Combining Cross-sectional Data on Prevalence with Risk Estimates from a Prediction Model

B. Engelhardt\textsuperscript{1,8}; J. König\textsuperscript{1}; M. Blettner\textsuperscript{1}; P. Wild\textsuperscript{2,3}; T. Münzel\textsuperscript{2}; K. Lackner\textsuperscript{4}; S. Blankenberg\textsuperscript{1,2}; N. Pfeiffer\textsuperscript{5}; M. Beutel\textsuperscript{6}; I. Zwiener\textsuperscript{1}

\textsuperscript{1}Department of Medical Biostatistics, Informatics and Epidemiology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany; \textsuperscript{2}Department of Medicine II, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany; \textsuperscript{3}Center for Thrombosis and Hemostasis, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany; \textsuperscript{4}Institute of Clinical Chemistry and Laboratory Medicine, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany; \textsuperscript{5}University Heart Center Hamburg, Department of General and Interventional Cardiology, Hamburg, Germany; \textsuperscript{6}Department of Ophthalmology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany; \textsuperscript{7}Department of Psychosomatic Medicine and Psychotherapy, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany; \textsuperscript{8}Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn-Aachen International Center for IT, Algorithmic Bioinformatics, Bonn, Germany

\section*{1. Introduction}

Death from cardiovascular diseases (CVD) is still the leading cause of death in Germany. About 271 people per ten thousand died of stroke and 179 of heart attack in 2008, both contributing to a combined percentage of about 37\% of all deaths \cite{1}. Well established risk factors for CVD include life style factors such as smoking, obesity, lack of physical activity, and diet factors \cite{2}. More recently, research has focused on genetic risk factors \cite{3}. Several models have been established to predict an individual’s risk of dying within the next five or ten years from CVD depending on the individual’s risk factors. The most important models for Germany are PROCAM \cite{4} and HeartScore Germany \cite{5}, with the latter considering the important risk factors sex, age, blood pressure, smoking status, and cholesterol level. As some of these risk factors depend on personal life style choices, changing one’s life style, e.g. by giving up smoking, may prevent death from CVD. The magnitude of the avoidable fraction depends on the strength of the association and the prevalence of the modifiable risk factors. While quantification of the fraction of avoidable death from CVD has been investigated by several authors in other countries, such as the USA \cite{6,7}, the Czech Republic \cite{8} or Japan \cite{9}, and in particular in...
the INTERHEART Study [10], little is known about the magnitude of avoidable deaths from CVD in Germany in contrast to e.g. the laryngeal risk on tumor subsites [11]. We used data from a large population-based cross-sectional study to estimate the prevalence of important risk factors in a representative population and used mortality predictions from the general population in Germany given by a risk prediction model to estimate the attributable proportion of fatal CVD risk according to different factors.

The purpose of our paper is to propose a method of combining data from two different sources, one for the prevalence of risk factors and one for the impact of the risk factors on the disease, to estimate the attributable risk. We propose a new presentation of the attributable risk and apply this method to estimate the fraction of avoidable deaths from CVD for Germany.

There are many situations in which only the prevalences of the risk factors and the absolute risks for the disease are available. Up-to-date prevalences may be estimated by a population-based cross-sectional study and absolute risks may be given by established literature of risk prediction modelling. Relating to these two types of data sources we want to calculate the attributable risks of the corresponding risk factors. This is not possible using existing formulas for the attributable risk, as they need a third data source: an up-to-date overall probability of the disease. We adapt existing methods that can calculate the attributable risk using only the two data sources and will deduce them in general. In addition we applied our new method to the case example of CVD in the Rhein-Main region in Germany.

The paper is structured as follows: In the methods section we review existing methods to calculate attributable risks and attributable fractions. In the results section we will then propose a new method to estimate the attributable fraction and its standard error by combining cross-sectional prevalence data with risk estimates from a prediction model, followed by a real data application in which we calculate the attributable fraction of dying from cardiovascular disease for a set of known risk factors. The paper concludes with a discussion of the proposed method.

