequent pregnancy. When the first pregnancy occurs at an early age, fewer cells are likely to have been initiated and the period of protection, afforded by the terminal differentiation of the breast glandular epithelium, covers a larger fraction of the woman's remaining lifetime [92].

Breastfeeding

Breastfeeding can lower breast cancer risk, especially if a woman breastfeeds for longer time. Breastfeeding may result in further terminal differentiation of the breast epithelium, thus making it more resistant to carcinogenic change. Additionally, breastfeeding delays the post-pregnancy reestablishment of the menstrual cycle and hence may reduce the risk [93].

Age at menopause

The positive relationship between age at menopause and breast cancer risk is well established. The reduction in risk associated with an earlier menopause is due to the cessation of ovarian function and the consequent reduction in circulating steroid hormone level [93].

Oral contraceptives and hormonal replacement therapy

Users of oral contraceptives and hormonal replacement therapy (HRT) have a higher risk of breast cancer. Hormones play a central role in the aetiology of breast cancer. For instance, after menopause, adipose tissue is the major source of estrogen, and obese postmenopausal women have both higher levels of endogenous estrogen and higher breast cancer risk [94, 95]. Further, estrogens and progesterone promote mammary tumors in animal [96].

Diet

Fruit and vegetable consumption is related to a decreased breast cancer risk [87, 97, 98]. Moreover, data support a protective role for vitamin A and carotenoids in the aetiology of breast cancer [99]. Inadequate folate levels may result in abnormal DNA synthesis and disrupted DNA repair and hence may influence breast cancer risk. Several studies suggest that adequate folate levels could reduce breast cancer risk [100-103].

6.5 Case-control data on breast cancer

Data came from an hospital-based case-control study on breast cancer conducted between 1991 and 2006 in the provinces of Milan, Genoa, Pordenone, and Forlì in Northern Italy, Latina and Rome in Central Italy, and Naples in Southern Italy, on a total of 2569 cases and 2588 controls (table 6.3) [51]. Cases were women aged 18 years or older with incident histologically confirmed breast cancer diagnosis admitted to major general hospitals. Controls were women aged 18 years or older admitted to the same network of hospitals as the cases for a wide spectrum of acute non-neoplastic and non-gynecologic diseases, and for other conditions unrelated to known or likely risk factors for breast cancer. All women enrolled in the study signed an informed consent, according to the recommendations of the Board of Ethics of study hospitals.

Trained interviewers administered a structured and validated questionnaire to cases and controls during their hospital stay [82-84]. The questionnaire collected information on socio-demographic characteristics, anthropometric measures, lifetime smoking and alcohol drinking habits, dietary habits related to two years before diagnosis/interview, reproductive factors, and family history of cancer. Information on anthropometric measures, smoking and alcohol drinking habits, and diet was the same as the questionnaire used for oral cavity cancer.

In a detailed section, women were asked to report their menstrual and reproductive histories, including age at menarche, menopausal status (menopause was defined as lack of menstruation for at least 12 months), type of menopause (natural or surgical), age at menopause, number of births and abortions, and age at each child delivery. Information was specifically collected on lifelong use of oral contraceptives (OCs) and hormonal replacement therapy (HRT), including age at start and duration of each episode of use.

Women were asked to report their family history of cancer in first-degree relatives (parents, siblings, and children), including the site of the tumor, type of relatives, and age at diagnosis.

6.6 Attributable fraction for breast cancer

We estimated average AF and Bruzzi's AF estimates for breast cancer. We set a 2019×10^{-5} prevalence of breast cancer [52] accounting for case-control data structure to estimate average AFs.

Table 6.3. Distribution of 2569 breast cancer cases and 2588 controls according to socio-demographic characteristics and selected risk factors. Italy, 1991-2006.

Variable	Cases		Controls	
	n	(%)	n	(%)
Study centre				
Pordenone	1046	(40.7)	1015	(39.2)
Milan	585	(22.8)	623	(24.1)
Genoa	290	(11.3)	310	(12.0)
Forlì	212	(8.3)	213	(8.2)
Naples	258	(10.0)	249	(9.6)
Rome/Latina	178	(6.9)	178	(6.9)
Age (years)				
<45	470	(18.3)	472	(18.2)
45-54	772	(30.1)	694	(26.8)
55-64	799	(31.1)	802	(31.0)
≥65	528	(20.6)	620	(24.0)
Education (years)				
<7	1273	(49.6)	1592	(61.5)
7-11	714	(27.8)	642	(24.8)
≥12	582	(22.7)	354	(13.7)
BMI (Kg/m^2)				
<25	1399	(54.5)	1353	(52.3)
25-<30	824	(32.1)	844	(32.6)
≥30	346	(13.5)	391	(15.1)
Smoking				
Never	1684	(65.6)	1759	(68.0)
Former	344	(13.4)	252	(9.7)
Current (cigarettes/day)				
<15	324	(12.6)	347	(13.4)
≥15	217	(8.5)	230	(8.9)
Alcohol drinking§				
Never	769	(29.9)	910	(35.2)
Ever (drinks/day)				
<2	1093	(42.6)	1009	(39.0)
≥2	703	(27.4)	666	(25.7)

