Original Contribution

Estimating the Proportion of Disease due to Classes of Sufficient Causes

Kurt Hoffmann, Christin Heidemann, Cornelia Weikert, Matthias B. Schulze, and Heiner Boeing

From the Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbrücke, Germany.

Received for publication April 21, 2005; accepted for publication August 17, 2005.

Disease can be caused by different mechanisms. A possible causal model proposed by Rothman is a complete causal mechanism or a so-called “sufficient cause” consisting of a set of component causes that can be illustrated in a pie chart. However, this model does not allow finding out what sufficient causes produce the majority of cases. The authors’ objective was to extend Rothman’s work by quantifying the proportion of disease that can be attributed to a class of sufficient causes. The underlying idea was to consider all combinations of a given set of known risk factors and to assign each combination to a class of sufficient causes. This assignment makes it possible to evaluate a class of sufficient causes by the population attributable fraction of the corresponding combination of risk factors. The approach presented was applied to sufficient causes of myocardial infarction by use of data on participants recruited between 1994 and 1998 into the European Prospective Investigation into Cancer and Nutrition-Potsdam Study. As a result, 51.8% of cases were attributed to only four different classes of sufficient causes. In conclusion, the statistical method described in the paper may be beneficial for quantifying the importance of different sufficient causes and for improving the efficiency of public health programs.

models, statistical; myocardial infarction; risk factors; statistics

Abbreviations: EPIC, European Prospective Investigation into Cancer and Nutrition; PAF, population attributable fraction; PDC, proportion of disease due to a class of sufficient causes.

Rothman’s model of sufficient and component causes (1, 2) introduced in 1976 is one of the most discussed causal models in epidemiology. Despite some inherent limitations (3, 4), this model seems to be appropriate to reflect the multiplicity of causal pathways and the biologic interaction among component causes. The pie-chart description of possible classes of sufficient causes is an illustrative method to differentiate among distinct etiologic mechanisms, leaving some unlabeled slices to represent unknown component causes. However, up to now, Rothman’s model is only a theoretical framework for epidemiology, with no direct connection to empirical data. Although epidemiologic studies on a specific disease should give information as to which of the possible sufficient causes probably exist and which do not, no approach has been suggested to get quantitative results. In other words, the application of the sufficient-component cause model in epidemiologic research is hampered by lack of an appropriate statistical method for estimating the proportion of disease due to different sufficient causes. Thus, an extension of Rothman’s work is needed.

An obvious approach to find out the most relevant sufficient causes is based on calculating the frequencies of component causes and of their combinations within cases (5). However, a high frequency of a specific combination of component causes within cases does not necessarily mean that this combination is a sufficient cause of high relevance, because the corresponding frequency within noncases may also be high. Moreover, this approach does not allow adjustment for other known factors and is not related to the common notions and concepts of epidemiology. In general, the importance of sufficient causes depends on both the relative risks and the frequencies of all possible component causes.
SUFFICIENT CAUSES AND POPULATION ATTRIBUTABLE FRACTION

A “sufficient cause” is defined as a set of minimal conditions that inevitably produce disease (2). It consists of a number of component causes. Each component cause is necessary for the completion of the sufficient cause. The completion of a sufficient cause is equivalent to the onset of the earliest stage of disease. In other words, the component cause that acts last determines the time of the onset. If one or more of the component causes are absent, the causal mechanism cannot operate.

Two sufficient causes can be similar because of having some of the same components. Differentiating between these sufficient causes can possibly be difficult, since some of the distinct component causes are often not known or not measured. Leaving some pieces unlabeled in a pie chart is not only a way to represent unknown component causes but also a possibility to group similar sufficient causes. Thus, any pie chart with an unlabeled piece can be considered a class of sufficient causes. The common characteristic of sufficient causes belonging to the same class is the same set of known component causes. Formally, such classes can also include sufficient causes with unpredictable components to include random events (6).