2. Methods

In this section we will provide the notation and review existing formulas to calculate the attributable risks. In section 2.1 we give some notation and in section 2.2 we give a quick review of common formulae for attributable risks pointing out the importance of a more suitable formula for our data sources.

2.1 Notation

Let $D$ be the binary indicator for the disease status with $D = 1$ if the person will have the disease and $D = 0$ otherwise. In the same way, let $E$ be a categorical exposure status with values $i = 0, \ldots, N_E$ for the different levels of exposure, while $i = 0$ denotes the non-exposed individuals. Generally, we define the probability that an individual will have the disease by $P(D = 1)$

The probability that an individual with exposure level $i$ will be deceased is denoted by $P(D = 1 | E = i)$. Using this notation, the prevalence data of a cross-sectional study may provide estimates for $P(E = i)$, and a prediction model may provide estimates for $P(D = 1 | E = i)$ in each stratum.

2.2 The Attributable Risk

Common concepts of the attributable risk (AR) are based on an idea first mentioned by Levin in 1953 [12]. In brief, the attributable risk [13] quantifies the proportion of the disease which is attributable to a considered exposition. In other words, the attributable risk is the proportion by which the risk of disease (or mortality) is reduced if the risk in one or more exposure categories is replaced by the risk observed in a reference category. In the real data application we consider the 10-year cardiovascular mortality risk $P(D = 1 | E = i)$ given by the HeartScore Germany and the prevalences $P(E = i)$ of the risk factors given by the cross-sectional Gutenberg Health Study [14]. Before introducing the general multistate case we start with two exposure groups and consider the exposure variable $E$ as binary with levels 0 and 1 (non-exposed/exposed). We also have a binary outcome variable $D$ with levels 0 and 1 (non diseased/diseased), so that the attributable risk is defined as

$$AR = \frac{P(D = 1) - P(D = 1 | E = 0)}{P(D = 1)}$$

In case of two different risk factors, the net effect of one risk factor, $E$, while adjusting for another one, $C$, with levels $k = 1, \ldots, N_C$, is of interest in our situation. Therefore the AR of $D$ due to $E$ adjusted for $C$ is defined as

$$AR_{adj}(C)(E) = \frac{P(D = 1) - \Delta}{P(D = 1)},$$

$$\Delta = \sum_{k=1}^{N_C} P(D = 1 | E = 0 \land C = k) P(C = k).$$

The adjusted attributable risks for several risk factors do not add up to the attributable risk associated with simultaneously eliminating all exposure variables to their reference. Therefore, attempts have been made to split the overall attributable risk of a set of risk factors into a sum of risk-factor-specific attributable risk constructs. One of these approaches is the introduction of the partial attributable risk (PAR), which is based on a concept from the cooperative game theory developed by Shapley [16]. It was formalized by Eide and Gefeller [17] and is nicely described by Rabe and Gefeller [18]. The partial attributable risk is the average of all marginal contributions of the risk factor in all possible permutations. The basic idea is to divide the combined population impact of multiple risk factors into components that can be attributed to the respective individual exposures while taking into account the potential interrelations between the factors [19].

In order to find the for the combination of risk factors (e.g. hypertension and cholesterol level) adjusted for several strata (e.g. smoking), the formulae above can be extended to several risk factors, exposure variables, or several confounders, each of which can possibly have more than two levels [19].

Similar approaches have been made by other authors, for example in estimating global burden of disease [20, 21].
contrast to these approaches, we estimate the attributable and partial attributable risks, respectively, on the basis of current local cross-sectional prevalence and combined them with well established risk estimates from a global prediction model.

3. Results

3.1 Attributable Risks: New Representation in Terms of Exposure Prevalences and Stratified Risks

These common formulae are well known [15, 22], but often data doesn’t fit these formulae because the overall prevalence of the diseased individuals \( P(D = 1) \) may not be given or is not up-to-date. Therefore we adapt the formula in such a way that only the information of the risks and up-to-date prevalences of the risk groups are needed.

The adapted formula can be applied separately within the strata defined by gender and age (or further covariates), or it can be applied to the overall population. We denote the set of observable combinations of exposure levels by \( i = 0, \ldots, N_E \), with the unexposed individuals are denoted by \( i = 0 \) and the set of combinations of confounding variables by \( k = 1, \ldots, N_C \).