Variable	Cases		Controls	
	n	(%)	n	(%)
Age at menarche§				
<14 years	1717	(66.8)	1636	(63.2)
≥14 years	848	(33.0)	949	(36.7)
Parity				
0	401	(15.6)	380	(14.7)
1	584	(22.7)	494	(19.1)
2	968	(37.7)	909	(35.1)
3	406	(15.8)	489	(18.9)
≥4	210	(8.2)	316	(12.2)
Breastfeeding				
No or <3 months	1114	(43.4)	1095	(42.3)
≥3 months	1455	(56.6)	1493	(57.7)
Hormonal replacement therapy				
No	2375	(92.5)	2396	(92.6)
Yes	194	(7.6)	192	(7.4)
Oral contraceptive use				
No	2208	(86.0)	2298	(88.8)
Yes	361	(14.1)	290	(11.2)
Family history				
No	2309	(89.9)	2465	(95.3)
Yes	260	(10.1)	123	(4.8)

§The sum does not add up to the total because of missing values.

The final model included alcohol drinking, parity, breastfeeding, use of OCs, and family history of breast cancer as risk factors; study centre, age, years of education, smoking, age at menarche, and use of HRT as adjustment factors.

Cases had higher years of education than controls. Cases were more likely drinkers and nulliparous or with one child that controls. Women with breast cancer used more likely OCs and had a first-degree relative with breast cancer. Smoking, BMI, age at menarche and HRT use were similar in cases and controls (table 6.3).

Women who consumed alcohol had a higher risk of breast cancer (OR=1.25 for those who consumed <2 drinks/day – 95% CI: 1.10; 1.43 and OR=1.27 for those who consumed ≥ drinks/day – 95% CI: 1.10; 1.48 – table 6.4). Parity increased breast cancer risk. In particular, nulliparous women (OR=1.54; 95% CI: 1.18; 2.00), women with one child (OR=1.72; 95% CI: 1.36; 2.16), and women with two children (OR=1.52; 95% CI: 1.24-1.88) had a significantly higher breast cancer risk compared with women with 4 or more children. The risk of breast cancer increased for women who used OCs with an OR of 1.16 (95% CI: 0.96-1.40) compared to non-users. Family history of breast cancer was strongly associated with the risk of breast cancer. Women with first-degree relatives with breast cancer had an higher risk (OR=2.28; 95% CI: 1.82-2.86) compared with women who did not have a family history of breast cancer. Age at menarche and breastfeeding were not associated with breast cancer. (data not shown).

Table 6.4. Odds ratios (ORs) and corresponding 95% confidence intervals (CIs), Bruzzi's attributable fraction (AF) estimates, and average AF estimates for breast cancer according to selected risk factors. Italy, 1991-2006.

Risk factor	OR (95% CI)§	Bruzzi's AF	Average AF (95% CI)
Alcohol drinking			
Never	Ref		
Ever (drinks/day)			
<2	1.25 (1.10; 1.43)		
≥2	1.27 (1.10; 1.48)	0.15	0.12 (0.06; 0.18)
Parity			
0	1.54 (1.18; 2.00)		
1	1.72 (1.36; 2.16)		
2	1.52 (1.24; 1.88)		
3	1.16 (0.94; 1.47)		
≥4	Ref	0.32	0.27 (0.16; 0.39)
Breastfeeding			
No or <4 months	1.11 (0.97; 1.27)		
≥4 months	Ref	0.04	0.04 (-0.02; 0.10)
Oral contraceptive use			
No	Ref		
Yes	1.16 (0.96; 1.40)	0.02	0.01 (-0.01; 0.03)
Family history			
No	Ref		
Yes	2.28 (1.82; 2.86)	0.05	0.04 (0.03; 0.06)
Joint		0.49	0.49 (0.35; 0.63)

§Adjusted for study centre, age (<45; 45-54; 55-64; ≥65 years), education (<7; 7-11; ≥12 years), BMI (<25; 25-<30; ≥30 Kg/m^2), smoking (never; former; current <15; current ≥15 cigarettes/day), age at menarche (<14; 14≥ years), use of HRT.

The joint AF was 0.49 (95% CI: 0.35; 0.63 – table 6.4) indicating that approximately half of breast cancer cases would not have occurred if all risk factors were eliminated simultaneously from the population. In particular, the greatest fraction of breast cancer cases was attributable to parity with an average AF of 0.27 (95% CI: 0.16; 0.39). Alcohol drinking accounted for 12% (95% CI: 6%; 18%) of breast cancer cases. The impact of the remaining risk factors was lower with average AFs of 0.04 (95% CI: -0.02; 0.10) for breastfeeding (No or <4 months), 0.01 (95% CI: -0.01; 0.03) for OCs, and 0.04 (95% CI: 0.03; 0.06) for family history (table 6.4). Attributable fractions of breast cancer cases for risk factors considered according to Bruzzi's formula were 0.15 for alcohol, 0.32 for parity, 0.04 for breastfeeding, 0.02 for OCs and 0.05 for family history.

Chapter 7

Conclusion

7.1 Discussion

Preventive strategies have to take into account the magnitude of risk factors and their prevalence in the population for which the intervention is planned. The AF provides a useful tool to address this issue. Disease aetiology involves multiple risk factors that may act simultaneously on the occurrence of disease and the problem of apportioning exposure-specific contributions in a population exposed to multiple risk factors is the primary interest. There is an intensive literature on the topic of estimating individual fractions in a population exposed to various risk factors.

