Estimating average attributable fractions with confidence intervals for cohort and case control studies

Estimating average attributable fractions with confidence intervals for cohort and case-control studies

John Ferguson[1], Alberto Alvarez-Iglesias[1], John Newell[1], John Hinde[2] and Martin O' Donnell[2]

[1]HRB Clinical Research Facility, NUI Galway
[2]School of Mathematics, Statistics and Applied Mathematics, NUI Galway

Abstract

Chronic diseases tend to depend on a large number of risk factors, both environmental and genetic. Average attributable fractions were introduced by Eide and Gefeller as a way of partitioning overall disease burden into contributions from individual risk factors; this may be useful in deciding which risk factors to target in disease interventions. Here we introduce new estimation methods for average attributable fractions that are appropriate for both case-control designs and prospective studies. Confidence intervals, derived using Monte Carlo simulation, are also described. Finally, we introduce a novel approximation for the sample average attributable fraction that will ensure a computationally tractable approach when the number of risk factors is large. An R package, *averisk*, implementing the methods described in this manuscript can be downloaded from the CRAN repository.

1: Introduction

The optimal approach to quantifying the individual and cumulative effects of differing risk factors on disease burden is one of the central questions in epidemiological research. Most researchers still report disease/risk factor associations using odds ratios, usually adjusted for confounders via a logistic model. Odds ratios (OR) are designed to measure the amplification in disease risk in an individual who carries a particular risk factor compared to a similar individual who is not a carrier. However, odds ratios say little about the cumulative impact of the risk factor on the burden of disease over all cases, primarily because they do not reflect the prevalence of disease. The attributable fraction[13], λ, is a more relevant metric for this purpose and quantifies the proportional reduction in disease prevalence that would be achieved if the risk factor could be somehow

eliminated from the population. More formally, λ is defined as:

$$\lambda = \frac{P(D) - P(D|E^c)}{P(D)}, \tag{1}$$

where the event D is that a randomly selected individual from the population has the disease, E is the corresponding event that they have the risk factor, and E^c is the complementary event that they do not have the risk factor. In contrast to the odds ratio, which for rare diseases closely approximates the relative risk, $RR = P(D|E)/P(D|E^c)$, the attributable fraction, λ, depends jointly on relative risk and the prevalence of the risk factor within the general population, $p = P(E)$. This fact can be highlighted by using Bayes formula to re-express (1) as:

$$\lambda = \frac{(RR-1)p}{1+(RR-1)p}. \tag{2}$$

Neither (1) nor (2) are directly estimable from case-control datasets, where direct estimates of disease prevalence are not available. However, another reexpression[7] together with the assumption that $RR \approx OR$, yields:

$$\lambda \approx \frac{P(E|D)(OR-1)}{OR}. \tag{3}$$

which can be used to approximate the attributable fraction in case-control studies, provided that the disease is rare.

Of course, disease etiology generally involves multiple risk factors and a pertinent question is how to apportion the total risk attributed to a collection of K risk factors into individual contributions. Individual attributable fractions for each risk factor, as given by (2), may add up to more than 1[24, 25, 17, 20]. In contrast, the associated average attributable fractions, $\boldsymbol{\lambda_a} = (\lambda_{a,1}...\lambda_{a,K})$ (defined in Section 2) partition a well-defined combined attributable fraction into individual contributions for each risk factor[6]. This seems an important calculation when describing the epidemiology for a disease affected by a number of known risk factors, yet despite this, a number of methodological and practical issues have inhibited the adoption of average attributable fractions into mainstream epidemiological research. First, to our knowledge, there is no known general method to generate confidence intervals for $\boldsymbol{\lambda_a}$. Second, computing $\boldsymbol{\lambda_a}$ may be prohibitively expensive when the number of risk factors being examined is large. Finally, while software to calculate $\boldsymbol{\lambda_a}$ is available for dichotomous risk factors[18], the software is designed for prospective studies and will give biased results when applied to case-control data. In addition, their formulation involves redundant summation, which can unnecessarily increase the computational burden. The techniques for calculating average attributable fractions demonstrated in this paper, and associated R-package[15]: *averisk* (downloadable from the CRAN archive), address all of these shortcomings.

The rest of the paper is outlined as follows. In Section 2, we give some background regarding the conceptual development of average attributable fractions.

In Section 3, we address methodological concerns regarding the theoretical construction, computation and interpretation of an appropriate confidence interval for $\boldsymbol{\lambda_a}$ and also the correct adaptation of the estimation procedure for case-control data. We also develop an approximation that estimates $\boldsymbol{\lambda_a}$ to a high degree of precision when the number of risk factors is large. In Section 4, we use simulations to demonstrate coverage properties of the suggested confidence interval and to indicate the computational speedup observed by the approximation. In the final section, we present our conclusions and suggest some extensions that could be implemented in the future.

2: Attributable Fractions

In the context of multiple risk factors, we would like the attributable fraction for a risk factor, E, to have an interpretation as the proportional decrease in disease prevalence if E were eliminated, but the distributions of other risk factors and confounders throughout the population remained unchanged. When the risk factors are correlated, λ, as defined in (2), corresponding to an unadjusted relative risk, cannot be interpreted in this way; instead, it may summarize the effects of both the risk factor of interest and several other correlated risk factors, potentially overstating the effect for E. Obviously a form of attributable risk that is adjusted for other relevant risk factors and potential confounders is necessary. There are a variety of differing approaches to calculating such a quantity, the simplest of which simply replaces the unadjusted relative risk in (2) with a predicted odds ratio (adjusted for confounding) from a model for disease[3]. An alternative approach involves first fitting a model for disease occurrence with all risk factors and confounders of interest[9]. This model is used to predict the total number of disease cases that would have been observed in the dataset under the scenario that no individual had the risk factor of interest (usually equivalent to a coding of '0' for all individuals for that risk factor, under the typical 0/1 coding for dichotomous risk factors), but with the values of all other risk factors left unchanged. We denote this predicted quantity by $\hat{N}$, and label the observed number of disease cases in the sample as N_{case}. An adjusted attributable fraction can be estimated using:

$$\hat{\lambda}^{adj} = (N_{case} - \hat{N})/N_{case}. \tag{4}$$

Note that (4) is not appropriate for use with case-control designs, since the ratio of cases to controls in the sample is fixed a priori and as a result the resulting predicted probabilities from the fitted models will be biased, as will (4) itself. While formulae and associated variance estimates for adjusted attributable fractions, estimable with case-control data, have been developed[2, 9, 4] (also see Supplementary Material, Section 1, for an adaptation of these formulae to estimating sequential attributable fractions), we discuss a differing approach in this manuscript, based on incorporating prevalence-based weighting. This approach has a couple of distinct advantages: first, it facilitates a single estimation procedure for both prospective-cohort and case-control studies and second, it does

not require approximating relative risks with odds ratios, and as a result is a valid estimation approach for common (i.e. high prevalence) diseases. The general approach of applying weighting to estimators appropriate for prospective studies to define consistent estimators in case-control studies is discussed in a more general context in Van der Laan (2008)[21].

