Variants of the Attributable Risk in a Multifactorial Situation: Theory and Computational Realization

Olaf Gefeller

Department of Medical Statistics, University of Göttingen

ABSTRACT

Statistical methods for quantifying the population impact of exposure factors by means of estimating some variant of the attributable risk parameter for a binary disease variable based on stratification methods or multivariable models have been extensively developed during the last decade. In epidemiologic studies, these risk parameters enjoy growing popularity as they allow a practice-oriented interpretation for public health purposes. Despite this demand by epidemiologic researchers, none of the leading statistical software packages including \textsc{SAS}\textsuperscript{TM} has incorporated any attributable risk parameter in the catalogue of measures of association to be automatically calculated in some procedure. In the paper, an introduction to the theory and application of attributable risk estimation from epidemiologic data is provided. Then, a general method and its corresponding practical realization are described how to calculate different attributable risk estimators using the \textsc{SAS} software. Data from the German GRIPS–study, a cohort study on cardiovascular risk factors for myocardial infarction, are used to illustrate the methodologic concept and its computational realization.

1. Introduction

Analytical epidemiologic studies are conducted to provide information on the determinants of some disease variable. Usually, the disease etiology under study is of a multifactorial nature so that several exposure factors have to be considered simultaneously. The effect of a particular exposure characteristic on the binary disease variable is then quantified by some measure of association among which, as epidemiologic risk measures, the relative risk (\textit{RR}), the odds ratio (\textit{OR}), and the attributable risk (\textit{AR}) are the most popular representatives. This paper focuses on the latter measure which can be informally introduced as the answer to the question, "What proportion of the observed cases in the study population suffers from the disease due to the exposure factor of interest?". In providing this information the attributable risk places the concept of relative risk commonly used...
in epidemiologic studies for the evaluation of the etiologic relevance of the exposure on
the individual level in a public health perspective.

The topic of attributable risk estimation has become an important subject of method-
ologic investigation in recent years (Gefeller, 1992a). The model situation in which several
exposure factors have to be taken into account has been analyzed extensively, and the
concept of adjusted attributable risks has been elaborated by several authors. This concept
applies to the situation in which the exposure impact of the corresponding factors has to
be controlled for the effects of other covariables. In this paper, the multifactorial case is
also considered, however, the focus lies here on the situation characterized by an equal-
ranking of exposure variables in the analysis. Thus, another conceptual approach to the
estimation of attributable risks for single exposure factors is required in this case. The
concepts of sequential and partial attributable risk for partitioning the disease risk into
components for the exposure variables are introduced and illustrated by an example from
a German cohort study on risk factors for myocardial infarction. Additionally, detailed
comments on the computational realization of the methods are provided in a separate
section.

2. Basic Definitions

Suppose a population can be divided into an exposed subpopulation (E) and an unexposed
one (\(\bar{E}\)) as well as a diseased part (D) and a non–diseased one (\(\bar{D}\)). Denote \(P(A)\) the
probability that a randomly chosen subject from this population belongs to subpopulation
A, and \(P(A \mid B)\) the corresponding conditional probability of A given B. The definition
of the attributable risk, originally introduced by Levin (1953), is then as follows:

\[
AR = \frac{P(D) - P(D \mid E)}{P(D)}.
\] (1)

Alternatively, the attributable risk can be expressed as

\[
AR = \frac{P(E) \cdot [RR - 1]}{P(E) \cdot [RR - 1] + 1},
\] (2)

or

\[
AR = P(E \mid D) \cdot \frac{RR - 1}{RR}.
\] (3)

Under the assumption of a causal relationship between exposure and disease the at-
tributable risk can be interpreted as the proportion of cases due to the exposure among
all cases of disease in the population. Alternatively, \(AR\) may also be understood as the
proportion of cases potentially preventable by a total elimination of the exposure in the
population. Already Levin recognized that these interpretations are based on the assump-
tion that exposed subjects, if they had not become exposed, would have had the same
probability of developing a disease event as that found among unexposed subjects.
3. Crude and Adjusted Attributable Risks

