Journal of Statistical Software

July 2011, Volume 43, Issue 7. http://www.jstatsoft.org/

SAS Macros for Calculation of Population Attributable Fraction in a Cohort Study Design

Maarit A. Laaksonen
National Institute for Health and Welfare

Esa Virtala
National Institute for Health and Welfare

Paul Knekt
National Institute for Health and Welfare

Hannu Oja
University of Tampere

Tommi Härkänen
National Institute for Health and Welfare

Abstract

The population attributable fraction (PAF) is a useful measure for quantifying the impact of exposure to certain risk factors on a particular outcome at the population level. Recently, new model-based methods for the estimation of PAF and its confidence interval for different types of outcomes in a cohort study design have been proposed. In this paper, we introduce SAS macros implementing these methods and illustrate their application with a data example on the impact of different risk factors on type 2 diabetes incidence.

Keywords: PAF, cohort study, risk factor, mortality, disease incidence, censoring, effect modification, piecewise constant hazards model.

1. Introduction

Quantification of the impact of exposure to modifiable risk factors on a particular outcome, such as death or disease occurrence, at the population level is a fundamental public health issue. Population attributable fraction (PAF) assesses the proportion of the outcome attributable to exposure to such modifiable risk factors in a given population by estimating the proportion of outcome that would not have occurred if all individuals had belonged to the low-risk reference category of those factors (for example, what proportion of deaths would not have occurred if nobody had started smoking). Both relative risk (RR) and the prevalence

of risk factors are taken into account in the calculation of PAF. Relative risk measures the strength of association between the risk factors and the outcome using a ratio of the probability of the outcome occurring in the exposed group to the probability of the non-exposed group outcome. The prevalence of the risk factors is the proportion of individuals with the risk factor in a population.

So far, the estimates of PAF and its confidence interval have been mainly obtained from cross-sectional and case-control studies (Benichou 1991; Coughlin, Benichou, and Weed 1994; Benichou 2001). Programs for the estimation of these static PAF estimates in different programming languages (e.g., SAS, Stata, and R/S-PLUS) are available (Mezzetti, Ferraroni, Decarli, La Vecchia, and Benichou 1996; Brady 1998; Kahn, O'Fallon, and Sicks 1998; Grömping and Weimann 2004; Eide 2006; Lehnert-Batar 2006; Rückinger, von Kries, and Toschke 2009; Rämsch, Pfahlberg, and Gefeller 2009). Three different approaches for estimating PAF and its confidence interval from cohort studies have been proposed. In the first approach, only the occurrence of the event of interest is observed whereas the timing of the event is ignored, i.e., the event outcomes are treated as binary (Benichou 2001). In this case, the only difference to the cross-sectional study is that the outcome is not observed simultaneously with the risk factors but after a fixed follow-up, and thus the same methods (i.e., the logistic model) and programs as for the estimation of PAF and its confidence interval in cross-sectional studies can be applied. This approach, however, may lose information and produces reliable estimates only in case of no censoring during follow-up. In the second approach, the time of the event or censoring time is observed, i.e., censored time-to-event data is used, but the effect of the hypothetical risk factor modification to the low-risk level is estimated at the instantaneous time point t (Chen, Hu, and Wang 2006; Samuelsen and Eide 2008). The estimate obtained thus describes the approximate proportion of events that could be prevented by the risk factor modification in question in a small time interval $[t, t + \Delta t]$, where $\Delta t \to 0$. As far as the authors of this paper know, one publicly available SAS macro for the estimation of this "instantaneous PAF" has been provided (Spiegelman, Hertzmark, and Wand 2007). Usually, however, it is more useful to demonstrate the effect of the risk factor modification during a longer time interval $(0, t]$ as is done in the third, most recently suggested, approach (Chen *et al.* 2006; Samuelsen and Eide 2008; Cox, Chu, and Muñoz 2009; Laaksonen, Knekt, Härkänen, Virtala, and Oja 2010b; Laaksonen, Härkänen, Knekt, Virtala, and Oja 2010a). For example, in case of an outcome, such as death, that is inevitable in time and can only be delayed, it would be useful to calculate PAF estimates for time intervals of different length in order to demonstrate the effect of the risk factor modification in the long run (Laaksonen *et al.* 2010b). When the outcome is disease occurrence, potential censoring due to death needs to be considered and the impact of censoring on the results can be observed in a longer follow-up (Laaksonen *et al.* 2010a). Furthermore, due to the inevitability of death the PAF in both cases will eventually approach zero as time goes to infinity and thus become meaningless, which further signifies the importance of specifying a certain time interval. Despite its importance, the last approach has, apparently due to difficulty in computation, received little theoretical attention. To the best of authors' knowledge, there are no publicly available programs which estimate the PAF for a time interval of $(0, t]$.