2.1 Combined and sequential attributable fractions

Sequential attributable fractions[6] are simplest to explain in the unrealistic, but easy to understand, scenario that only 2 risk factors influence the disease. In this case, it represents the proportion of the (original) disease prevalence that can be eliminated by removing a second risk factor from the population, over and above that which has already been eliminated by removing the first risk factor. Suppose the risk factors are labeled as 1 and 2, and that the attributable fraction for 1 is denoted λ_1, with the combined attributable fraction for both 1 and 2 denoted as λ_{12}. The combined attributable fraction is defined as the proportional decrease in disease prevalence from removing both 1 and 2 from the population. We denote the sequential attributable fraction for risk factor 2, after removal of 1, as $\lambda_{2|1}$ which represents the difference in these 2 quantities:

$$\lambda_{2|1} = \lambda_{12} - \lambda_1. \tag{5}$$

It can then be estimated (again assuming a prospective design) using:

$$\hat{\lambda}_{2|1} = (\hat{N}_1 - \hat{N}_{12})/N_{case}, \tag{6}$$

where $\hat{N}_1$ and $\hat{N}_{12}$ represent estimates for the number of disease cases assuming risk factor 1, and both risk factors, 1 and 2, are removed from the population. In fact, these definitions can be extended to any number of risk factors in the obvious way. Suppose that there are K risk factors of interest, labeled $1, 2, \ldots, K$. Then the sequential attributable fraction for risk factor k, $2 \leq k \leq K$, after the removal of risk factors $1, \ldots, (k-1)$ is:

$$\lambda_{k|1\ldots(k-1)} = \lambda_{12\ldots k} - \lambda_{1\ldots(k-1)}, \tag{7}$$

where $\lambda_{12\ldots k}$ denotes the combined attributable fraction for risk factors $1, \ldots, k$. Note that the combined attributable fractions $\lambda_{12\ldots k}$ is estimated from models that include all the risk factors, $(1, \ldots, K)$. For this reason, in the context of K known risk factors, λ_1 is the attributable fraction for risk factor 1, adjusted for risk factors $2, 3, \ldots, K$.

2.2 Average attributable fraction

The above decomposition has the nice property that the sum of the sequential contributions for each risk factor equals the combined attributable fraction for all risk factors in the collection. In contrast, often the individual attributable fractions, as calculated using (1) or (2), sum to more than 1[25, 17, 20]. However, sequential attributable fractions have the disadvantage that they directly

depend on the order in which risk factors are removed from the population, which may be arbitrary. For instance, even for diseases where epidemiology has been established, such as cardiovascular disease, sequencing of risk factors is challenging, as many occur simultaneously. Average attributable fractions mitigate this problem by averaging the differing sequential attributable fractions derived from every possible permutation by which the risk factors could be eliminated from the population, while still keeping the nice property that the component attributable fractions sum to the combined attributable fraction. More specifically, let σ represent some permutation function over the integers $1, \ldots, K$, that is an invertible function $\sigma : \{1, ...K\} \to \{1, ..., K\}$. For example, a permutation σ with $\sigma(1) = 5$ represents a removal order where risk factor '1' is the 5^{th} risk factor to be eliminated from the population. We denote the set of all possible permutation functions by $\boldsymbol{S_K}$. Using the notation above, the average attributable fraction for risk factor k, can be represented as:

$$\lambda_{a,k} = \left\{ \sum_{\sigma \in \boldsymbol{S_K}: \sigma(k)=1} \lambda_k + \sum_{\sigma \in \boldsymbol{S_K}: \sigma(k) \neq 1} \lambda_{k|\sigma^{-1}(1)...\sigma^{-1}(\sigma(k)-1)} \right\} / K!. \quad (8)$$

While the expression $\lambda_{k|\sigma^{-1}(1)...\sigma^{-1}(\sigma(k)-1)}$ looks complex, it is simply the sequential attributable fraction from removing risk factor k having already removed the risk factors labeled: $\sigma^{-1}(1)...\sigma^{-1}(\sigma(k)-1)$. Note that when $\sigma(k) \neq 1$, $\sigma^{-1}(1)...\sigma^{-1}(\sigma(k)-1)$ is the set of risk factors that appear before risk factor k, according to the order specified by σ. To make this more concrete, consider a situation with $K = 4$ risk factors, where we want to calculate the sequential attributable fraction corresponding to removing 2, after having removed first 4, then 3 and then 1, that is: $\lambda_{2|4,3,1}$. Then since 4 is the first risk factor to be removed, $\sigma(4) = 1$. Similarly, $\sigma(3) = 2$, $\sigma(1) = 3$ and $\sigma(2) = 4$. Applying the inverse permutation, it follows then that $\sigma^{-1}(1) = 4$, $\sigma^{-1}(2) = 3$, $\sigma^{-1}(3) = 1$, which are the 3 risk factors that are removed before removing $\sigma^{-1}(4) = 2$.

As mentioned in Section 1, (8) can be cumbersome to calculate. In the next section, we adapt the above derivations for case-control studies, develop some computational approximations that can be used when the number of risk factors is large, and construct a Monte Carlo based confidence interval that can be used in either case-control or prospective studies.