For the sake of simplicity, we first regard the case of only two known component causes, for example, $X_1$ and $X_2$. Then, four different classes of sufficient causes can be distinguished as shown in figure 1. The class $S_1$ comprises all sufficient causes that have $X_1$, but not $X_2$, as the component cause. The symbol $U_1$ stands for one or more unknown component causes. Analogously, $S_2$ is defined as the class of sufficient causes that contains $X_2$ as the only known component. $S_{12}$ stands for causal mechanisms that require the presence of both $X_1$ and $X_2$. Finally, the class $S_0$ comprises all sufficient causes that produce disease in the absence of $X_1$ and $X_2$. Obviously, any single sufficient cause belongs to one and only one of the four classes depicted in figure 1.

Now, consider the PAF of the component causes $X_1$ and $X_2$. The PAF of a single condition is defined here as the fraction of cases that would not have occurred if the condition were not present (7). If stratification successfully removes confounding, the PAF can be validly estimated by the formula of Miettinen (8) that requires estimates of the condition prevalence among cases and the risk ratio standardized to the condition (2, p. 295). Analogously, the joint PAF of two conditions defined as the fraction of cases that would not have occurred if both conditions were not present can also be estimated by Miettinen’s formula. Denoting the proportion of disease due to a class $S$ of sufficient causes by $PDC(S)$, the following four equations hold.

$$PAF(X_1) = PDC(S_1) + PDC(S_{12})$$
$$PAF(X_2) = PDC(S_2) + PDC(S_{12})$$
$$PAF(X_1, X_2) = PDC(S_1) + PDC(S_2) + PDC(S_{12})$$
$$PAF(X_1, X_2) = 1 - PDC(S_0).$$

The equations can be verified by the pie charts of figure 1. First, consider the case that $X_1$ has not occurred. Then, sufficient causes of types $S_1$ and $S_{12}$ could not have been completed, whereas the completion of all other sufficient causes would not have been affected. Therefore, the fraction of cases that can be attributed to $X_1$ must be the sum of the proportions of disease caused by $S_1$ and $S_{12}$ as stated by the first equation. The second equation has a quite similar interpretation, but here $X_2$ plays the role of $X_1$. The remaining two equations refer to the joint PAF of both component causes. A generalization of these equations to the situation of more than two known component causes is straightforward. In general, the joint PAF of any $m$ conditions can be calculated by summing up the PDC values of all sufficient causes that contain at least one of the $m$ conditions as component cause.
The equations 1 are not suitable to calculate the frequencies of sufficient causes from the single and joint PAF of component causes. For this purpose, the equations 1 should be rewritten in the following form:

\[
\begin{align*}
\text{PDC}(S_1) &= \text{PAF}(X_1, X_2) - \text{PAF}(X_2) \\
\text{PDC}(S_2) &= \text{PAF}(X_1, X_2) - \text{PAF}(X_1) \\
\text{PDC}(S_{12}) &= \text{PAF}(X_1) + \text{PAF}(X_2) - \text{PAF}(X_1, X_2) \\
\text{PDC}(S_0) &= 1 - \text{PAF}(X_1, X_2).
\end{align*}
\] (2)

To get an impression of equations 2, we considered a very strong component cause \(X_1\) with a PAF of 0.5 and a strong component cause \(X_2\) with a PAF of 0.3. For simplicity, we assumed that both components are risk factors and that the presence of one of the components is not protective for disease in any etiologic mechanism. Then, the joint PAF cannot be smaller than the larger of each individual PAF and cannot be larger than their sum. Therefore, in this example, the joint PAF can vary only between 0.5 and 0.8.