The maximum likelihood estimator of $AR$ can be easily obtained in $2 \times 2$ tables from cohort and cross-sectional studies by substituting sample proportions for the respective probabilities in (1) leading to what has been termed crude estimators of the attributable risk (Walter, 1976). Some additional approximation (i.e. replacing $RR$ by $OR$ in (3)) is necessary for the case-control design (Benichou, 1991). Typically in observational epidemiologic studies the exposure-disease relationship is affected by other covariables leading to a biased attributable risk estimate when the estimation is based on $2 \times 2$ tables (Walter, 1980; Walter, 1983). In the epidemiologic context this distortion of the estimation is referred to by the terms “confounding” and “effect modification”. In both (conceptually very different) situations the estimation of the attributable risk has been approached by some adjustment procedure taking the effects of the covariable(s) into account which is based on the following ideas:

Suppose the explanatory variables can be defined into a set of $L$ exposure factors of primary interest and another set of $M$ adjusting variables. Let all combinations of values for the adjusting variables define a total of $S$ strata $C_s, s = 1, \ldots, S$. Furthermore, let all combinations of values for the exposure factors define $K + 1$ exposure classes $E_k, k = 0, 1, \ldots, K$. Usually, $k = 0$ indicates the class of no (or lowest) exposure previously denoted by $E$. The ordering of the latter $k$ classes is arbitrary. Then, definition (1) generalizes to the adjusted attributable risk given by

$$AR_{adj} = 1 - \frac{\sum_{s=1}^{S} P(D \mid E_0, C_s) \cdot P(C_s)}{P(D)}. \quad (4)$$

This adjusted attributable risk for the total effect of all $L$ exposure factors may be interpreted as the proportion of the diseased population that is potentially preventable if the risks of disease in the exposed subpopulations were changed to the risks of the unexposed ($E_0$) population in all strata of the adjusting variables, respectively. Estimation of $AR_{adj}$ based on stratification methods (Whittemore, 1982; Gefeller, 1992b) or on a logistic regression approach (Bruzzi et al., 1985) have been investigated in detail during the last decade. Mostly, the special case $L = 1$ is considered so that the adjusted attributable risk for a single exposure factor is calculated, all other variables are then used for the purpose of an appropriate adjustment. This model scenario is only reasonable whenever the specific aim of the study is to evaluate the role of one particular exposure factor. Otherwise, the implicit hierarchy imposed on the variables involved in the calculation is not justified and another approach to attributable risk estimation is required (Gefeller & Eide, 1993).
4. Sequential Attributable Risks

In this section we will describe a stepwise algorithm for calculating attributable risks in a multifactorial setting leading to the definition of sequential attributable risks and a proposal of the optimal preventive strategy for the elimination of exposures with respect to the greatest population impact for a given number of exposure factors to be eliminated. Suppose the $K + 1$ exposure classes are generated by $L$ exposure factors each with $K_i + 1$ exposure categories, i.e. $K + 1 = \prod_{i=1}^{L}(K_i + 1)$. Our interest lies in the potential reduction of cases when preventing the $L$ exposures, one at a time, in a given sequence, for instance starting with exposure no. 1, then exposure no. 2, and so on, until all exposures are eliminated from the population. A reasonable way to accomplish this will be first to calculate the adjusted attributable risk as shown in the previous section with all exposure factors but the first one included among the adjusting variables. This results in an adjusted attributable risk denoted by $AR_{adj}^{(1)}$ derived from a situation with $S \cdot \prod_{i=2}^{L}(K_i + 1)$ strata and $K_1 + 1$ exposure classes. Thereafter, define $AR_{adj}^{(2)}$ as the adjusted attributable risk calculated for the combined effect of first and second exposure variable (creating $(K_1 + 1) \cdot (K_2 + 1)$ exposure classes), and the remaining exposures and the adjusting variables forming the $S \cdot \prod_{i=3}^{L}(K_i + 1)$ strata. This stepwise procedure of calculating adjusted attributable risk for different sets of exposure variables can be continued until all $L$ exposure factors are incorporated among the exposure classes generating variables. The last term in this sequence $AR_{adj}^{(L)}$ corresponds to the total population impact of all $L$ exposures controlled for the effect of the $M$ adjusting variables forming the $S$ strata.

Any difference $AR_{adj}^{(r)} - AR_{adj}^{(p)}$, $p < r$, describes the additional effect of considering the $(p + 1)$st, $(p + 2)$nd, ..., $r$-th exposure after having previously taken into account the effect of the first $p$ exposures in the specified sequence. These differences may be called sequential attributable risks (SAR). Notice that the SAR of a specific exposure factor may differ even for the same set of $L$ exposures according to the sequence of exposure variables considered during the stepwise process of calculation.