3: Estimation and confidence intervals

Modification for case-control studies

As mentioned in the previous section, estimators such as (4), which rely on fitted probabilities for disease, are biased when applied to case control data due to the fixed ratio of cases to controls in the sample. Provided an estimate of disease prevalence is available, a possible way to correct this imbalance is to weight the likelihood function used when estimating the disease prediction

model. Essentially, this requires us to find the coefficient vector $\hat{\boldsymbol{\beta}}$ to maximize:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i l_i(\boldsymbol{\beta}), \tag{9}$$

with $l_i(\boldsymbol{\beta})$ corresponding to the individual log-likelihood contribution for the i^{th} observation, and $w = (w_1, \ldots, w_N)$ a vector of weights. For instance, suppose that the ratio of controls to cases in the sample is $r : 1$ and the prevalence of the disease in the population is p. If each case is given a weight of 1, then each control should be given a weight of $(1-p)/rp$. In the case that $(1-p)/rp$ is an integer, the estimated probabilities of disease from the fitted model, weighted in this way, are identical to those estimates that would be found from a unweighted model, with an altered design matrix where the row for each control is repeated $(1-p)/rp$ times, and the row for each case is repeated only once. In addition, the estimated sequential attributable fractions, that constitute the summands for the estimate, also need to be adjusted to account for the imbalance between cases and controls in the following way:

$$\hat{\lambda}_{k|1..k-1} = (\sum_{i=1}^{N} w_i \hat{p}_{i1} - \sum_{i=1}^{N} w_i \hat{p}_{i2})/N_{case}, \tag{10}$$

where $\hat{p}_{i1}$ and $\hat{p}_{i2}$ are the predicted probabilities that the i^{th} individual is a case from the weighted model, assuming the values of the risk factors $1, \ldots, k-1$ (for $\hat{p}_{i1}$) and values of $1, ..., k-1, k$ (for $\hat{p}_{i2}$) are set to their reference levels. Note that the same formula is valid in prospective studies, provided the weight terms, w, satisfy $w_i = 1$ for all $i = 1 \ldots N$. The average attributable fraction for risk factor k, adjusting for possible case-control structure, can then be estimated by substituting the calculated values for (10) into (8):

$$\widehat{\lambda}_{a,k} = \left\{ \sum_{\sigma \in \boldsymbol{S_K}: \sigma(k)=1} \hat{\lambda}_k + \sum_{\sigma \in \boldsymbol{S_K}: \sigma(k) \neq 1} \hat{\lambda}_{k|\sigma^{-1}(1)\ldots\sigma^{-1}(\sigma(k)-1)} \right\} / K!. \tag{11}$$

Reorganizing the exact computation

There is some redundancy in the term by term summation of (11). Take for example a term: $\hat{\lambda}_{k|321}$ corresponding to a permutation σ with $\sigma^{-1}(1) = 3, \sigma^{-1}(2) = 2, \sigma^{-1}(1) = 3$ and $\sigma^{-1}(4) = k$. It represents the proportional decrease in disease prevalence from removing risk factor k from the population, given that the risk factors $3, 2$ and 1 were already removed, in that order. However, this calculation is the same no matter the order in which risk factors $1, 2$ and 3 were removed. More specifically: $\hat{\lambda}_{k|321} = \hat{\lambda}_{k|123} = \hat{\lambda}_{k|132} = \hat{\lambda}_{k|213} = \hat{\lambda}_{k|231} = \hat{\lambda}_{k|312}$. Similarly, if there are 10 risk factors in the problem, the sequential attributable risk $\hat{\lambda}_{k|321}$ does not depend on which of the $6! = (10-1-3)!$ orders the remaining risk factors are deleted from the population, after deleting k. In general, let $C_r^{K,-k}$ denote the set of all subsets of size r (i.e. unordered

choices of r integers) from the set $1, 2, \ldots, K \backslash \{k\}$. For $\boldsymbol{s} \in C_r^{K,-k}$, we write $\hat{\lambda}_{k|\boldsymbol{s}}$ to be the estimated sequential attributable fraction calculated for risk factor k, given that the risk factors corresponding to the subset $\mathbf{s}$ were already removed from the population. Denoting the cardinality of a set $\mathbf{s}$, as $|\boldsymbol{s}| = r$, the above argument shows $\hat{\lambda}_{k|s}$ corresponds to exactly $(K-1-r)!r!$ permutations in (11). We can re-express (11) as:

$$\hat{\lambda}_{a,k} = \sum_{r=0}^{K-1} \left\{ (K-1-r)!r! \sum_{\boldsymbol{s} \in C_r^{K,-k}} \hat{\lambda}_{k|\boldsymbol{s}} \right\} / K!, \tag{12}$$

defining $\hat{\lambda}_{k|\boldsymbol{\phi}} = \hat{\lambda}_k$ for the empty subset $\boldsymbol{\phi} \in C_r^{K,-k}$. The number of summands in (12) is 2^K as opposed to $K!$ for (11) and thus can result in a substantial computational saving for large K. Note that this more computationally efficient formula for the average attributable fraction has been identified before by a number of authors[19, 12].

Approximate Average Population Attributable Fraction ($\tilde{\lambda}_a$)

When the number of risk factors is large, an exact calculation of $\hat{\lambda}_{a,k}$ may be prohibitively expensive, even after re-expressing the formula as (12). As an alternative, one can take a random sample of permutations from the permutation group $\boldsymbol{S_K}$, that is the set of all permutations of the integers $1, \ldots, K$, and instead average the sequential attributable fractions for the sampled permutations. Suppose $(\sigma_1 \ldots . \sigma_m)$ are such a random sample of permutations. The approximate average population attributable fraction for risk factor k is defined as:

$$\tilde{\lambda}_{a,k} = \left\{ \sum_{i \leq m : \sigma_i(k) = 1} \hat{\lambda}_k + \sum_{i \leq m : \sigma_i(k) \neq 1} \hat{\lambda}_{k|\sigma_i^{-1}(1) \ldots \sigma_i^{-1}(\sigma_i(k)-1)} \right\} / m. \tag{13}$$

Assuming m is chosen to be large enough, there will be negligible difference between this and the exact calculation, given by (11). In addition, the additional imprecision in this estimate can be incorporated into the confidence interval as is suggested below:

Monte Carlo confidence interval for $\lambda_{a,k}$

Note that $\hat{\lambda}_{a,k}$ can be regarded as a function, $f(\mathbf{X}, \hat{\boldsymbol{\beta}})$, of both the estimated coefficient vector, $\hat{\boldsymbol{\beta}}$, that is used to generate predicted probabilities for (10), and the observed design matrix $\mathbf{X}$. However it turns out that the sampling variance of $\hat{\boldsymbol{\beta}}$ is the most important factor in the determination of $Var(\hat{\lambda}_{a,k})$ in the sense that $Var(\hat{\lambda}_{a,k}) = Var_{X,\hat{\boldsymbol{\beta}}}(f(\mathbf{X}, \hat{\boldsymbol{\beta}})) \approx Var_{\hat{\boldsymbol{\beta}}}(f(X, \hat{\boldsymbol{\beta}})|\mathbf{X})$ (see the supplementary material for more discussion). With this in mind, we simulate B vectors: $\boldsymbol{\beta_1} \ldots . \boldsymbol{\beta_B}$ from the estimated asymptotic sampling distribution of