Levin proposed the concept of AF [2] that describes the proportion of cases that could be attributable to eliminating a risk factor from the population. The original formula ignored the presence of other factors (i.e., other risk factors that may act together to cause disease, adjustment variables or confounders), considering the AF as a univariate parameter. This unadjusted estimator is generally biased. Water discussed the conditions under which unadjusted AFs differ from adjusted ones [6]. The adjusted AFs quantify the effect of one after controlling for other factors. Stratification and modeling approach are the main approaches. The Mantel-Haenszel approach, based on stratification, allows one to control for confounders but not for effect modifications. The weighted-sum approach, also based on stratification, allows one to control both for confounders and effect modifications, but bias in estimating the AF occurs when the data are sparse [21]. The approach based on regression model is more flexible and general [5, 10, 22]. It includes stratification approaches as special cases and provides a unified framework for estimation. However, the sum of individual AFs usually exceeds the joint AF and in some situation might be more than 1. Adjusted AFs should not be used for partitioning the joint risk into exposure-specific contributions.

Because of a similar problem in game theory, Cox Jr. [14] and later Eide and Gefeller [12] suggested a solution to estimate the individual shares attributable to multiple risk factors by calculating the sequential and average AFs. The epidemiological problem of partitioning the joint AF into individual contributions for each risk factor is formally analogous to the economic problem of dividing the profit among several players of different companies that act together in a coalition. Game-theoretic results on "fair" allocation rules have been used to develop sequential and average AFs as a "reasonable procedure of partitioning the joint risk in epidemiology". Sequential AF is the AF for eliminating a risk factor in a particular order from the population. It quantifies the additional effect of one risk factor after the preceding risk factors have already been removed in a specified order from the population. The sequential AFs depend on the order in which risk factors are removed. Average AF overcome this shortcoming by averaging sequential AFs for a risk factor over all possible orders by which risk factors can be removed from the population. Average AFs quantify the additional effect of one risk factors after the preceding factors selected randomly have already been removed from the population. Sequential and average AFs satisfy some mathematical properties such as symmetry, marginal rationality and internal marginal rationality. In addition, these parameters share another nice property, they guarantee that the individual contributions sum up to the joint AF (component-additivity property).

Although these parameters overcome the mathematical problem of partitioning the joint effect ascribable to each risk factor, some considerations are required. Sequential AFs are meaningful numbers from a public health preventive perspective, but average AFs may not be so meaningful. Indeed, sequential AFs indicate what the effect would be for a particular intervention order (i.e., we first will do a smoking campaign, then an alcohol campaign, and so on until the last campaign will be performed). Average AFs, instead, might not represent the actual proportion of cases caused by each risk factor as average AFs assume that risk factors are removed in a random order [104]. In some circumstances, the assumption of random removal order can be implausible. Taking an epidemiological view, risk factors are not all equally modifiable; some risk factors are easier to target via public health interventions. Suppose, for example, that the disease of interest is myocardial infarction and the analyzed risk factors are tobacco smoking, hypertension and hypercholesterolemia. The problem is that the probability of removing the three risk factors in a given order (for example smoking-hypercholesterolemia-hypertension) could be different from the probability of any other order (for example hypertension-smoking-hypercholesterolemia), whereas the average AF assume that all orders have the same probability of occurring [47].

In the original notation, sequential and average AFs was designed for prospective studies. Case-control studies differ from cohort studies in that they sampled diseased (cases) and non-diseased (controls) subjects rather than exposed and unexposed subjects. Thus, the ratio of controls to cases in the sample is fixed a priori and the resulting AF estimates will be biased. Several methods to estimate AFs accounting for case-control data structure have been developed [10, 22, 23]. Ferguson and colleagues proposed an interesting approach to estimate sequential and average AFs in case-control studies. This method consists in weighting the likelihood function of the model, used to estimate sequential and average AFs, for the disease prevalence [16]. Although this method can be easily applied, an important issue regarding the selection process of the cases and controls should be considered. In case-control studies, it is often difficult to ensure that cases and controls are a sample of the same source of population [28]. The assumption that there are no factors influencing the selection of controls other than the prevalence of disease used to weight the regression model is crucial [50].

The problem of estimating confidence intervals for AFs has been intensively researched. Asymptotic approximation and simulation are the main approaches. Variance estimation for AFs is a complex task because it involves covariances between relative risk estimates and risk factor prevalence that are related implicitly through score equations. Benichou and Gail derived estimates of standard errors for AFs using an extension of the delta method for implicitly related random variables [15, 54]. However, their computational formulas are complicated and difficult to implement. Simulation methods, also known as Monte Carlo simulation, do not make any distributional assumption. The Monte Carlo method is a computer-based approach for constructing variance parameter via simulation in place of the theoretical analysis. The idea is the generation of AFs using simulated datasets that are similar to the experimental one, but each one with different random noise normally distributed. Llorca and Delgado-Rodriguez compared several procedures to estimate confidence intervals for AF under different scenarios, discussing their asymptotic behavior, strengths, and limitations [105]. When risk factor prevalence is low, the delta method tends to fail. If the AF is close to zero or one, asymptotic normality cannot be assumed and log-transformed confidence intervals are preferred. Generally, Monte Carlo confidence intervals showed the more accurate estimates. As has been shown by Efron and Tibshirani, the delta method “can be viewed as approximation to the Monte Carlo estimate of variance” [106].