For different values of the joint PAF, table 1 gives the frequency distribution for sufficient causes. Obviously, for the increasing value of the joint PAF, the proportions of \(S_1\) and \(S_2\) also increase, whereas the proportions of \(S_{12}\) and \(S_0\) decrease. In the case of the minimal joint PAF, no sufficient cause that contains only the weaker component \(X_2\) can exist. Thus, the proportion of disease due to \(S_2\) has to be zero. More interesting is the opposite case of the maximal joint PAF. Here, the joint PAF is equal to the sum of single PAFs, which means that no interaction in public health contexts between \(X_1\) and \(X_2\) exists (9, 10). As we can see in the last row of table 1, the additivity of attributable fractions implies that both causes \(X_1\) and \(X_2\) cannot belong to a common sufficient cause. In other words, they act in different etiologic mechanisms and have no biologic interaction (2). Thus, the two concepts of public health and biologic interaction have the same referent point corresponding to PAF(\(X_1, X_2\) = 0.8.

The equations 2 can be generalized to the case of more than two component causes. Let \(k\) be the number of known component causes, and then \(2^k\) classes of sufficient causes can be distinguished. To calculate the PDC for all \(2^k\) classes from the PAF of the component causes, the same number of equations is necessary. These \(2^k\) equations are derived in appendix 1.

### EVALUATING SUFFICIENT CAUSES

Up to now, we have only a vague impression of PDC as the proportion of cases that is due to a specific class of sufficient causes. Indeed, we need a precise definition of this notion by using the terminology of probability theory. For this purpose, consider the class of sufficient causes that contains from the \(k\) known factors \(X_1, \ldots, X_k\) only the first \(m\) ones as component causes. We denote this multifactorial event by \(E = (X_1 = 1, \ldots, X_m = 1, X_{m+1} = 0, \ldots, X_k = 0)\) and the corresponding class of sufficient causes by \(S_E\). Obviously, the PDC of \(S_E\) cannot be larger than the proportion of cases that had the event \(E\) at the onset of disease. Formally, the PDC is bounded above by \(P(E|D)\) that denotes the conditional probability of the event \(E\) under the condition that the individual develops the disease \(D\). However, within the stratum of cases that have the event \(E\), a fraction of individuals would also develop the disease if all known component causes of \(S_E\) were absent before. The remaining fraction in this stratum is equal to the proportion of cases that is due to \(S_E\). Thus, we get the definition,

\[
\text{PDC}(S_E) = P(E|D) \frac{P(D|E) - P(D|E_0)}{P(D|E)},
\] (3)

where \(E_0\) stands for the zero event \((X_1 = 0, \ldots, X_k = 0)\), whereas \(P(D|E)\) and \(P(D|E_0)\) are the disease rates in the strata defined by \(E\) and \(E_0\), respectively. Throughout the paper, we assume that the disease rate among the individuals with event \(E\) exceeds the disease rate among the unexposed. Dividing the numerator and the denominator of the right-hand term by \(P(D|E_0)\) yields the representation,

\[
\text{PDC}(S_E) = P(E|D) \frac{\text{RR} - 1}{\text{RR}},
\] (4)

with \(\text{RR}\) denoting the relative risk of the disease in individuals with event \(E\) versus individuals with \(E_0\). Obviously, equation 4 is similar to the formula of Miettinen (8) for attributable fraction. Actually, the PDC of the class \(S_E\) of sufficient causes is equal to the PAF of the multifactorial event \(E\). By definition, the class of sufficient causes and the multifactorial event form an inseparable pair. The \(2^k\) different combinations for presence and absence of the \(k\) known factors define the same number of strata in the sample and the same number of classes of sufficient causes.

It can be proven that the definition of PDC by equation 3 or equation 4 is consistent with the relations between PDC and PAF represented before. Actually, equation 4 gives a solution of the equation systems 1 and 2 and its generalization to the multicomponent case (appendix 2). Because \(2^k\) equations for \(2^k\) different PDC variables cannot have more than one solution for each PDC, the solution just given is the unique one.

### TABLE 1. Relation between the population attributable fraction of component causes and the frequency distribution of sufficient causes in a fictive example

<table>
<thead>
<tr>
<th>Population attributable fraction</th>
<th>Frequency of classes of sufficient causes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>(X_2)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
</tr>
</tbody>
</table>

* Frequencies were calculated from the population attributable fractions by use of equation 2 given in the paper.