In terms of prevalences of the exposure \( p_{ik} = P(E = i, C = k) \) and the confounders and the stratum-specific risks (a stratum is given by the exposure and confounder combination) \( r_{ik} = P(D = 1|E = i, C = k) \), we write the adjusted AR as

\[
AR^{EC} = \frac{\sum_{i=0}^{N_E} \sum_{k=1}^{N_C} (r_{ik} - p_{ik}) p_{ik}}{\sum_{i=0}^{N_E} \sum_{k=1}^{N_C} p_{ik}}. \tag{2}
\]

That means we compare the weighted differences in stratum-specific risks between each exposure level and their reference with the weighted overall risk (for the derivation of \ref{eq:2}, see the \ref{appx}). The summands in the numerator of \ref{eq:2} can be interpreted as the risk difference between an individual with exposure level \( i \) and an individual in the reference within each confounder stratum. They are multiplied by the frequency of the exposure level \( i \) and stratum \( k \). Thus the denominator gives the overall risk in all exposure levels weighted by their prevalence.

Note that by relating all risk strata to one common reference, the adjusted AR can also be expressed in terms of exposure-confounder prevalences and relative risks

\[
RR_{ik} = \frac{r_{ik}}{r_{0k}}.
\]

3.2 Standard Errors for the Proposed Representation of the Attributable Risk

Standard errors for attributable risk estimates were obtained using the delta method. We denote the numerator and denominator of \ref{eq:2} as \( X \) and \( Y \), their expectations as \( \mu_X \) and \( \mu_Y \), and the variances as \( \sigma_X^2 \) and \( \sigma_Y^2 \). We use the delta method approximation, following Eide and Gefeller [17],

\[
V(AR^{EC}) = V\left( \frac{\sum_{i=0}^{N_E} \sum_{k=1}^{N_C} (r_{ik} - p_{ik}) p_{ik}}{\sum_{i=0}^{N_E} \sum_{k=1}^{N_C} p_{ik}} \right) = \frac{1}{\mu_Y^2} V(X) + \frac{\mu_X^2}{\mu_Y^2} V(Y) - 2 \frac{\mu_X}{\mu_Y^2} \text{COV}(X, Y). \tag{3}
\]

Both the numerator and denominator are linear combinations of multinomial distributed random variables whereby the risks are considered as known. To derive formula (3) which is equivalent to their expected value we used the Taylor series expansion [23]. \( V(AR^{EC}) \) can also be estimated by

\[
V\left( \frac{\mu_X^2}{\mu_Y^2} \right) \approx \frac{1}{\mu_X^2} V(\hat{X}) + \frac{\mu_Y^2}{\mu_X^2} V(\hat{Y}) - 2 \frac{\mu_Y}{\mu_X^2} \text{COV}(\hat{X}, \hat{Y}).
\]
which is exactly the point of Formula 3. We write

\[ X = \sum_{i=0}^{N_x} \sum_{k=1}^{N_y} (r_{ik} - r_{0k}) \bar{p}_{ik} = \sum_{i=0}^{N_x} \sum_{k=1}^{N_y} (r_{ik} - r_{0k}) \frac{w_{ik}}{N} m_{ik} = a^T m \]

and

\[ Y = \sum_{i=0}^{N_x} \sum_{k=1}^{N_y} r_{ik} \bar{p}_{ik} = b^T m \]

Here, \( m \) denotes the set of observed absolute frequencies rearranged as a vector, \( w_{ik} \) denotes the weights attached to each group and \( a \) and \( b \) denote the corresponding sets of linear coefficients.

Therefore \( V(X) = a^T \Sigma_m a, V(Y) = b^T \Sigma_m b \) and \( \text{cov}(X, Y) = a^T \Sigma_m b \).