The $L$ exposure factors will lead to $L!$ different sequences of variables. Then, ignoring all complicating issues (e.g. costs, availability of intervention measures) for the moment, an optimal strategy of eliminating exposures to reduce the disease burden of the population would obviously be to remove first the single exposure among all $L$ variables which gives rise to the highest $AR_{adj}^{(1)}$, and next, to remove the exposure among the remaining $L - 1$ which, combined with the eliminated one, leads to the highest $AR_{adj}^{(2)}$, and so forth, until all exposures are removed (Eide & Gefeller, 1994).

5. Partial Attributable Risks

As the sequential attributable risks introduced in the previous section do not yield a unique value for a particular exposure factor, the problem of an unambiguous quantification of
the contribution of one exposure to the disease load in a population in a multifactorial situation under the assumption of an equal-ranking of factors needs further examination. The total of \( L \) exposure variables will give rise to \( L! \) sequential attributable risks for a given exposure. However, among the \( L! \) different orderings some will yield an identical value of the \( SAR \) for the exposure under study as all permutations of exposure variables within the two groups of factors, those to calculate an adjusted attributable risk for and those to adjust for, respectively, do not change the particular \( SAR \). Nevertheless, \( 2^{L-1} \) different values for the \( SAR \) of interest are possible due to the variety of constellations of other exposure variables which have to be taken into account when calculating adjusted attributable risks. A way out of this dilemma is to average the \( SARs \) for one exposure over all orderings. This procedure is an application of the principle suggested by Kruskal (1987) for determining relative importance for independent variables in a multiple regression setting. Also Cox (1985) gives a formal justification for averaging \( SAR \)-like parameters.

The resulting unique measure of attributable risk in a multifactorial setting can be used for the purpose of partitioning the total attributable risk for a set of exposure variables among the elements of the set. Adapting Cox's arguments drawn from mathematical game theory to this context it can be easily proved that the arithmetic mean of all \( SAR \) values for a given exposure, which we will call partial attributable risk (\( PAR \)), is the only solution to the problem of attributable risk partitioning which satisfies three natural conditions: (i) exposure factors independent from the disease variable receive a \( PAR \) value of zero, (ii) numbering of exposure factors does not influence the risk partition, (iii) additivity of risk partitions holds for independent components. Further discussion of this approach in a legal context can be found in Lagakos & Mosteller (1986). A practical application of this concept mentioned by Cox relates to tort-law liability issues when the responsibility for the disease load in a population is shared between several parties who may be held responsible for the different exposures and it is the intention to quantify the exact contribution of each factor as an aid for individual compensation (Cox, 1987).

6. Aspects of the Computational Realization

Although the statistical theory for inference on the attributable risk has been worked out in considerable detail and practical applications of the concept can be frequently found in the epidemiologic literature, the computational realization of these methods has not received much attention until now. None of the leading statistical software packages (e.g. SAS, BMDP, SYSTAT, SPSS) offers procedures to analyze epidemiologic data in terms of the attributable risk. Even specialized software packages for epidemiologic applications as, for example, EGRET, EPISTAT and EPILOG do not comprise the opportunity to calculate attributable risks. Five simple spreadsheet programs (LIVESAVER, SAMNEC, PAR, PARUBC, CANTROL) have been developed for elementary applications, but none
of these incorporate any more sophisticated statistical features (Milsum & Jones, 1988).

A general method of adapting statistical software originally designed for categorical data analysis according to the Grizzle, Starmer & Koch (1969) approach to the problem of estimating complex measures of association has been recently described by Gefeller & Woltering (1993). Using the SAS package (SAS Institute Inc., 1990) this idea can be easily realized employing the CATMOD procedure of the SAS/STAT™ software (Gefeller & Woltering, 1991). A detailed description of the application of this general technique to the problem of estimating crude and adjusted attributable risks defined in (1) and (4) using SAS, giving the necessary SAS programming code, has been published previously (Gefeller, 1991). The additional computational effort required to calculate sequential and partial attributable risks from the particular set of adjusted attributable risks is negligible and can be approached by data step programming or, equivalently, by special SAS/IML™ programs. The most cumbersome part of the computational process relates to the calculation of all adjusted attributable risks for different orderings of exposure variables. This task may become excessive as the number of exposure factors considered in the analysis increases.