† \(S_i\) is the class of sufficient causes containing \(X_1\) and not \(X_2\) as component cause. \(S_2\) contains \(X_2\) and not \(X_1\). \(S_{12}\) contains both \(X_1\) and \(X_2\), and \(S_0\) contains neither \(X_1\) nor \(X_2\).
Equation 4 is suitable to get a valid estimate of PDC from any valid estimate of RR. Because equation 4 is the PAF formula of Miettinen for the event $E$, it holds also if adjustment is needed (2). The estimate of PDC can be adjusted for confounding factors by applying any one of the available statistical methods (11–13). We chose the most flexible and general adjustment method based on RR estimates from a logistic regression model that includes all the confounders assessed in the study. Confidence limits for the estimated PDC can be derived from the theory of PAF by using an implicit delta method (14, 15), the delta method ignoring some variances and covariances (16, 17), the asymptotic variance formula for the maximum likelihood estimator (18), or a variance formula originally proposed for case-control studies (2, 19, 20). In this paper, we utilized the last one.

To differentiate between relevant and nonrelevant sufficient causes and to simultaneously estimate the PDC for all relevant classes of sufficient causes, we propose the following multistep procedure.

1. Define indicator variables (dichotomous variables) for all $2^n − 1$ multifactorial events different from the zero event $E_0$.
2. Define all covariates that are planned for the analysis.
3. Starting with a model with all the covariates defined before, apply stepwise regression on the indicator variables only, to include all significant indicator variables in the final model.
4. Calculate the relative frequencies of the remaining significant multifactorial events within all cases.
5. Estimate the relative risks for the significant multifactorial events and insert these estimates into equation 4 to estimate the corresponding PDC.

This procedure includes statistical tests for the stratum-specific effects (step 3). A significant rejection of the hypothesis that the stratum-specific effect parameter is zero is equivalent to a significant rejection of the corresponding hypothesis $RR = 1$ and is a necessary condition for the significant rejection of the hypothesis $PDC = 0$. Thus, exclusion of nonsignificant indicator variables ignores only such classes of sufficient causes for which the confidence interval for PDC does contain zero. The subsequent statistical analysis was performed with SAS software (21).

All potential cases of incident myocardial infarction were identified by self-reports or death certificates. Cases were defined as participants who developed myocardial infarction according to International Classification of Diseases, Tenth Revision, codes I21.0–I21.9 (24). All incident cases of nonfatal or fatal myocardial infarction were verified by patients’ medical records or death certificates according to World Health Organization Monitoring of Trends and Determinants in Cardiovascular Disease (MONICA) criteria (25). Participants with missing follow-up status or prevalent myocardial infarction at baseline were excluded, leaving 26,972 participants (10,470 men and 16,502 women) for analyses. Among these, we identified 159 newly diagnosed cases of myocardial infarction (116 nonfatal and 43 fatal) that occurred between baseline and April 30, 2004. Mean follow-up was 4.6 years.

**Risk factors**

From the potentially modifiable risk factors associated with myocardial infarction (26–30), the following four were considered: smoking, hypertension, obesity, and lack of exercise. We defined individuals at risk from smoking if they were current smokers or had quit smoking within the past 5 years. Hypertension was defined as having a systolic blood pressure of 140 mmHg or more, having a diastolic blood pressure of 90 mmHg or more, taking antihypertensive medication, or self-reporting a hypertension diagnosis. According to previous studies (30), obesity was considered abdominal obesity and defined by a waist/hip ratio of 0.9 or more in men and 0.8 or more in women. Individuals were judged to have a lack of exercise if they engaged in less than 2 hours a week in sporting activities.

**Estimated risks and frequencies**

Indicator variables were defined for all 15 combinations of the four risk factors that include at least one present risk factor. The prevalence of each combination within cases and noncases is given in table 2. As expected, the percentage of individuals with three or four risk factors was markedly higher within cases than within noncases. Indeed, 78.1 percent (48.5 percent plus 29.6 percent) of cases had at least three risk factors, whereas the corresponding percentage within noncases was only 37.6 percent (30.0 percent plus 7.6 percent). However, the inverse relation was observed for individuals with none, one, or two risk factors.