In this paper, we will present macros for the estimation of PAF for a time interval $(0, t]$ and its confidence interval in a cohort study design both for total mortality and disease incidence, adjusted for potential confounding factors and accounting for potential effect modifying factors. Proportional hazards models with a piecewise constant baseline hazard functions for

death and disease occurrence are assumed. In Section 2, we give a definition of PAF for a time interval $(0, t]$. In Section 3, we propose an estimate of PAF for total mortality and disease incidence and derive its asymptotic variance, both in the total population and in subpopulations. In Section 4, we explain the SAS macros for the estimation of PAF for total mortality and disease incidence in the presence of potential effect modification. In Section 5, we illustrate the application of these macros. Finally in Section 6, we discuss the strengths and weaknesses related to our program and its application.

2. Concept of PAF in cohort study design

Consider the occurrence of an outcome A in a population of n individuals with risk factor values $X_i = (x_{i1}, \ldots, x_{im})^\top$, $i = 1, \ldots, n$. In a cohort study design, PAF is defined to be the proportion of the outcome occurrence that could be avoided during a certain follow-up time (T), which is determined as the time from baseline $(t = 0)$ to the time of the event of interest or censoring (whichever comes first), if it was possible to change some risk factor values to their chosen target values, $X_i = (x_{i1}, \ldots, x_{im})^\top \to X_i^* = (x_{i1}^*, \ldots, x_{im}^*)^\top$ (Laaksonen *et al.* 2010b). In this notation, X_i is the vector of all risk factors of the ith individual considered relevant (modifiable, non-modifiable, and confounding factors), and thus only the modifiable risk factors whose effect we wish to measure will have a different value in X_i^* while the rest of the factors retain their values. The PAF is then

$$\mathrm{PAF}(A) = \frac{\sum_{i=1}^n \mathsf{P}\{A_i|X_i\} - \sum_{i=1}^n \mathsf{P}\{A_i|X_i^*\}}{\sum_{i=1}^n \mathsf{P}\{A_i|X_i\}} = 1 - \frac{\sum_{i=1}^n \mathsf{P}\{A_i|X_i^*\}}{\sum_{i=1}^n \mathsf{P}\{A_i|X_i\}},$$

where $\mathsf{P}\{A_i|X_i\}$ is the probability of the occurrence of outcome (A) for the ith individual with the risk factors X_i.

In this study, we are interested in calculating PAF both for occurrence of a terminal outcome, such as death, and for order of occurrence of two terminal outcomes, such as occurrence of a chronic disease before death, during a time interval $(0, t]$. If the outcome of interest is death, PAF is the proportion of mortality that could hypothetically be avoided during a time interval $(0, t]$ if its risk factors were modified (Laaksonen *et al.* 2010b). Let T^M denote the time of death. Then the proportion of excess mortality up to time t due to certain modifiable risk factors in X_i is given by

$$\mathrm{PAF}(T^M \le t) = 1 - \frac{\sum_{i=1}^n \mathsf{P}\{T_i^M \le t|X_i^*\}}{\sum_{i=1}^n \mathsf{P}\{T_i^M \le t|X_i\}}, \tag{1}$$

where $\mathsf{P}\{T_i^M \le t|X_i\}$ is the probability of death up to time t, given the risk factor values X_i. If, however, the outcome of interest is incidence of disease, PAF is the proportion of disease cases that could hypothetically be avoided during a time interval $(0, t]$ if its risk factors were modified. In this case, mortality before contracting the disease of interest causes selection in the population during follow-up (Laaksonen *et al.* 2010a). If the risk factors that are related to the incidence of the disease of interest are also related to mortality, the modification of these risk factors is likely to affect both the risk of the disease and the risk of death. Therefore, censoring due to death needs to be taken into account in the definition of PAF for disease incidence. The importance of considering censoring due to death in the estimation of PAF for disease incidence has been demonstrated elsewhere (Laaksonen *et al.* 2010a). Let T^D denote

the time of the occurrence of the disease. Then the proportion of excess disease incidence up to time t due to certain modifiable risk factors in X_i is given by

$$\mathrm{PAF}(T^D \leq \min(T^M, t)) = 1 - \frac{\sum_{i=1}^{n} \mathsf{P}\{T_i^D \leq \min(T_i^M, t)|X_i^*\}}{\sum_{i=1}^{n} \mathsf{P}\{T_i^D \leq \min(T_i^M, t)|X_i\}}, \tag{2}$$

where $\mathsf{P}\{T_i^D \leq \min(T^M, t)|X_i\}$ is the probability of the disease incidence up to time t, given the risk factor values X_i.