$\hat{\boldsymbol{\beta}}$, that is the multivariate normal distribution with mean $\hat{\boldsymbol{\beta}}$, the MLE for the weighted log likelihood, and covariance $\hat{\boldsymbol{C}}$. In the examples considered in this manuscript we have used $B = 100$, in accordance with previous suggestions for the number of bootstrap replicates required to estimate a standard error[5]. When estimating $\hat{\boldsymbol{C}}$ we use the covariance matrix from the unweighted model, since the effective sample size according to the weighted likelihood is artificially high, while the unweighted likelihood correctly respects the actual number of cases and controls sampled. Suppose we fix a variable of interest, k. Separate estimates $\hat{\lambda}^1_{a,k} \ldots . \hat{\lambda}^B_{a,k}$ for the average attributable fraction of that variable are produced from each sampled $\boldsymbol{\beta_b}$ vector ($b = 1, \ldots B$), by setting $\hat{\lambda}^b_{a,k} = f(\mathbf{X}, \boldsymbol{\beta_b})$. In practice, when a small number of risk factors (say $K \leq 9$) are present, each Monte Carlo simulate $\hat{\lambda}^b_{a,k}$ is calculated using (12), where the component estimated sequential fractions are calculated using the coefficient vector $\boldsymbol{\beta_b}$ rather than $\hat{\boldsymbol{\beta}}$. The variance of $\hat{\lambda}_{a,k}$ can be estimated as:

$$\widehat{Var}_k = \sum_{b=1}^{B} (\hat{\lambda}^b_{a,k} - \overline{\lambda}_{a,k})^2/(B-1), \tag{14}$$

with $\overline{\lambda}_{a,k} = (\sum_{b=1}^{B} \hat{\lambda}^b_{a,k})/B$. One can view the above estimator as a Monte Carlo alternative to the Delta Method, in that the variance of a function of $\hat{\boldsymbol{\beta}}$ is estimated via simulation rather than using an analytic approximation.

In the case that K is large, (say with at 10 or more risk factors), we instead use approximate average population attributable fractions: $\tilde{\lambda}^1_{a,k} \ldots \tilde{\lambda}^B_{a,k}$, when estimating the variance, with each estimate, $\tilde{\lambda}^j_{a,k}$, computed with a different $\boldsymbol{\beta_b}$. We suggest using a total of m sampled permutations for $\tilde{\lambda}^b_{a,k}, b \leq B$ to ensure equivalence of the computational cost of calculating the estimated variance and point estimate (recall that m was is defined as the number of sampled permutations that are used when calculating the point estimate). In this case, these m permutations, need to be evenly shared over $\tilde{\lambda}^b_{a,k}, j = 1 \ldots B$, with the result that each $\tilde{\lambda}^b_{a,k}$ will use m/B permutations. An estimated variance for the estimate $\tilde{\lambda}_{a,k}$ is then given by:

$$\widehat{Var}_k = \sum_{b=1}^{B} (\tilde{\lambda}^b_{a,k} - \overline{\tilde{\lambda}}_{a,k})^2/(B-1) - \hat{v}(B-1)/m, \tag{15}$$

where $\hat{v}$ is the average of the sample variances of the estimated sequential attributable fractions over the sampled permutations for a fixed $\boldsymbol{\beta_b}$, and $\overline{\tilde{\lambda}}_{a,k} = (\sum_{b=1}^{B} \tilde{\lambda}^b_{a,k})/B$. The correction factor adjusts the variance for the fact that $\tilde{\lambda}_{a,k}$ is an approximation for (13), with $\hat{v}/m$ estimating the variance of the approximation error, and also for the fact that the point estimate given in (13), $\tilde{\lambda}_k$, uses m permutations, whereas $\tilde{\lambda}^b_k$ uses only m/B permutations. See the Supplementary Material for more details regarding the derivation of this correction factor.

In either case a $100(1-\alpha)\%$ confidence interval is produced using

$$\hat{E}_k \pm t_{B-1,(1-\alpha/2)} \cdot \sqrt{\widehat{Var}_k} \tag{16}$$

where $\hat{E}_k$ is either calculated by either (12) or (13) depending on whether it was necessary to sample permutations, and $t_{B-1,(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the t distribution with $B-1$ degrees of freedom. Note that a t-distribution quantile, as opposed to a normal quantile, is used to reflect the fact that the variance is estimated. When estimated average attributable fractions are used, calculating the point estimate, $\hat{E}_k$, using (13) and the variance term, $\sqrt{\widehat{Var}_k}$, using (15) requires separate computations, each involving the calculation of m sequential attributable fractions. An alternative more computationally efficient approach is to use a point estimate, $\hat{E}$, that instead averages over the Monte Carlo simulates, $\tilde{\lambda}^b_{a,k}$, for the approximate average attributable fraction. That is to use: $\hat{E}_k = (\sum_{b=1}^{B} \tilde{\lambda}^b_{a,k})/B$ instead of (13). In this case the same m sequential attributable fractions are used in calculating $\hat{E}_k$ and $\sqrt{\widehat{Var}_k}$. So long as B is relatively large, there will be negligible difference between the two estimates.