Relative risks for the 15 combinations were estimated using a logistic regression model adjusted for age, sex, prevalence of diabetes, and history of dyslipidemia based on self-reports of a diagnosis or of taking cholesterol-lowering medication. The results are given in the last column of table 2. The relative risks for combinations with one or two present risk factors were not significantly different from one. In contrast, lack of exercise together with any two other risk factors was associated with a significantly elevated relative risk for myocardial infarction. The highest relative risk of 5.4 (95 percent confidence interval: 3.5, 8.5) was associated with the simultaneous presence of all four risk factors.

**APPLICATION TO REAL DATA**

**Study population**

The study population was the EPIC-Potsdam cohort, which is one of the two German cohorts contributing to the EPIC Study, a multicenter cohort study into diet and chronic diseases (22). A total of 27,548 subjects aged from 35 to 65 years were recruited from the general population between 1994 and 1998 (23). Baseline examinations included anthropometric measurements, blood sampling, the completion of a food frequency questionnaire, and a personal interview on lifestyle habits and medical history. Follow-up questionnaires are sent to the study participants every 2–3 years.
Following the procedure described before, PDC was estimated for all classes of sufficient causes that correspond to significant multifactorial events. The estimates and the 95 percent confidence intervals are given in figure 2. As expected from table 2, the class of sufficient causes containing all four risk factors as component causes was the most important one. The estimated proportion of disease due to this class was 24.1 percent. Somewhat surprisingly, the class of sufficient causes that requires the presence of hypertension, obesity, and lack of exercise ranked second with 14.6 percent, although the corresponding relative risk was only 2.0. This high percentage can be explained by the high prevalence of the corresponding multifactorial event. The PDCs of the other two classes with three known component causes were 9.6 percent and 3.5 percent. The PAF of the single risk factors can be estimated by summing up the PDC values of all sufficient causes that contain the risk factor as component cause. Thus, lack of exercise is the risk factor with the highest PAF of 51.8 percent, followed by obesity (48.3 percent), hypertension (42.2 percent), and smoking (37.2 percent).

**DISCUSSION**

The concept of sufficient cause and component causes introduced by Rothman is a tempting approach to model causality. In contrast to a completely deterministic model, stochastic elements for treating unmeasured or unpredictable...
component causes can be included in the pie charts by unlabeled slices. Thus, sufficient causes can be modeled without knowing the component causes completely. In this paper, we classified sufficient causes by multifactorial events that reflect specific combinations of known risk factors. Therefore, sufficient causes are not distinguishable if they belong to the same class. However, sufficient causes have different known causal components if they belong to different classes. Based on such a classification scheme, a statistical approach was described to estimate the proportion of disease that is due to a class of sufficient causes.

Although the PDC is defined as attributable fraction of a multifactorial event, it clearly differs from the joint PAF of multiple factors because the known component causes of the sufficient causes are fixed. The estimated PDC of all possible classes allows a partition of cases into strata characterized by combinations of known risk factors, whereas PAF estimates refer to overlapping events in general. Therefore, the practical value of the presented estimation procedure consists in deriving more detailed and precise results on the associations between the presence of risk factors and development of disease. To illustrate this issue, consider abdominal obesity as a risk factor for myocardial infarction in the EPIC-Potsdam Study. A recommendation to lose weight would be addressed to all obese individuals if only the PAF estimates are given, whereas PDC estimates refer to restricting the recommendation to obese individuals lacking exercise who are smokers or have hypertension. For both strategies, the proportion of disease that can be potentially prevented is 48.3 percent. However, the efficiency of the second strategy is much higher, because only 33.7 percent of subjects have the described risk profile compared with 53.1 percent of individuals who are obese (refer to table 2 and summing up respective prevalence within noncases).