In this work, we proposed an alternative method to estimate average AF confidence intervals. Our approach is a modification of the Ferguson's method that is based on Monte Carlo simulation. Our method accounts for sequential AF variability on the total AF variability. We compared our and Ferguson's methods to estimate average AF variance using simulated data. Standard deviation increment (i.e., the relative difference between standard deviations of our method and the Ferguson's one) became gradually larger with increasing number of independent risk factors. Conversely, standard deviation increment decreased with increasing number of correlated risk factors. Although the contribution of our method on the total AF variability could have a substantial relative impact (up to 88%), the absolute standard deviation differences are very small indicating a limited contribution of our method.

Ferguson and colleagues proposed the "averisk" R package [58] to estimate average AFs and corresponding confidence intervals for a set of risk factors in both prospective and case-controls studies. This package allows one to consider either binary or ordinal risk factors. The "averisk" R package, however, yielded biased estimates when two risk factors were considered. We analyzed the architecture of "averisk" and found the bug in the code. During the third year of this Ph.D., we proposed to the authors some code to estimate average AF correctly (appendix A.4).

Finally, although AF for continuous risk factors is well defined, the available statistical software allows one to consider only binary or categorical risk factors. Although dichotomizing a continuous risk factor, such as blood pressure, into two (or more) categories represents a practical solution, the resulting estimated AF will be probably underestimate the effect of optimal blood pressure control on the disease risk.

7.2 Conclusions

Sequential and average AFs are useful tools to apportion exposure-specific contributions in a population exposed to multiple risk factors. Sequential and average AFs share some mathematical properties such as component-additivity, symmetry, marginal rationality, and internal marginal rationality. Average AFs, however, do not represent the actual amount of disease ascribable for each risk factors because they assume that risk factors are removed from the population in a random order. Nevertheless, average AFs could be useful parameters to estimate the average burden of disease for each risk factors across all possible removal orders.

In this work, we proposed an alternative approach to estimate the average AF confidence interval accounting for sequential AF variability on the total AF one. We compared the performance between our and Fergusons' methods to estimate AF variance. Although our method could have a relative impact on total AF variability, the absolute standard deviation differences suggest a limited contribution of our method. This final topic should be further analyzed.

Appendix

A.1

MacMahon and Pugh attributable fraction formula is:

$$\frac{P(D)-P(D|\bar{E})}{P(D)}. \tag{1}$$

Since

$$\begin{aligned} P(D) &= P(E)\cdot P(D|E)+P(\bar{E})\cdot P(D|\bar{E})= \\ &= P(E)\cdot P(D|E)+\left[1-P(E)\cdot P(D|\bar{E})\right], \end{aligned}$$

substituting for $P(D)$ in the equation (1)

$$\begin{aligned} &\frac{P(E)\cdot P(D|E)+\left[1-P(E)\right]\cdot P(D|\bar{E})-P(D|\bar{E})}{P(E)\cdot P(D|E)+\left[1-P(E)\right]\cdot P(D|\bar{E})}= \\ &=\frac{P(E)\cdot\left[P(D|E)-P(D|\bar{E})\right]}{P(E)\cdot\left[P(D|E)-P(D|\bar{E})\right]\cdot P(D|\bar{E})} \end{aligned} \tag{2}$$

Since

$RR=\dfrac{P(D|E)}{P(D|\bar{E})}$ and $P(D|E)=RR\cdot P(D|\bar{E})$

substituting for $P(D|E)$ in equation (2)

$$\begin{aligned} &\frac{P(E)\cdot\left[RR\cdot P(D|E)-P(D|\bar{E})\right]}{P(E)\cdot\left[RR\cdot P(D|E)-P(D|\bar{E})\right]+P(D|\bar{E})}= \\ &=\frac{P(E)\cdot P(D|\bar{E})\cdot[RR-1]}{P(E)\cdot P(D|\bar{E})\cdot[RR-1]+P(D|\bar{E})} \end{aligned} \tag{3}$$

Dividing by $P(D|\bar{E})$ in equation (3)

$$\frac{(RR-1)\cdot P(E)}{(RR-1)\cdot P(E)+1}.$$

which is the attributable fraction formula proposed by Levin.

A.2

The type II strategy allows adjustment for one (or more) variable. Formally, the adjustment is represented by a stratum variable C with K level, $C = c_1, c_2, \ldots, c_K$. By definition of the case-load weighting and of AF:

$$1 - {}_{adj}AF = \sum_{k=1}^{K} w_k \cdot (1 - AF_k) = \sum_{k=1}^{K} P(C = c_k \mid D) \cdot \frac{P(D \mid \bar{E}, C = c_k)}{P(D \mid C = c_k)}.$$