4: Applications to real and simulated datasets

Hordaland study

As a first example, we consider the Hordaland study of obstructive lung disease, utilized previously to demonstrate average attributable risk [1, 6, 8]. The original study consisted of a random sample of 4,270 residents of Hordaland County in Norway. The variable of interest was the presence (yes or no) of chronic cough. In the final model for the occurrence of cough considered by Eide and Gefeller[6], 3 risk factors were included: residence (urban or rural), smoking habits (never, ex, 1-9 cigarettes/day, 10-19 cigarettes/day and $\geq$ 20 cigarettes/day) and finally occupational exposure to gas/dust (yes or no). We used the R-package *averisk* to estimate the average and combined attributable risks corresponding to these risk factors and the associated 95% confidence intervals. Note that the package first fits a weighted logistic model for disease status, and uses equations (10) and (11) to calculate the average attributable fractions. In this case, since the design of the study is a random sample of the population, rather than a case-control study, weights of $w_i = 1$ were used for all 4,270 individuals constituting the sample. 95% confidence intervals using the approach discussed in Section 3 were computed using 1,000 Monte Carlo simulates. The respective average attributable fractions for these 3 risk factors and the respective 95% confidence intervals were calculated to be 0.127 (0.048,0.207), 0.350 (0.253,0.446) and 0.125 (0.072,0.178), generating a combined attributable fraction of 0.602 (0.496,0.707). A rough interpretation of these results, is that an estimated 60.2% of the prevalence of chronic cough in Hordaland county was due to the 3 risk factors. This 60.2% partitions into roughly equal contributions

from residence and occupational exposure (12.7% and 12.5%), and a much larger contribution due to smoking (at 35%), indicating that reducing both the degree and prevalence of smoking should be a priority in any intervention targeting chronic cough. Note that the point estimates given here are identical to those previously reported in the literature[8].

As a second use for this dataset, we treat the joint distribution of risk factors and disease estimated from the Hordaland study as a new population with the average attributable and combined attributable risks estimated in the previous paragraph now representing parameters for that population. The distribution of risk factors within cases and controls for this population was derived via Bayes rule, using information detailed in [8]. Using these distributions, risk factor sets were simulated for $N/2$ cases and $N/2$ controls, with N set as either 2,000 or 10,000 depending on the simulation. For each simulated dataset, a weighted logistic regression was fit for the response, with weights specificed according to a disease prevalence of 0.09. Monte Carlo confidence intervals were then produced using (16) based on the variance in (14). Coverage percentages for the procedure were estimated from 1,000 independent simulations as the percentage of the associated confidence intervals that contained the true λ_a, see Table 1. In most cases, the coverage percentages are a little higher than the target nominal levels. For instance, for N=2,000, the empirical coverage percentages are 94.9%, 96.1%, 97.1% and 95.4%, when the target is 95% confidence. Note though that the Monte Carlo confidence interval is only approximate, and may exhibit slightly different behaviour on other datasets.

What might happen if we used weights of 1 for everybody, effectively ignoring the fact that disease prevalence was known to be 0.09? It turns out that $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ will be badly biased in such a situation. For instance, over the 1000 simulations, the averages of the estimates $\hat{\lambda}_{a,k}$, $k \leq 3$ were 0.128, 0.350 and 0.125 totaling to 0.603 when incorporating prevalence-based weights into the estimation, equaling the true fractions used for the simulation after rounding to 3 decimal places. In contrast, when weighting was not incorporated the average estimates and associated true values (in parentheses) were 0.091 (vs. 0.1275), 0.274 (vs. 0.3496) and 0.090 (vs. 0.1247), indicating substantial relative biases of -28.6%, -21.6% and -27.8% in the estimated average attributable fractions and highlighting the practical necessity to correct for case-control structure.

Genetic simulation

We next consider an example motivated by complex disease genetics. While in contrast to environmental risk factors, genetic risk factors are not modifiable, the breakdown between the genetic and environmental contributions to disease is still a question of great scientific interest. The separate genetic and environmental contributions are generally measured using the heritability[22] coefficient, but in the case that a known set of Single Nucleotide Polymorphisms (SNPs) are associated with the disease, attributable fractions represent another principled method to partition total disease prevalence between environmental and known genetic contributors, and even into the contribution of individual

Table 1: Risk factors (urban or rural residence, smoking status and occupational exposure) were simulated for $N/2$ cases with obstructive lung disease and $N/2$ healthy controls based on the empirical distributions for the 4,270 individuals constituting the original Hordaland dataset. Confidence intervals for average attributable fractions were computed for each of 1,000 independent simulations. Coverage percentages were estimated by the percentage of the various confidence intervals that included the true average attributable fraction or combined attributable fraction.

N=2,000

Confidence Level	Urban/Rural	Smoking	Occupational Exposure	Combined
99%	99.0	99.5	99.6	99.4
95%	94.9	96.1	97.1	95.4
90%	89.5	92.2	93.3	91.4
80%	80.8	83.7	85.2	82.3

N=10,000

Confidence Level	Urban/Rural	Smoking	Occupational Exposure	Combined
99%	99.3	99.6	99.4	99.0
95%	94.9	97.4	96.9	95.8
90%	90.4	91.9	93.3	90.7
80%	80.3	82.8	85.1	81.4

SNPs to total disease prevalence[16]. (Note that SNPs are single-base positions in the human genome showing substantial allelic variation from person to person. For instance, perhaps 80% of DNA molecules in the population have a Guanine (G) nucleotide at that position, whereas 20% have a Thymine (T) nucleotide. The minor allele frequency is defined as the proportion of molecules with the less common allele, which would be 0.2 in this example.) Here, we simulate genetic data from a hypothetical complex genetic disease, with a population prevalence of 0.01. Disease occurrence is assumed to depend causally on 25 known SNPs, each with minor allele frequency 0.2. The genotype at SNP position $j, j \leq 25$, for individual i in the population can be represented as the number of copies, X_{ij}, that i has for the less frequent allele at position j. Note that each person has 2 homologs (i.e. 2 DNA molecules), and as a result 2 alleles at each SNP position so that $X_{ij} \in 0, 1, 2$, with probabilities (0.64, 0.32, 0.04) assuming independence of the values for these two alleles in the population. We assume the following model relating the probability of disease to genotype:

$$logit(P(D_i|\boldsymbol{X}_i)) = -5.63 + 0.2 \sum_{j \leq 25} X_{ij}, \tag{17}$$

with $\boldsymbol{X_i}$ representing the vector $(X_{i1}, \ldots, X_{i25})$ and D_i the event that individual i in the population has the disease. The intercept term, -5.63, ensures the population prevalence is 0.01. We can also calculate the combined attributable fraction for the 25 SNPs as: $(0.01 - e^{-5.63}/(1 + e^{-5.63}))/0.01$ =0.644 by using

(1). Arguing by symmetry (the effect size and prevalence of each SNP is the same and the SNPs are independent), the average attributable fraction associated with each SNP must be $0.644/25 \approx 0.026$. We sampled genotype values for $N/2$ cases and $N/2$ controls (with N=2,000 or N=20,000) from the above probability model, adjusted for case-control structure via Bayes Rule, for either 40 (when N= 2,000) or 8 (when N=20,000) independent simulations. Due to the symmetry of the problem, these set of simulations generate $40 \times 25 = 1000$ and 8×25 estimates for the average attributable fraction (which is the same for all SNPs). For each simulation, we calculated $\tilde{\lambda}_{a,k}$ for each SNP, according to (13) using 1,000, 10,000 and 100,000 permutations, under the assumption that the underlying prevalence was known. The number of rare alleles $X_{ij} \in 0, 1, 2$ at locus j was coded as a categorical variable with 3 levels. Here, the number of terms in (12) is extremely large: $2^{25} = 33,554,432$ and taking a random subset of permutations may be necessary, even on a powerful computer.