Some limitations for the described estimation procedure should be noted. First, the simultaneous consideration of a lot of risk factors and the differentiation among many classes of sufficient causes require a high number of cases. If the sample size and especially the number of cases are too low, stepwise regression will exclude many indicator variables, and the estimated PDC of some important classes of sufficient causes may not be significantly different from zero. This limitation was apparent in the EPIC-Potsdam Study of incident myocardial infarction in which all combinations with one and two risk factors present and one combination with three risk factors present were excluded by stepwise regression. Large studies, such as the INTERHEART study (30) with about 15,000 cases, do not encounter this problem and should show a higher number of important classes of sufficient causes. Second, the estimation procedure for PDC supposes dichotomous risk factors. If originally continuous variables are dichotomized, the cutoffs chosen may be arbitrary, and change of cutoffs may lead to different results. Although it is possible to include different possible cutoffs by different types of sufficient causes, a conclusive strategy to handle this problem is not available.

Different concepts and interpretations of PAF have been proposed and discussed over the last 50 years (2, 31, 32). The concept used in our paper is that of “excess fraction,” defined as the fraction of cases that would not have occurred if the exposure (risk factor) had not occurred. The advantage of this concept is that an excess fraction can be validly estimated by the formulas of Miettinen (8) and Bruzzi et al. (33) by inserting an adjusted odds ratio from logistic regression, provided that there are no unobserved confounders and disease is rare (2). Because we have defined the PDC of a specific class $S_E$ as the PAF of the corresponding multifactorial event $E$, it can be interpreted as the excess
fraction of cases. Clearly, from the standpoint of biology, the fraction of cases that are etiologically attributable to \( S_E \) seems to be more relevant than the fraction of excess cases. Unfortunately, one cannot estimate the etiologic fraction without resorting to very strong biologic assumptions. The etiologic fraction is larger than the excess fraction (2, 31), because some cases caused by \( S_E \) still have become cases within the considered time period had \( E \) never occurred and, therefore, do not belong to the excess fraction. On the other hand, because the etiologic fraction attributable to \( S_E \) cannot be larger than the proportion of cases with the event \( E \), the current approach gives lower and upper bounds for the unknown etiologic fraction. In our example of myocardial infarction, for the class of sufficient causes with all four known risk factors as component causes, the etiologic fraction cannot be smaller than 24.1 percent (figure 2) and not larger than 29.6 percent (table 2).

Besides sufficient-component cause models, three other major types of causal models have been applied in health-sciences research: causal diagrams (34, 35), potential-outcome models (36, 37), and structural-equations models (38). Whereas the last two models provide a basis for quantitative analysis of effects, Rothman’s model of sufficient and component causes was considered only to illustrate the specific hypotheses about mechanisms of action (39). The extension of Rothman’s work presented may be helpful for quantitative analyses and for using the sufficient-component cause model beyond teaching examples. Furthermore, because a class of sufficient causes is characterized by a specific combination of present and absent risk factors, recommendations aimed to reduce the risk of disease can be specified to subgroups of individuals with the same risk profile. An application of the method to a study of myocardial infarction demonstrated the possible public health benefit of quantifying the importance of different sufficient causes.

**REFERENCES**

APPENDIX 1

Equation System for PDC in the Multicomponent Case

This appendix provides the equations for calculating the frequency distribution for sufficient causes from the single and joint PAF of component causes. Let \( X_1, \ldots, X_k \) be known risk factors of disease, let \( I \) be a subset of \( \{1, \ldots, k\} \), and let \( S_I \) be the class of sufficient causes that contain all \( X_i, i \in I \), as component causes but that do not contain \( X_i \) for \( i \in I \).