Therefore:

$$1 - {}_{adj}AF = \frac{1}{P(D)} \cdot \sum_{k=1}^{K} P(C = c_k) \cdot P(D \mid \bar{E}, C = c_k),$$

from Bayes' theorem:

$$\frac{1}{P(D)} \cdot \sum_{k=1}^{K} P(C = c_k) \cdot P(D \mid \bar{E}, C = c_k) \cdot \left\{ P(\bar{E} \mid C = c_k) + P(E \mid C = c_k) \right\} =$$
$$= \frac{1}{P(D)} \cdot \sum_{k=1}^{K} P(D \mid \bar{E}, C = c_k) \cdot \left\{ P(\bar{E}, C = c_k) + P(E, C = c_k) \right\},$$

by definition of conditional probability:

$$\frac{1}{P(D)} \cdot \sum_{k=1}^{K} \left\{ P(D, \bar{E}, C = c_k) + \frac{1}{RR} \cdot P(D, E, C = c_k) \right\},$$

by definition of conditional probability and of the common relative risk RR:

$$\frac{1}{P(D)} \cdot \sum_{k=1}^{K} \left\{ P(D, \bar{E}) + \frac{1}{RR} \cdot P(D, E) \right\} =$$
$$= P(\bar{E} \mid D) + \frac{1}{RR} \cdot P(E \mid D) =$$
$$= 1 - P(E \mid D) \cdot \frac{RR - 1}{RR}.$$

Therefore, if the relative risk is to be assume common to all K strata, the AF estimated through type I adjustment strategy with case-load weighting approach has the same expression as the AF estimated through type II adjustment strategy with Mantel-Heanszel approach.

A.3

The law of the total variance can be proved using the law of the total expectation. First,

$$Var\left(Y\right)=E\left(Y^{2}\right)-E\left(Y\right)^{2}$$

from the definition of variance. Then, applying the law of the total expectation to each term by conditioning on the random variable X :

$$E_{x}\left[E\left(Y^{2}\mid X\right)\right]-E_{x}\left[E\left(Y\mid X\right)\right]^{2}.$$

Now, the conditional second moment of Y can be rewritten in terms of its variance and first moment:

$$E_{x}\left[Var\left(Y\mid X\right)+E\left(Y\mid X\right)^{2}\right]-E_{x}\left[E\left(Y\mid X\right)\right]^{2}.$$

Since the expectation of a sum is the sum of expectations, the terms can now be regrouped:

$$E_{x}\left[Var\left(Y\mid X\right)\right]+\left\{E_{x}\left[E\left(Y\mid X\right)^{2}\right]-E_{x}\left[E\left(Y\mid X\right)\right]^{2}\right\}.$$

Finally, the term $\left\{E_{x}\left[E\left(Y\mid X\right)^{2}\right]-E_{x}\left[E\left(Y\mid X\right)\right]^{2}\right\}$ is the variance of the conditional expectation $E\left(Y\mid X\right)$, and then:

$$E_{x}\left[Var\left(Y\mid X\right)\right]+Var_{x}\left[E\left(Y\mid X\right)\right].$$

A.4

We analyzed the “averisk” R package and discovered a bug in the average AFs estimates when the user considered two risk factors.

The problem lied in the “create_frame” function. This function creates all possible permutations of removal order. In particular, the following code:

```
if(n==1) return(matrix(c(1,1),nrow=1)),
```

yielded only one out of two possible removal orders.

We suggested to replace the code above, as follows:

```
if(n==1) return(matrix(c(1,1,0,1),nrow=2,byrow=F)).
```

In a updated version of the “averisk” R package, published 03-20-2017, the authors modified some bugs in the package including our code.

References

1. Uter, W. and A. Pfahlberg, *The concept of attributable risk in epidemiological practice.* Biometrical J, 1999. **41**(8): p. 985-93.
2. Levin, M.L., *The occurence of lung cancer in man.* Acta Unio Internationalis contra Cancrum 1959. **9**(3): p. 531-41.
3. Miettinen, O.S., *Proportion of disease caused or prevented by a given exposure, trait or intervention.* Am J Epidemiol, 1974. **99**(5): p. 325-32.
4. Walter, S.D., *The estimation and interpretation of attributable risk in health research.* Biometrics, 1976. **32**(4): p. 829-49.
5. Deubner, D.C., et al., *Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia.* Am J Epidemiol, 1980. **112**(1): p. 135-43.
6. Walter, S.D., *Prevention for multifactorial diseases.* Am J Epidemiol, 1980. **112**(3): p. 409-16.
7. Rothman, K.J., *Causes.* Am J Epidemiol, 1976. **104**(6): p. 587-92.
8. Whittenmore, A.S., *Statistical methods for estimating attributable risk from retrospective data.* Stat Med, 1982. **1**(3): p. 229-43.
9. Greenland, S., *Bias in methods for deriving standardized morbidity ratio and attributable fraction estimates.* Stat Med, 1984. **3**(2): p. 131-41.
10. Bruzzi, P., et al., *Estimating the population attributable risk for multiple risk factors using case-control data.* Am J Epidemiol, 1985. **122**(5): p. 904-14.
11. Benichou, J., *A review of adjusted estimators of attributable risk.* Stat Methods Med Res, 2001. **10**(3): p. 195-216.
12. Eide, G.E. and O. Gefeller, *Sequential and average attributable fractions as aids in the selection of preventive strategies.* J Clin Epidemiol, 1995. **48**(5): p. 645-55.
13. Kruskal, W., *Relative importance by averaging over orderings.* Am Stat, 1987. **41**(1): p. 6-10.
14. Cox, L.A., *A new measure of attributable risk for public health applications.* Managment Sci, 1985. **31**(7): p. 800-13.
15. Benichou, J. and M.H. Gail, *A Delta Method for Implicitly Defined Random-Variables.* American Statistician, 1989. **43**(1): p. 41-44.
16. Ferguson, J., et al., *Estimating average attributable fractions with confidence intervals for cohort and case-control studies.* Stat Methods Med Res, 2016.
17. MacMahon, B. and T.F. Pugh, *Epidemiology Principles and Methods.* Boston, MA, Little, Brown and Company, 1970.
18. Leviton, A., *Letter: Definitions of attributable risk.* Am J Epidemiol, 1973. **98**(3): p. 231.
19. Walter, S.D., *The distribution of Levin's measure of attributable risk.* Biometrika, 1975. **62**(2): p. 371-4.
20. Walter, S.D., *Calculation of attributable risks from epidemiological data.* Int J Epidemiol, 1978. **7**(2): p. 175-82.