The covariance matrix \( \Sigma_m \) is the covariance matrix of multinomial distribution.

\[ \Sigma_m = \left[ \text{diag} \left( E(m) \right) - \frac{1}{M} E(m) E(m)^T \right] \]

if only one age stratum is involved or contains diagonal blocks of multinomial variances for each age-gender stratum. The expectations \( E(m) \) are replaced by observed frequencies \( m_{ik} \) arranged as a vector to obtain approximate standard errors. For our analyses we also estimated the standard error without any weights to check the effect of weighting. The above formulae then apply with vectors and modified accordingly.

4. Real Data Application

The underlying issue is to calculate an up-to-date attributable risk of cardiovascular mortality due to different risk factors. Therefore, we combine an established risk prediction model with up-to-date prevalences for the risk factors. For this, we apply the developed representation of the attributable risk in section 3.

We use data from two different sources to obtain estimates of attributable risks. Stratum-specific estimates of cardiovascular mortality risks are obtained from the German adaption of the Euro Heart Score by the European Society of Cardiology [24], and the distribution of cardiovascular risk factors was estimated using data from the cross-sectional baseline examination of the Gutenberg Health Study [14], a population-based cohort study including 15,000 participants in the Rhein-Main region. In section 4.1 and 4.2 we will discuss these two different data sources that were eligible for our analysis.
4.1 HeartScore Germany for Risk Estimates of Prediction Model

The Euro Heart Score combines a pool of datasets from twelve European cohort studies. It provides risk predictions for individual patients based on the most prominent risk factors: gender, age, systolic blood pressure, cholesterol level, and smoking. The hazard ratios of these risk factors were estimated based on data of twelve European countries. To calculate the ten-year risk of fatal cardiovascular disease for residents in Germany the baseline hazard was recalculated with specific mortality rates from Germany in 1999 and results from the German National Health Interview and Examination Survey of 1998 [25, 26]. The result is the HeartScore Germany [5, 27] (copyright by the European Society of Cardiology (ESC), www.heartscore.org), see Figure 1, which predicts the individual ten-year risk of fatal cardiovascular disease for residents in Germany for different combinations given the five most important risk factors age (5 categories: < 45 years = reference, 45–52.5, 52.5–57.5, 57.5–62.5, 62.5–67), gender (male, female), systolic blood pressure (4 categories: ≤ 130 mmHg = reference, 130–150, 150–170, > 170), the ratio of the overall cholesterol level to HDL (5 categories: ≤ 3.5 = reference, 3.5–4.5, 4.5–5.5, 5.5–6.5, > 6.5), and smoking (2 categories: non-current smoker = reference vs. current smoker). This score is especially designed for general practitioners to enable them to easily calculate the CVD risk of their patients using a table. We use this score for estimating the risk of fatal CVD for the three modifiable risk factors blood pressure, cholesterol level and smoking.

Figure 1 provides the Heart Score Germany table with ten-year risks of dying of cardiovascular disease for males and females in different age groups and risk groups. We only use the risk of the exposed vs. the non-exposed rather than the overall 10-year mortality, as the baseline hazards used in this model are not up-to-date.

### Table 1

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>14.9</td>
<td>16.4</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>20.0</td>
<td>30.1</td>
</tr>
<tr>
<td>Smoking</td>
<td>12.8</td>
<td>15.4</td>
</tr>
<tr>
<td>Smoking and blood pressure</td>
<td>26.3</td>
<td>29.8</td>
</tr>
<tr>
<td>Smoking and cholesterol</td>
<td>29.9</td>
<td>40.6</td>
</tr>
<tr>
<td>Blood pressure and cholesterol</td>
<td>31.9</td>
<td>41.7</td>
</tr>
<tr>
<td>Overall</td>
<td>40.7</td>
<td>50.8</td>
</tr>
</tbody>
</table>