The left part of Figure 1 shows the clustering of $\tilde{\lambda}_{a,k}$, $k \leq 25$, around the true $\lambda_{a,k}$, 0.026, for N=2,000 and N=20,000, with $\tilde{\lambda}_{a,k}$ calculated using 100,000 permutations in each case. Using a sample size of $N = 20,000$ results in approximate average attributable fractions that are typically much closer to $\lambda_{a,k}$, as indicated by the smaller sampling variation around $\lambda_{a,k}$. If we calculated $\hat{\lambda}_{a,k}$ for each sample, we would still expect such sampling variation around the true $\lambda_{a,k}$. Similar results for the combined attributable fraction: $\lambda_{12...25}$ are shown in on the right side of the Figure. Figure 2 illustrates the approximation error in using $\tilde{\lambda}_{a,k}$ as an estimate of $\hat{\lambda}_{a,k}$ by comparing $\tilde{\lambda}_{a,k}$s calculated with 1,000 and 10,000 permutations to that calculated with 100,000 permutations, where the $\tilde{\lambda}_{a,k}$ calculated with 100,000 permutations is treated as a proxy for an exact calculation of $\hat{\lambda}_{a,k}$. The top-pane of the Figure shows that the approximation error is generally less than 0.001, even when only using 1,000 permutations. These issues are discussed further in the following section

How many permutations do we need?

In most settings 1000 permutations should ensure $\tilde{\lambda}_{a,k}$ is a sufficiently accurate approximation of $\hat{\lambda}_{a,k}$, but more permutations can be used if necessary. Regarding the $K!$ possible sequential attributable fractions constituting (11) as a population, $\hat{\lambda}_{a,k}$ and $\tilde{\lambda}_{a,k}$ can be regarded as the respective true and sample means for this population. Using the standard results regarding the standard error of a sample mean, the approximation error should be roughly $\sqrt{(v/N)}$, where $\sqrt{v}$ is the estimated standard deviation from the empirical distribution of the sampled sequential attributable fractions (i.e. the sample standard deviation for this population). Going further, a 95% confidence interval for the approximation error has margin of error $ME = 1.96 \times \sqrt{(v/N)}$, indicating that using $N = v \times 1.96^2/ME^2$ permutations ensures an approximation error of ME with 95% confidence. An estimate $\hat{v}$ can be found from the sample variance for the set of sequential attributable fractions from a small number of permutations (in the *averisk* package, 100 permutations are used). As an example,

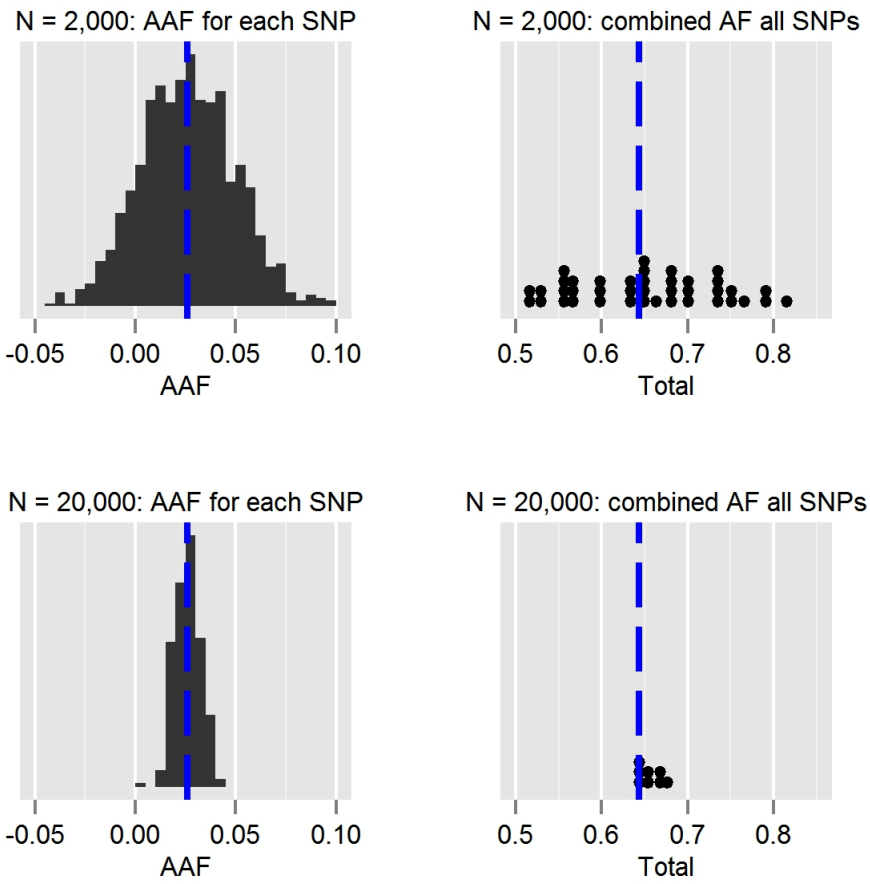

Figure 1: Approximate average (AAF) and combined attributable fractions (combined AF) for genetic example. Each plot summarizes 40 simulations from the genetic model described in the text. Since 25 SNPs are involved in the genetic model, and the average attributable fraction is 0.026 for each SNP, there are 25×40=1,000 estimates for the true average attributable fraction (L.H.S. of plot) and 40 estimates for the true combined attributable fraction (R.H.S. of plot). All estimates were computed using approximate average attributable fractions using 100,000 permutations.