Obviously, the joint PAF of \( X_1, \ldots, X_k \) must be greater than or equal to the joint PAF of \( X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k \). The difference between these two PAFs is equal to the fraction of cases that is due to the class \( S_I \) of sufficient causes containing \( X_i \) as the only known component cause. Thus, the following equations hold:

\[
PDC(S_I) = \text{PAF}(X_1, \ldots, X_k) - \text{PAF}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_k).
\]  

(\text{A1})

Now consider any permutation \( X_{i(1)}, \ldots, X_{i(k)} \) of the risk factors and let \( X_{(m+1)}, \ldots, X_{(k)} \) be the last \( k - m \) factors in this arrangement. Then, the difference between the PAF of \( X_1, \ldots, X_k \) and the PAF of \( X_{(m+1)}, \ldots, X_{(k)} \) can be attributed to sufficient causes containing only risk factors as known component causes that have to belong to \( \{X_{(1)}, \ldots, X_{(m)}\} \). Therefore, we obtain the following relations:

\[
\sum_{I \subseteq \{1, \ldots, m\}} PDC(S_I) = \text{PAF}(X_1, \ldots, X_k) - \text{PAF}(X_{(m+1)}, \ldots, X_{(k)}),
\]  

(\text{A2})

where the sum on the left-hand side runs over all subsets of \( \{1, \ldots, m\} \). If we isolate the class of sufficient causes with the largest number of known component causes, equation \( \text{A2} \) can be rewritten as

\[
PDC(S_I) = \text{PAF}(X_1, \ldots, X_k) - \sum_{I \subseteq \{1, \ldots, m\}} \text{PAF}(X_{(m+1)}, \ldots, X_{(k)}),
\]  

(\text{A3})

where the sum on the right-hand side runs over all proper subsets of \( \{1, \ldots, m\} \). In the special case \( m = k \), the second term on the right-hand side of equation \( \text{A3} \) must be set equal to zero. The recursive equations \( \text{A3} \), together with the initial equations \( \text{A1} \), allow the calculation of almost all frequencies of sufficient causes. The remaining frequency refers to \( S_0 \) that can be calculated by the simple additional formula:

\[
PDC(S_0) = 1 - \text{PAF}(X_1, \ldots, X_k).
\]  

(\text{A4})

The equations \( \text{A1}, \text{A3}, \text{A4} \) together form a system of \( 2^k \) equations appropriate to determine the frequencies of the \( 2^k \) classes of sufficient causes.

APPENDIX 2

Solution of the Equation System in the Multicomponent Case

Again, let \( X_{(1)}, \ldots, X_{(k)} \) be any permutation of the risk factors \( X_1, \ldots, X_k \) and let \( S_I, I \subseteq \{1, \ldots, k\} \), be the class of sufficient causes that contains all \( X_i, i \in I \), as component causes but that does not contain \( X_i \) for \( i \notin I \). Then, in the generalization of equation 1, the joint PAF of the first \( m \) risk factors \( X_{(1)}, \ldots, X_{(m)}, m \leq k \), can be obtained by summing up the PDC values of all sufficient causes that contain at least one of the \( m \) risk factors as the component cause. Thus, we have the following equation:

\[
\text{PAF}(X_{(1)}, \ldots, X_{(m)}) = \sum_{I \subseteq \{1, \ldots, m\}} \text{PDC}(S_I).
\]  

(\text{A5})

On the other hand, applying the formula of Bruzzi et al. (33) that was formally proven by Benichou (12), the PAF of \( X_{(1)}, \ldots, X_{(m)} \) is equal to

\[
\text{PAF}(X_{(1)}, \ldots, X_{(m)}) = \sum_{I \subseteq \{1, \ldots, m\}} P(D|E_I) \frac{RR_I - 1}{RR'_I},
\]  

(\text{A6})

where \( E_I \) is the multifactorial event with \( X_i = 1 \) for \( i \in I \) and \( X_i = 0 \) for \( i \notin I \), and where \( RR_I \) is the relative risk of \( E_I \) versus \( E_0 \). Thus, PDC defined by equation 4 is a solution of the equation system A5. Since equation A5 is equivalent to equations A1 and A2, equation 4 also represents the solution for the latter ones.