7. Example: The G.R.I.P.S. Study

Data of the G.R.I.P.S. study (Göttingen, Risk, Incidence, and Prevalence Study), a cohort study on 6029 men aged 40 to 60 years, designed to analyze the influence of potential risk factors for myocardial infarction (Cremer et al., 1991), are used to illustrate the different concepts of attributable risk. For the purpose of illustration, we focus on the effect of the three lipoprotein fractions (HDL-, LDL- and VLDL-cholesterol) and cigarette smoking as the exposure variables of interest. Age, familiar disposition to myocardial infarction, alcohol consumption, physical activity, blood glucose, and blood pressure were included in the analysis as adjusting variables. All analyses were performed within the SAS software (see previous section). Table 1 shows estimated values for crude attributable

<table>
<thead>
<tr>
<th>Exposure factor</th>
<th>Definition of &quot;unexposed&quot;</th>
<th>Estimators of crude $AR$ and asy. variance in ($)</th>
<th>Estimators of $AR_{adj}$ and asy. variance in ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL-cholesterol</td>
<td>$&lt; 160$ mg/dl</td>
<td>0.6121 (0.0037)</td>
<td>0.5773 (0.0046)</td>
</tr>
<tr>
<td>HDL-cholesterol</td>
<td>$&gt; 35$ mg/dl</td>
<td>0.2044 (0.0025)</td>
<td>0.1722 (0.0032)</td>
</tr>
<tr>
<td>VLDL-cholesterol</td>
<td>$&lt; 30$ mg/dl</td>
<td>0.2169 (0.0038)</td>
<td>0.1669 (0.0062)</td>
</tr>
<tr>
<td>cigarette smoking</td>
<td>no</td>
<td>0.3710 (0.0067)</td>
<td>0.3697 (0.0080)</td>
</tr>
</tbody>
</table>

Table 1: Estimates of crude and adjusted attributable risks and their asymptotic variances for four exposure factors of myocardial infarction in the G.R.I.P.S. study
risks (derived from the corresponding $2 \times 2$ tables) and adjusted attributable risks (derived by incorporating a logistic regression modelling approach) of the four exposure variables of primary interest as well as the corresponding estimates of the asymptotic variances. From the comparison of crude and adjusted values it is evident that some confounding of the relationship between the lipoprotein exposure variables and myocardial infarction by the adjusting variables is present, however, this observation does not hold for cigarette smoking.

As an example for the stepwise procedure of estimating sequential attributable risks, the effect of eliminating LDL-cholesterol as first, second, third and fourth exposure is shown in Table 2. Notice that the value of the estimated adjusted attributable risk for elevated LDL-cholesterol given in Table 1 can also be found in the first row of Table 2. All other SAR values quantify the additional impact of eliminating the LDL-cholesterol exposure after the effect of other exposure factors has already been taken into account.

**Table 2**: Sequential attributable risks of elevated LDL-cholesterol for myocardial infarction in the G.R.I.P.S. study

<table>
<thead>
<tr>
<th>Elimination of LDL-cholesterol as $j$-th variable</th>
<th>Exposures previously eliminated</th>
<th>Estimates of the sequential attributable risk SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j = 1$ $i = 0$ none</td>
<td></td>
<td>0.5773</td>
</tr>
<tr>
<td>$j = 2$ $i = 1$ HDL-cholesterol</td>
<td></td>
<td>0.4675</td>
</tr>
<tr>
<td></td>
<td>VLDL-cholesterol</td>
<td>0.4603</td>
</tr>
<tr>
<td></td>
<td>cigarette smoking</td>
<td>0.3412</td>
</tr>
<tr>
<td>$j = 3$ $i = 2$ HDL + VLDL</td>
<td></td>
<td>0.4097</td>
</tr>
<tr>
<td></td>
<td>HDL + cig.</td>
<td>0.3106</td>
</tr>
<tr>
<td></td>
<td>VLDL + cig.</td>
<td>0.3018</td>
</tr>
<tr>
<td>$j = 4$ $i = 3$ HDL, VLDL + cig.</td>
<td></td>
<td>0.2780</td>
</tr>
</tbody>
</table>

Finally, averaging all SAR values for the different exposure factors yields the partial attributable risks as shown in Table 3. However, it should be recognized that not only the eight numerically different SAR values (displayed for elevated LDL-cholesterol in Table 2) have to be averaged. Altogether $4! = 24$ SAR values originate from a model situation with four factors for each exposure variable. As explained in section 5, several of these values are identical due to the invariance of the SAR for certain permutations.
of exposures. For example, the first SAR values in Table 2 (0.5773) and also the last one (0.2780) both appear six times in the complete list of all 24 SAR values (all other SAR values of Table 2 appear only twice). Thus, when averaging only over all numerically different SAR values, an appropriate weighting scheme has to be employed.