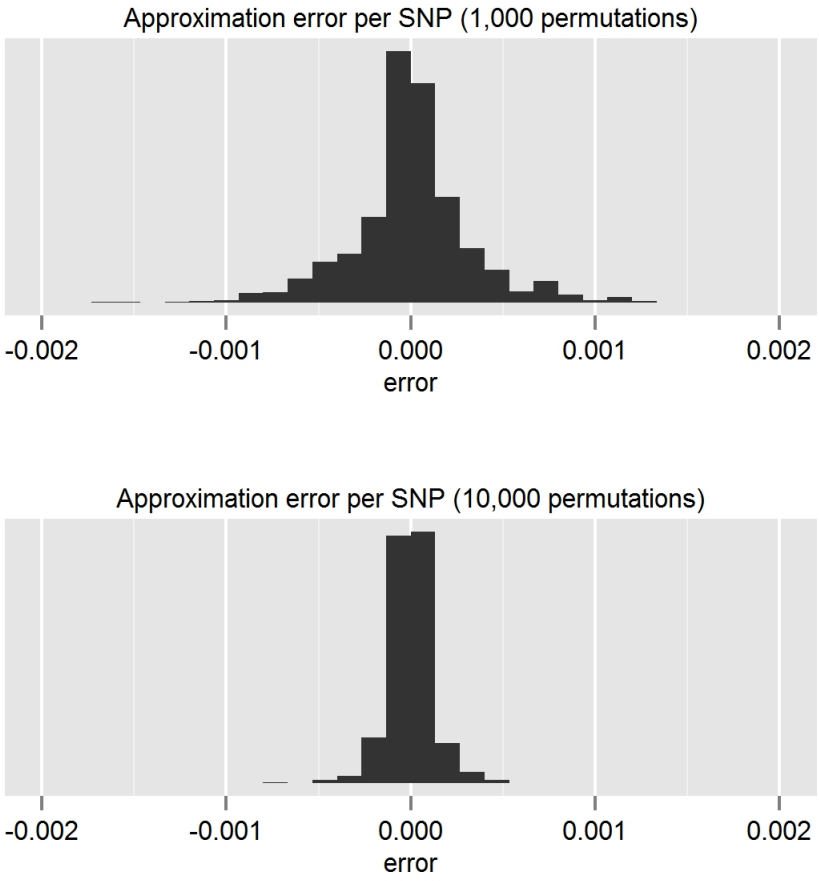

Figure 2: Error of $\tilde{\lambda}_a$ in approximating $\hat{\lambda}_a$ for genetic example. Approximate average attributable fractions using 100,000 permutations were used as a proxy for $\hat{\lambda}_a$. Approximation error was assessed by examining the difference between this quantity and approximate average attributable fractions calculated using 1,000 and 10,000 permutations.

for the genetic example, the estimated standard deviation of the sequential attributable fraction is about 0.01 for each SNP. Using this formula, we find that $0.01^2 \times 3.84/(0.001)^2 = 384$ permutations are necessary to ensure an approximation error of 0.001. The results shown in Figure 2 empirically show the approximation errors when using 1,000 and 10,000 permutations (which from the preceding discussion, we would expect to be less than 0.001). Using the R-package, *averisk*, calculating $\tilde{\lambda}_{a,k}$ with 1000 permutations for the genetic example with 25 SNPs (with 1000 cases and 1000 controls) took roughly 2 minutes, when running on a single core of a Desktop Intel i7-4770 processor. Calculating $\tilde{\lambda}_{a,k}$ for the problem with 10,000 cases and 10,000 controls took 13 minutes using the same machine. In contrast, calculating the exact average attributable risk using (12), for each SNP would take roughly $25 \times (2^{25} - 1)/1000 = 838,860$ times longer - that is just over 3 years (with 1,000 cases and 1,000 controls), and just over 20 years (with 10,000 cases and 10,000 controls).

5: Discussion

Average attributable fractions represent a principled way to partition the contribution to disease prevalence from a set of risk factors into individual components. In addition, they represent the only method of risk allocation satisfying several desirable rationality criteria[12] . Nevertheless, the method has not been adopted in practice by researchers, perhaps as a result of (a) programming issues due to a lack of available software (b) computational difficulties when the number of risk factors is large and (c) no known procedure to calculate confidence intervals. We address all of these deficiencies in this paper.

As mentioned in Section 2, publicly available software calculate average attributable fractions in the case that all risk factors are 2-level categorical variables does exist [18]. Our R package, *averisk*, recently submitted to the CRAN repository, can also be used to estimate average attributable fractions but allows much greater flexibility. First, we allow adjustment for categorical risk factors with 3 or more levels, in addition to binary risk factors. We also allow adjustment for confounders that not modifiable risk factors – age and race might be two typical variables that fall within this category. When calculating $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$, we use the more computationally efficient formula (12) as opposed to (11), which will result in a large computational speed-up when the number of risk factors is large. Finally, we introduce the idea of weighting the log-likelihood into the context of attributable risk, allowing our methods to be used on both prospective and case-control studies. Here we also propose the idea of approximate average attributable fractions, $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$, defined as an approximation for the average attributable fraction calculated by randomly sampling permutations, enabling a tractable computation of the average attributable fraction even when the number of risk factors is large, and demonstrate the size of the approximation error as a function of the number of sampled permutations. While an asymptotic expression for the variance of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ has been calculated for the case of binary risk factors in prospective studies[10], the formula is not easy to calculate and cannot

be used for non-binary categorical risk factors, case-control sampling designs or when other variables (not regarded as modifiable risk factors) are confounding the relationships between risk factors and response. Here, we demonstrate a more general Monte Carlo method to calculate confidence intervals. In contrast to bootstrapping, the method requires no additional computation over that required to calculate the point estimate, $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$, since the same set of sequential attributable fractions (for various permutations) can be used when calculating the point estimate of the average attributable fraction and its standard error. Note that the R-package *averisk* utilizes logistic (generalized linear models with logit link) models to predict the per-individual probabilities of disease that are used when estimating sequential attributable fractions, as given by (10). It should be noted that other plausible statistical models can also be used at this step if desired.