#### 4.2 Gutenberg Health Study for Cross-sectional Prevalence of Risk Factors

Data from the Gutenberg Health Study (GHS) is used to assess the up-to-date prevalence of risk factors needed for the Euro Heart Score respectively HeartScore Germany. The GHS is a population-based cohort study in the Rhein-Main area which began recruitment in 2007. In this study, various information was collected from each participant. The GHS recruited people of age 35 to 64. For every decade of age (35–44, 45–54, 55–64) the same number of persons have been recruited into the study. At the time of our analysis, information about the first 10,000 participants was available. We excluded 66 participants because of missing values on one or more of the three risk factors. A further 1,752 participants aged over 67 years were excluded because the HeartScore Germany is limited to this age range. In the end we could analyze data from 8,182 participants. We categorized the participants into groups in the same way as the HeartScore Germany and estimated their prevalences in per cent (Figure 2). The GHS sample is an age-stratified random sample population taken from the city of Mainz and the surrounding district of Mainz-Bingen. For all decades of age, the same number of participants was included. The age structure in the underlying population of Mainz and Mainz-Bingen has more people in the younger age decades than in the older ones. To accommodate for this, the age- and gender-specific prevalences were aggregated by weighting each person according to the age structure of the underlying population of Rhineland-Palatinate in 2007. The data was provided by the national statistical office of Rhineland-Palatinate.

The data come from the national statistical office of Rhineland-Palatinate.

#### 4.3 Results

In this section we present the results of our analyses regarding stability and accuracy of the estimates based on the HeartScore Germany and the GHS.

The overall attributable risks (Table 1) and deduced partial attributable risks (Table 2) based on the HeartScore Germany and the GHS can be estimated with Formula 2. This method is flexible so that the adaption to other sources and exposure set-ups is simple. It is also possible to fit our method to odds ratios. Note that in contrast to the common methods [15, 19, 22], we don’t need the overall 10-year risk of fatal CVD.

For illustration, we will estimate the attributable risk of an increased systolic blood pressure for males in the age range of 62.5 and 67 (Figure 3). Smoking and cholesterol are the set of confounders. They split each exposure level into ten strata (2 levels for smoking, 5 levels for cholesterol). We have four different levels for the exposure (one reference group = 0 = 120 mmHg, three exposed groups: 1 = 140 mmHg, 2 = 160 mmHg, and 3 = 180 mmHg). The risks of dying from cardiovascular disease within the next ten years for the reference group (those with a

© Schattauer 2014

Methods Inf Med 5/2014
good blood pressure) are given in the bottom line of ▶ Figure 3, namely: 4, 5, 6, 7 and 8 percent for the non-smokers and 7, 9, 11, 13 and 14 percent for the smokers. This means that a non-smoker with the lowest cholesterol level has a 10-year-risk of fatal CVD of 4 percent and for a smoker with an equal cholesterol level has a 10-year-risk of fatal CVD of 7 percent. To calculate the attributable risk you compare the ten-year risk of a fatal CVD with the risk in the reference group within the same stratum for the confounders and multiply the value with the corresponding prevalence from the GHS (▶ Figure 2). More precisely, the risk of dying for a smoker having cholesterol level is 23 percent if his blood pressure is around 180, 17 percent if his blood pressure is around 160 and 12 percent if is blood pressure is around 140, respectively.

A man with blood pressure around 160 could reduce his blood pressure to around 120, his risk of dying would be reduced from 17 percent to 9 percent. So for this individual the risk attributable to his systolic blood pressure is (17 – 9)/17 = 47%.

In addition, the GHS prevalences must be age-weighted to apply ▶ Formula 2. The weights attached to each participant of the GHS when aggregating over different age groups, were \( w_s = g_s \frac{N}{n_s} \) with \( n_s \) denoting the sample size of age stratum \( s \) within the GHS which was about 250 samples per stratum, \( N \) is the size of the analyzed sample, 8,182 individuals, and is the proportion of that age group within the population of Rhineland-Palatinate in 2007 restricted to the age range of 35 – 67 years. Weights were determined separately for each gender.

Because weights are constant within age strata, the average exposure prevalences or mortality risks can simply be written as

\[
f = \sum_{s=1}^{S} f_s g_s \text{ where } f \text{ and } f_s \text{ denote averaged and stratum-specific proportions, respectively.}
\]

In addition to the overall AR, we calculated the for different age classes individually (▶ Table 3) and deduced the PAR (▶ Table 4), too. From the medical point of view, the estimates for the different age groups seem plausible. ▶ Table 3 shows that the attributable risk due to smoking decreases with age. This is due to the decrease of the prevalences of smoking for increasing age. We can see the opposite correlation between age and blood pressure. The attributable risk due to increased systolic blood pressure increases with higher age, as higher age goes along with a higher proportion of hypertension.