21. Benichou, J., *Methods of adjustment for estimating the attributable risk in case-control studies: a review.* Stat Med, 1991. **10**(11): p. 1753-73.

22. Greenland, S. and K. Drescher, *Maximum likelihood estimation of the attributable fraction from logistic models.* Biometrics, 1993. **49**(3): p. 865-72.

23. Coughlin, S.S., J. Benichou, and D.L. Weed, *Attributable Risk-Estimation in Case-Control Studies.* Epidemiologic Reviews, 1994. **16**(1): p. 51-64.

24. van der Laan, M.J., *Estimation based on case-control designs with known prevalence probability.* Int J Biostat, 2008. **4**(1): p. Article 17.

25. Gefeller, O., *Comparison of adjusted attributable risk estimators.* Stat Med, 1992. **11**(16): p. 2083-91.

26. Miettinen, O.S., *Components of the crude risk ratio.* Am J Epidemiol, 1972. **96**(2): p. 168-72.

27. Kleinbaum, D.G., L.L. Kupper, and H. Morgenster, *Epidemiologic Reasearch: Principles and Quantitative Methods.* Lifetime Learning, Belmont, 1982.

28. Mantel, N. and W. Haenszel, *Statistical aspects of the analysis of data from retrospective studies of disease.* J Natl Cancer Inst, 1959. **22**(4): p. 719-48.

29. Greenland, S., *Variance estimators for attributable fraction estimates consistent in both large strata and sparse data.* Stat Med, 1987. **6**(6): p. 701-8.

30. Birch, M.W., *The detection of partial associations, I: the 2 X 2 case.* J Roy Stat Soc. Series B, 1964. **26**(2): p. 313-24.

31. Breslow, N.E., *Odds ratio estimators when the data are sparce.* Biometrika, 1981. **68**(1): p. 73-84.

32. Greenland, S. and H. Morgenstern, *Morgenstern corrects a conceptual error.* Am J Public Health, 1983. **73**(6): p. 703-4.

33. Ejigou, A., *Estimation of attributable risk in presence of confounding.* Biom J, 1979. **21**(2): p. 155-65.

34. Sturmans, F., P.G. Mulder, and H.A. Valkenburg, *Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage.* Am J Epidemiol, 1977. **105**(3): p. 281-9.

35. Fleiss, J.L., *Inference about population attributable risk from cross-sectional studies.* Am J Epidemiol, 1979. **110**(2): p. 103-4.

36. Basu, S. and J.R. Landis, *Model-based estimation of population attributable risk under cross-sectional sampling.* Am J Epidemiol, 1995. **142**(12): p. 1338-43.

37. O'Neill, T.J., *Positive bias of the combined effect of risk factors estimated by marginal aetiological fractions.* Int J Epidemiol, 1991. **20**(4): p. 1137-9.

38. Gefeller, O. and G.E. Eide, *The attributable fraction of the combined effect of two risk factors.* Int J Epidemiol, 1992. **21**(4): p. 819-20, 823-4.

39. Rowe, A.K., K.E. Powell, and W.D. Flanders, *Why population attributable fractions can sum to more than one.* Am J Prev Med, 2004. **26**(3): p. 243-9.