While we anticipate that our work will be helpful in the application of average attributable fractions, there are still a number of methodological issues that need to be addressed. First, while in theory the definition of attributable fraction for a continuous risk factor is well defined, no current software allows its calculation. While dichotomizing a risk factor such as blood pressure into two (or more) groups representing normal and elevated values represents a practical solution, the resulting estimated attributable fraction will probably underestimate the effect of optimal blood pressure control on disease risk. A second problem that we do not address here relates to the calculation of average attributable fractions based on survival data. Extensions to the formulas for population attributable fraction for censored data have been proposed[11], and could be adapted to calculate average attributable fractions. Another issue that has been discussed in the literature relates to the sequence in which risk factors are eliminated[23]. Taking an epidemiological view, risk factors are not all equally modifiable; some are easier to target via public health interventions than others. Hypothetically, if there was only one possible order in which the risk factors could be removed from the population and this order was known, the correct partition of the combined attributable fraction for the collection of risk factors would correspond to the sequential attributable fractions associated with this correct removal order. In general, every permutation corresponds directly to a particular order of elimination. With this in mind, if certain permutations are more plausible than others (for example, if the risk factors appearing in the first couple of positions are most easily targeted by interventions), a modified version of $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ might be calculated using a weighted random sample of permutations with such plausible permutations having higher probability of being sampled. An alternative approach to partitioning the combined attributable risk which ignores removal order altogether, was considered in Llorca and Rodríguez[14] in 2004.

The idea of sampling permutations might be worrying for some researchers, since it introduces a small extra amount of randomness into the calculated $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$. We do not view this as a serious problem, as the approximation error can usually be well controlled with just 1000 permutations, and will usually be negligible compared to the variance in the sampling distribution of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$, and is incorporated

into our confidence interval. However, another idea that is worth investigating is to choose a subset of permutations judiciously, rather than taking a random sample. Choosing m permutations so they are maximally spaced from each other according to some metric, and averaging the sequential attributable fractions for these carefully chosen permutations, allows us to remove the randomness from the calculated $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$, and most probably improve the approximation error compared to m randomly sampled permutations. This is certainly an idea for future investigation; although it is unclear if the approximation error can be incorporated into a confidence interval in a theoretically convenient way as in the case for randomly sampled permutations. Our methods work both for case-control and prospective studies; although when using $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ or $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ for case-control studies as suggested in this manuscript, an estimate of disease prevalence is needed. If there is uncertainty regarding the correct estimate, a sensitivity analysis can be performed, where $\hat{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ or $\tilde{\boldsymbol{\lambda}}_{\boldsymbol{a}}$ is calculated over a range of plausible prevalences. In our experience, the prevalence-based weighting procedure described in this this manuscript is quite robust to small errors in the specified prevalence, provided the prevalence lies within reasonable bounds. As an example, for the simulations based on the Hordaland study discussed in Section 4, we can show that the estimated average attributable fractions computed using any estimated prevalence between 0.08 and 0.10 differ by less than 0.002 for all risk factors. See the Supplementary Material for more details.

References

[1] P Bakke, GE Eide, R Hanoa, and A Gulsvik. Occupational dust or gas exposure and prevalences of respiratory symptoms and asthma in a general population. *European Respiratory Journal*, 4(3):273–278, 1991.

[2] Paolo Bruzzi, Sylvan B Green, David P Byar, Louise A Brinton, and Catherine Schairer. Estimating the population attributable risk for multiple risk factors using case-control data. *American journal of epidemiology*, 122(5):904–914, 1985.

[3] Philip Cole and Brian MacMahon. Attributable risk percent in case-control studies. *British journal of preventive and social medicine*, 25(4):242–244, 1971.

[4] Steven S Coughlin, Jacques Benichou, and Douglas L Weed. Attributable risk estimation in case-control studies. *Epidemiologic Reviews*, 16(1):51–64, 1994.

[5] B Effron and R Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):p54–77, 1986.

[6] Geir Egil Eide and Olaf Gefeller. Sequential and average attributable fractions as aids in the selection of preventive strategies. *Journal of clinical epidemiology*, 48(5):645–655, 1995.

[7] Geir Egil Eide and Ivar Heuch. Attributable fractions: fundamental concepts and their visualization. *Statistical Methods in Medical Research*, 10(3):159–193, 2001.

[8] Geir Egil Eide and Ivar Heuch. Average attributable fractions: a coherent theory for apportioning excess risk to individual risk factors and subpopulations. *Biometrical journal*, 48(5):820–837, 2006.

[9] Sander Greenland and Karsten Drescher. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics*, pages 865–872, 1993.

[10] Ulrike Grömping and Ulla Weimann. The asymptotic distribution of the partial attributable risk in cross-sectional studies. *Statistics*, 38(5):427–438, 2004.

[11] MA Laaksonen, T Härkänen, P Knekt, E Virtala, and H Oja. Estimation of population attributable fraction (paf) for disease occurrence in a cohort study design. *Statistics in medicine*, 29(7-8):860–874, 2010.

[12] Matthias Land and Olaf Gefeller. A game theoretic approach to partitioning attributable risks in epidemiology. *Biometrical Journal*, 39(7):777–792, 1997.

[13] Mark L Levin. The occurrence of lung cancer in man. *Acta-Unio Internationalis Contra Cancrum*, 9(3):531–541, 1953.

[14] Javier Llorca and Miguel Delgado-Rodríguez. A new way to estimate the contribution of a risk factor in populations avoided nonadditivity. *Journal of clinical epidemiology*, 57(5):479–483, 2004.

[15] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[16] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[17] Alexander K Rowe, Kenneth E Powell, and W Dana Flanders. Why population attributable fractions can sum to more than one. *American journal of preventive medicine*, 26(3):243–249, 2004.

[18] Simon Ruckinger, Rudiger von Kries, and Andre M Toschke. An illustration of and programs estimating attributable fractions in large scale surveys considering multiple risk factors. *BMC medical research methodology*, 9(1):7, 2009.

[19] Lloyd S Shapley. A value for n-person games. Technical report, DTIC Document, 1952.

[20] Yvo M Smulders, Abel Thijs, and Jos W Twisk. New cardiovascular risk determinants do exist and are clinically useful. *European Heart Journal*, 29(4):436–440, 2008.

[21] Mark J van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, 4(1), 2008.

[22] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

[23] SD Walter. Attributable risk in practice. *American journal of epidemiology*, 148(5):411–413, 1998.

[24] Stephen D Walter. Prevention of multifactorial disease. *American Journal of Epidemiology*, 112(3):409–416, 1980.

[25] Stephen D Walter. Effects of interaction, confounding and observational error on attributable risk estimation. *American Journal of Epidemiology*, 117(5):598–604, 1983.