To validate the formula for the approximate standard error in the real data application we drew 1000 bootstrap samples (independently for each age group and each gender) from the GHS sample and record the empirical standard deviation of estimates based on the bootstrap samples.

Furthermore, we estimated standard errors using our approximation in ▶ Formula 3 and compared them to the estimate obtained by the bootstrap samples (▶ Table 5). Comparing the standard errors, the bootstrapped standard error is smaller than the calculated standard error in most cases. This is a characteristic of the bootstrap method. In conclusion we might assume that our approximate formula is a conservative estimate. The largest difference of 0.6 is seen for the overall AR for the 58 – 62 year olds. The SE is very small in all cases. In this context we also analyzed the influence of age weighting of the standard error and found that the weights can be neglected because of their maximal deviation of 0.25 percentage points (data not shown). Owing to the small SEs, the estimates for the AR are very exact, which can be explained by the large number of participants used to calculate the prevalences.
(N ≈ 8,500 in the GHS) and the assumption of fixed mortality rates. If we do not assume fixed mortality risks, formula 3 must be modified by risk distributions. This could be done by taking into account the SEs of the HeartScore Germany estimates for the mortality risks $r_g$ into account.

5. Discussion

The analyses show a high potential for prevention of CDV in Germany, especially related to smoking, cholesterol level and blood pressure. The prevention scheme should be age-dependent, as we have seen different impacts in the different age groups. From the statistical point of view we obtained a realistic estimate of the risk-reduction potential within ten years and are able to calculate approximate standard errors.

Due to the underlying different data sources, the method has its limitations. On the one hand, we used the HeartScore Germany, an adaptation of the Euro Heart Score for the individual risks, and on the other hand the GHS for the prevalence of each risk group. The Euro Heart Score as a Europe-wide study is well known and accepted, and the GHS is a current local study which is still under way and whose participants are healthier than the German average. The reason for this is the study structure; the incentive for the participants to join the study is a free health check. Despite this, the GHS is very attractive because it is a representative, up-to-date study, and the GHS is a current local study which is still under way and whose participants are healthier than the German average. The reason for this is the study structure; the incentive for the participants to join the study is a free health check. Despite this, the GHS is very attractive because it is a representative, up-to-date study, and it is one of the first of its kind.

The real potential of risk reduction might be influenced by the fact that not all people can be treated in such a way that all risk factors are reduced to a normal level. Reducing the blood pressure of a patient requires balancing the benefits and potential side effects of the therapy, and therefore, in practice, reducing the blood pressure to the normal level of 130 mmHg might not always be possible. However, in this paper we assume full reduction of the risk factor. Thus the should be seen as an upper bound on what is practically achievable.

In contrast to what is generally assumed, smoking is no longer the leading risk factor. Thus the should be seen as an upper bound on what is practically achievable.

The ongoing prevention of hypertension shows positive effects in our study, resulting in cholesterol as the leading factor of the preventable proportion for cardiovascular mortality. This is in agreement with current studies, which show an increase of obesity. Obesity is one of the main factors for high cholesterol levels [28].

The SEs estimated with our methods are small, and therefore the estimates of the attributable risks are very exact, but the analysis is limited by the underlying data sources. We assume fixed exposure risks and unbiased prevalence estimation. Nevertheless, our formula and its SEs are easy to implement and are flexible. It is very simple to replace the data sources in order to modify the analysis. We obtain stable estimators for global attributable risks of the most important risk groups, which can be expressed more simply as.