40. Gefeller, O., M. Land, and G.E. Eide, *Averaging attributable fractions in the multifactorial situation: assumptions and interpretation.* J Clin Epidemiol, 1998. **51**(5): p. 437-41.
41. Land, M., C. Vogel, and O. Gefeller, *Partitioning methods for multifactorial risk attribution.* Stat Methods Med Res, 2001. **10**(3): p. 217-30.
42. Eide, G.E. and I. Heuch, *Average attributable fractions: a coherent theory for apportioning excess risk to individual risk factors and subpopulations.* Biom J, 2006. **48**(5): p. 820-37.
43. Cox, L.A., *Probability of causation and the attributable proportion of risk.* Risk Analysis, 1984. **4**(3): p. 221-30.
44. von Neumann, J. and O. Morgenster, *Theory of games and economic Behaviour.* Princeton University Press, Princeton, 1944.
45. Shapley, L.S., *A value for n-person games. In: Kuhn H, Turcker A, Eds. Contributions to the Theory Games, II.* Ann Math Studies, 1953. **28**: p. 307-17.
46. McElduff, P., et al., *Estimating the contribution of individual risk factors to disease in a person with more than one risk factor.* J Clin Epidemiol, 2002. **55**(6): p. 588-92.
47. Llorca, J. and M. Delgado-Rodriguez, *A new way to estimate the contribution of a risk factor in populations avoided nonadditivity.* J Clin Epidemiol, 2004. **57**(5): p. 479-83.
48. Cosslett, S.R., *Maximum likelihood estimator for choice-based samples* Econometrica, 1981. **49**(5): p. 1289-1316.
49. Manski, C.F. and S.R. Lerman, *The estimation of choice probabilities from choice based samples.* Econometrica, 1977. **45**(8): p. 1977-88.
50. Haaf, K.T. and E.W. Steyerberg, *Methods for individualized assessment of absolute risk in case-control studies should be weighted carefully.* Eur J Epidemiol, 2016. **31**(11): p. 1067-1068.
51. Bosetti, C., et al., *Diabetes mellitus and cancer risk in a network of case-control studies.* Nutr Cancer, 2012. **64**(5): p. 643-51.
52. Associazione Italiana dei Registri Tumori (AIRTUM), *I tumori in Italia - Rapporto 2014. Prevalenza e guarigione da tumore in Italia.* Milano, Inferenze Edizioni, 2014.
53. Denman, D.W. and J.J. Schlesselman, *Interval estimation of the attributable risk for multiple exposure levels in case-control studies.* Biometrics, 1983. **39**(1): p. 185-92.
54. Benichou, J. and M.H. Gail, *Variance Calculations and Confidence-Intervals for Estimates of the Attributable Risk Based on Logistic-Models.* Biometrics, 1990. **46**(4): p. 991-1003.
55. Kooperberg, C. and D.B. Petitti, *Using logistic regression to estimate the adjusted attributable risk of low birthweight in an unmatched case-control study.* Epidemiology, 1991. **2**(5): p. 363-6.
56. Graubard, B.I. and T.R. Fears, *Standard errors for attributable risk for simple and complex sample designs.* Biometrics, 2005. **61**(3): p. 847-55.

57. Natarajan, S., S.R. Lipsitz, and E. Rimm, *A simple method of determining confidence intervals for population attributable risk from complex surveys.* Stat Med, 2007. **26**(17): p. 3229-39.
58. R Core Team, *R: A language and enviroment for statistical computing.* Vienna, Austria, 2016. **URL https://www.R-project.org**.
59. Taylor, A.E. and W.R. Mann, *Advanced Calculus (3rd ed.).* New York: John Wiley, 1983.
60. Rao, C.R., *Linear Statistical Inference and Its Application (2nd ed.).* New York: John Wiley, 1973.
61. Efron, B. and R.J. Tibshirani, *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.* Statistical Science, 1986. **1**: p. 54-77.
62. World Health Organization (WHO), *International Classification of Diseases for Oncology (ICD-0). Third edn. First Revision.* Geneva: World Health Organization, 2013.
63. Ferlay, J., et al., *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet].* Lyon, France: International Agency for Research on Cancer. Available from: http://globocan.iarc.fr 2013.
64. La Vecchia, C., et al., *Epidemiology and prevention of oral cancer.* Oral Oncology, 1997. **33**(5): p. 302-12.
65. de Camargo Cancela, M., et al., *Oral cavity cancer in developed and in developing countries: population-based incidence.* Head Neck, 2010. **32**(3): p. 357-67.
66. International Agency for Research on Cancer (IARC), *IARC Monographs on the evaluation of carcinogenic risks to humans. Vol. 100E.* Lyon, France: International Agency for Research on Cancer, 2012.
67. Merchant, A., et al., *Paan without tobacco: an independent risk factor for oral cancer.* Int J Cancer, 2000. **86**(1): p. 128-31.
68. Nair, U., H. Bartsch, and J. Nair, *Alert for an epidemic of oral cancer due to use of the betel quid substitutes gutkha and pan masala: a review of agents and causative mechanisms.* Mutagenesis, 2004. **19**(4): p. 251-62.
69. Chaturvedi, A.K., et al., *Incidence trends for human papillomavirus-related and -unrelated oral squamous cell carcinomas in the United States.* J Clin Oncol, 2008. **26**(4): p. 612-9.
70. Franceschi, S., et al., *Food groups, oils and butter, and cancer of the oral cavity and pharynx.* Br J Cancer, 1999. **80**(3-4): p. 614-20.
71. Goldenberg, D., *Mate: a risk factor for oral and oropharyngeal cancer.* Oral Oncol, 2002. **38**(7): p. 646-9.
72. Guha, N., et al., *Oral health and risk of squamous cell carcinoma of the head and neck and esophagus: results of two multicentric case-control studies.* Am J Epidemiol, 2007. **166**(10): p. 1159-73.
73. Radoi, L. and D. Luce, *A review of risk factors for oral cavity cancer: the importance of a standardized case definition.* Community Dent Oral Epidemiol, 2013. **41**(2): p. 97-109, e78-91.

74. Hecht, S.S., *Tobacco carcinogens, their biomarkers and tobacco-induced cancer.* Nat Rev Cancer, 2003. **3**(10): p. 733-44.

75. Warren, G.W. and A.K. Singh, *Nicotine and lung cancer.* J Carcinog, 2013. **12**: p. 1.

76. Hukkanen, J., P. Jacob, 3rd, and N.L. Benowitz, *Metabolism and disposition kinetics of nicotine.* Pharmacol Rev, 2005. **57**(1): p. 79-115.

77. Tutka, P., J. Mosiewicz, and M. Wielosz, *Pharmacokinetics and metabolism of nicotine.* Pharmacol Rep, 2005. **57**(2): p. 143-53.