Due to their construction the partial attributable risks in Table 3 reveal an additive property, which means that the sum of all PAR values (=0.7697) equals the total effect of all four exposures measured by the corresponding adjusted attributable risk as calculated from expression (4). Consequently, in any situation the sum of PAR values cannot exceed the natural limit of one. In the light of the confused discussion about adding crude and adjusted attributable risks of exposure variables (O’Neill, 1991; Gefeller & Eide, 1992) this property of the partial attributable risk approach has to be viewed as a strong advantage of the concept with respect to the interpretation of the measure in a multifactorial setting.

Table 3: Partial attributable risks of four exposure factors for myocardial infarction in the G.R.I.P.S. study

<table>
<thead>
<tr>
<th>Factor</th>
<th>Definition of “unexposed”</th>
<th>Estimates of PAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL–cholesterol</td>
<td>&lt; 160 mg/dl</td>
<td>0.4048</td>
</tr>
<tr>
<td>HDL–cholesterol</td>
<td>&gt; 35 mg/dl</td>
<td>0.0796</td>
</tr>
<tr>
<td>VLDL–cholesterol</td>
<td>&lt; 30 mg/dl</td>
<td>0.0742</td>
</tr>
<tr>
<td>Cigarette smoking</td>
<td>no</td>
<td>0.2111</td>
</tr>
<tr>
<td>Total effect of all four factors</td>
<td>combining all four above by “∧”</td>
<td>0.7697</td>
</tr>
</tbody>
</table>

8. Discussion

The estimation of attributable risks from epidemiologic data forms an integral part of modern analytical approaches quantifying the relationship between some binary disease variable and exposure factors. Whereas the relative risk may be thought of as indicating the strength of the physiologic effects of an exposure to the disease, the attributable risk combines this information on the individual hazard of the exposure with its prevalence in the population under study. Thus, the attributable risk is the appropriate epidemiologic risk measure for evaluating the public health importance of an exposure factor for the disease load of a population and should guide health administrators to rational planning of prevention campaigns.

Statistical methods for estimating the attributable risk in a variety of design situations have been developed in the methodologic literature. Especially the multifactorial situation has attracted considerable research activities, since mostly the disease variable under study exhibits a multifactorial etiology. In addition to adjustment procedures frequently
investigated during the last decade, this paper has delineated the concepts of sequential and partial attributable risk which supplements the repertoire of epidemiologic approaches in this situation.

As a final remark, we would like to draw the attention to the inherent problems of the interpretation of any attributable risk estimate (Greenland & Robins, 1988). The calculation of an attributable risk in a given study revealing some positive value for AR, say 50%, does by no means guarantee that one half of the observed cases in the study population can be causally attributed to the exposure of interest and are “saved” when the exposure is eliminated from the population, even if all obstacles of the statistical calculation have been properly taken. The statistical analysis of epidemiologic data is only one (important) part belonging to the evaluation of the relationship between exposure and disease which has to be supplemented by medical reasoning. The particular problem of the interpretation of attributable risk estimates results from the hardly solvable linguistic dilemma that in each simple description of the estimates it cannot be avoided to implicitly anticipate a causal relationship by using such phrases as “attributable/due to”. Consequently, statisticians have to be very cautious in communicating their results of attributable risk calculations to physicians and other public health professionals.

Acknowledgements

The author gratefully acknowledges the support given by Prof. Dr. G.E. Eide, University of Bergen (Norway), in the form of stimulating discussions on the topic of this paper. The author’s “Habilitationsschrift” at the Medical Faculty of the Georg-August-University of Göttingen is partly reflected in this paper.

References


**Address for correspondence:**

Priv.-Doz. Dr. Olaf Gefeller  
Abteilung für Medizinische Statistik  
Georg-August-Universität Göttingen  
Humboldtallee 32  
D-37073 Göttingen  
Germany

*SAS, SAS/IML, and SAS/STAT* are registered trademarks of SAS Institute Inc., Cary, NC, USA.