For further research, it will be interesting to compare the mortality rates that can

<table>
<thead>
<tr>
<th>Age group</th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35 to 44 years</td>
<td>45 to 52 years</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>1.9</td>
<td>5.8</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>12.5</td>
<td>14.6</td>
</tr>
<tr>
<td>Smoking</td>
<td>17</td>
<td>16.7</td>
</tr>
<tr>
<td>Overall</td>
<td>31.5</td>
<td>37.2</td>
</tr>
<tr>
<td></td>
<td>35 to 44 years</td>
<td>45 to 52 years</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>5.2</td>
<td>7.4</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>26.8</td>
<td>27.4</td>
</tr>
<tr>
<td>Smoking</td>
<td>17.5</td>
<td>16.5</td>
</tr>
<tr>
<td>Overall</td>
<td>49.6</td>
<td>51.3</td>
</tr>
</tbody>
</table>

Table 5 Calculated and simulated standard errors of the adjusted attributable risks for the main risk factors, age adjusted, women only.

<table>
<thead>
<tr>
<th>Age group</th>
<th>35 to 44 years</th>
<th>45 to 52 years</th>
<th>53 to 57 years</th>
<th>58 to 62 years</th>
<th>63 to 67 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood pressure</td>
<td>0.34</td>
<td>0.38</td>
<td>0.59</td>
<td>0.66</td>
<td>0.94</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>0.54</td>
<td>0.56</td>
<td>0.61</td>
<td>0.53</td>
<td>0.83</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.77</td>
<td>0.82</td>
<td>0.83</td>
<td>0.9</td>
<td>1.12</td>
</tr>
<tr>
<td>Overall</td>
<td>0.85</td>
<td>0.97</td>
<td>0.9</td>
<td>0.85</td>
<td>1.15</td>
</tr>
</tbody>
</table>
be estimated by our formula with the mortality rates of the follow-up data in the GHS in order to prove the accuracy of our calculation. However, the GHS’s follow-up data is not available yet.

The results above demonstrate that our representations for attributable and partial attributable risks are useful to easily compare different data sets.

Acknowledgments

The Gutenberg Heart Study is funded through the government of Rhineland-Palatinate ("Stiftung Rheinland Pfalz für Innovation"); contract number AZ 961-386261/733) and the research programs “Wissenschaft Zukunft” and “Schwerpunkt Vaskuläre Prävention” of the Johannes Gutenberg-University of Mainz and its contract with Boehringer-Ingelheim and Philips Medical Systems, including an unrestricted grant for the Gutenberg Heart Study.

References

Appendix

To illustrate the equivalence of Equations 1 and 2 following Eide and Gefeller [17], we first consider the denominators. We will start with our formula:

\[ \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} r_{ik} p_{ik} = \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} P(D = 1 | E = i \land C = k) P(E = i \land C = k) \]

Because \( r_{ik} \) is the disease risk under condition of exposure level \( i \) and strata \( k \). Analogously \( p_{ik} \) is the conditional probability, given strata \( k \). This product is equivalent to the probability of disease \( P(D = 1) \).

Now consider the numerator of our formula:

\[ \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} \left( r_{ik} - r_{0k} \right) p_{ik} = \]

\[ = \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} P(E = i \land C = k) P(D = 1 | E = i \land C = k) \]

\[ - \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} P(E = i \land C = k) P(D = 1 | E = 0 \land C = k) \]

The explanation is analogous to the denominator.

It is easy to see that the first subtrahend is equivalent to the probability of disease \( P(D = 1) \), and we get:

\[ P(D = 1) - \sum_{i=0}^{N_e} \sum_{k=1}^{N_c} P(E = i \land C = k) P(D = 1 | E = 0 \land C = k) \]

We also know that the equation \( \sum_{i=0}^{N_e} p_{ik} = P(C = k) \) must be true for all \( k \in (1, \ldots, N_c) \).

Now we get:

\[ P(D = 1) - \sum_{k=1}^{N_c} P(C = k) P(D = 1 | E = 0 \land C = k) \]

As consequence, we arrive at the equation:

\[ \frac{\sum_{i=0}^{N_e} \sum_{k=1}^{N_c} \left( r_{ik} - r_{0k} \right) p_{ik}}{\sum_{i=0}^{N_e} \sum_{k=1}^{N_c} r_{ik} p_{ik}} = \frac{P(D = 1) - \sum_{k=1}^{N_c} P(C = k) P(D = 1 | E = 0 \land C = k)}{P(D = 1)} \]