78. Xue, J., S. Yang, and S. Seng, *Mechanisms of Cancer Induction by Tobacco-Specific NNK and NNN.* Cancers (Basel), 2014. **6**(2): p. 1138-56.

79. Turati, F., et al., *A meta-analysis of alcohol drinking and oral and pharyngeal cancers: results from subgroup analyses.* Alcohol Alcohol, 2013. **48**(1): p. 107-18.

80. Hashibe, M., et al., *Alcohol drinking in never users of tobacco, cigarette smoking in never drinkers, and the risk of head and neck cancer: pooled analysis in the International Head and Neck Cancer Epidemiology Consortium.* J Natl Cancer Inst, 2007. **99**(10): p. 777-89.

81. Seitz, H.K. and F. Stickel, *Molecular mechanisms of alcohol-mediated carcinogenesis.* Nat Rev Cancer, 2007. **7**(8): p. 599-612.

82. D'Avanzo, B., et al., *Reliability of information on cigarette smoking and beverage consumption provided by hospital controls.* Epidemiology, 1996. **7**(3): p. 312-5.

83. Decarli, A., et al., *Validation of a food-frequency questionnaire to assess dietary intakes in cancer studies in Italy. Results for specific nutrients.* Ann Epidemiol, 1996. **6**(2): p. 110-8.

84. Ferraroni, M., et al., *Validity and reproducibility of alcohol consumption in Italy.* Int J Epidemiol, 1996. **25**(4): p. 775-82.

85. Ellisen, L.W. and D.A. Haber, *Hereditary breast cancer.* Annu Rev Med, 1998. **49**: p. 425-36.

86. Daly, B. and O.I. Olopade, *Race, ethnicity, and the diagnosis of breast cancer.* JAMA, 2015. **313**(2): p. 141-2.

87. Hunter, D.J. and W.C. Willett, *Diet, body size, and breast cancer.* Epidemiol Rev, 1993. **15**(1): p. 110-32.

88. Longnecker, M.P., *Alcoholic beverage consumption in relation to risk of breast cancer: meta-analysis and review.* Cancer Causes Control, 1994. **5**(1): p. 73-82.

89. Willett, W.C., G. Colditz, and M. Stampfer, *Postmenopausal estrogens--opposed, unopposed, or none of the above.* JAMA, 2000. **283**(4): p. 534-5.

90. Henderson, B., et al., *In Schottenfeld D., Fraumeni J. Jr (Eds): Cancer Epidemiology and Prevention.* New York, Oxford University Press, 1996: p. 1022-39.

91. Adami, H.O., L.B. Signorello, and D. Trichopoulos, *Towards an understanding of breast cancer etiology.* Semin Cancer Biol, 1998. **8**(4): p. 255-62.

92. Russo, J., et al., *Comparative study of human and rat mammary tumorigenesis.* Lab Invest, 1990. **62**(3): p. 244-78.
93. Kelsey, J.L., M.D. Gammon, and E.M. John, *Reproductive factors and breast cancer.* Epidemiol Rev, 1993. **15**(1): p. 36-47.
94. Harris, J.R., et al., *Breast cancer (1).* N Engl J Med, 1992. **327**(5): p. 319-28.
95. Huang, Z., et al., *Dual effects of weight and weight gain on breast cancer risk.* JAMA, 1997. **278**(17): p. 1407-11.
96. Briand, P., *Hormone-dependent mammary tumors in mice and rats as a model for human breast cancer (review).* Anticancer Res, 1983. **3**(4): p. 273-81.
97. Trichopoulou, A., et al., *Consumption of olive oil and specific food groups in relation to breast cancer risk in Greece.* J Natl Cancer Inst, 1995. **87**(2): p. 110-6.
98. Zhang, S., et al., *Dietary carotenoids and vitamins A, C, and E and risk of breast cancer.* J Natl Cancer Inst, 1999. **91**(6): p. 547-56.
99. Adami, H.O., D.J. Hunter, and D. Trichopoulos, *Textboob of Cancer Epidemiology.* New York, Oxford University Press, 2008: p. 411-412.
100. Zhang, S., et al., *A prospective study of folate intake and the risk of breast cancer.* JAMA, 1999. **281**(17): p. 1632-7.
101. Zhang, S.M., et al., *Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer.* J Natl Cancer Inst, 2003. **95**(5): p. 373-80.
102. Rohan, T.E., et al., *Dietary folate consumption and breast cancer risk.* J Natl Cancer Inst, 2000. **92**(3): p. 266-9.
103. Sellers, T.A., et al., *Dietary folate intake, alcohol, and risk of breast cancer in a prospective study of postmenopausal women.* Epidemiology, 2001. **12**(4): p. 420-8.
104. Fuchs, C. and V.W. Berger, *Quantifying The Proportion Of Cases Attributable To An Exposure.* Journal of Modern Applied Statistical Methods, 2004. **3**: p. 54-64.
105. Llorca, J. and M. Delgado-Rodriguez, *A comparison of several procedures to estimate the confidence interval for attributable risk in case-control studies.* Stat Med, 2000. **19**(8): p. 1089-99.
106. Efron, B. and R.J. Tibshirani, *An Introduction to the Bootstrap.* New York, Chapman and Hall, 1993.