We thus need two different definitions of PAF, $\mathrm{PAF}(T^M \leq t)$ and $\mathrm{PAF}(T^D \leq \min(T^M, t))$, depending on the outcome of interest. Furthermore, to analyze the impact of some potential effect modifying factor on the relationship between the risk factor and the outcome of interest at the population level, we can calculate PAFs in the subpopulations defined by different categories of the effect modifying factor. By calculating the differences between these PAF estimates and their confidence intervals we can further assess the statistical significance of the effect modification.

3. Estimation of PAF in cohort study design

3.1. General model assumptions

The following assumptions in the calculation of PAF for total mortality or for disease incidence in the cohort study design are made in this study. Proportional hazards models are applied. The hazard of death is $h^M(t)$ and the hazard of disease incidence $h^D(t)$. The corresponding cumulative hazard functions are then $H^M(t) = \int_0^t h^M(u)\,du$ and $H^D(t) = \int_0^t h^D(u)\,du$, and the conditional survival functions $S^M(t) = \exp\{-H^M(t)\}$ and $S^D(t) = \exp\{-H^D(t)\}$. For each individual, the hazard functions are assumed to depend on all relevant risk factors (X) for both mortality and disease incidence: $h^M(t) := h^M(t; X)$ and $h^D(t) := h^D(t; X)$. The time of death T^M and the time of the occurrence of disease T^D are assumed to be conditionally independent given X. The hazard function for disease-free survival at time t, $\min(T^M, T^D) > t$, is assumed to be $h^M(t; X) + h^D(t; X)$. Then, the probability that the first event is disease is $\mathsf{P}\{\min(T^M, T^D) = T^D | \min(T^M, T^D) > t\} = h^D(t; X)/\left(h^M(t; X) + h^D(t; X)\right)$.

There may still be right-censoring by T^C which is assumed to be conditionally independent of T^M and T^D given X. If the outcome of interest is death, we then observe $T^C = \min\{T^C, T^M\}$ in case of right-censoring or $T^M = \min\{T^C, T^M\}$ in case of death. If the outcome of interest is incidence of disease, we observe $T^C = \min\{T^C, T^M, T^D\}$, $T^M = \min\{T^C, T^M, T^D\}$, $T^D < T^C = \min\{T^C, T^M\}$, or $T^D < T^M = \min\{T^C, T^M\}$. It is important to note that the definition of PAF does not depend on T^C.

3.2. Piecewise constant hazards model

In the calculation of PAF, the times T^D and T^M are assumed to follow a proportional hazards model with piecewise constant baseline hazard functions, given X. A parametric piecewise constant hazards model is chosen due to its flexibility in accommodating to the shape of the underlying survival curve and ease of computation (Laaksonen *et al.* 2010a,b). In a parametric piecewise constant hazards model, the follow-up time is partitioned into J intervals

$(0 = a_0, a_1]$, ..., $(a_{j-1}, a_j]$, ..., $(a_{J-1}, a_J]$, where $a_{j-1} < a_j$ for all j and the hazard for the ith individual

$$h(t; X_i) = e^{X_i^\top \beta} \prod_{j=1}^{J} \lambda_{0j}^{1\{a_{j-1} < t \leq a_j\}}$$

is allowed to depend on time by letting the baseline hazard λ_{0j} change at times a_j (Friedman 1982). A log-linear function between the risk factors and the hazard function is thus assumed. The effect of age can be taken into account by dividing the range of individual dates of birth into $B-1$ birth cohorts $(v_1, v_2]$, ..., $(v_{b-1}, v_b]$, ...,$(v_{B-1}, v_B]$, and then further stratifying the baseline hazard by them (λ_{0jb_i}), where b_i is the birth cohort for the ith individual (Korn, Graubard, and Midthune 1997).

Let us thus denote the hazard of death at time t for the ith individual given the birth cohort b_i and the risk factors $X_i = (x_{i1}, \ldots, x_{im})^\top$ as

$$h^M(t; b_i, X_i) = \prod_{j=1}^{J} \left(\lambda_{ij}^M\right)^{1\{a_{j-1} < t \leq a_j\}},$$

and the hazard of disease incidence as

$$h^D(t; b_i, X_i) = \prod_{j=1}^{J} \left(\lambda_{ij}^D\right)^{1\{a_{j-1} < t \leq a_j\}},$$

where

$$\lambda_{ij}^M = \lambda_{0jb_i}^M e^{X_i^\top \beta^M} = e^{\alpha_{jb_i}^M + X_i^\top \beta^M} = e^{Z_{ij}^\top \gamma^M}, \tag{3}$$

and

$$\lambda_{ij}^D = \lambda_{0jb_i}^D e^{X_i^\top \beta^D} = e^{\alpha_{jb_i}^D + X_i^\top \beta^D} = e^{Z_{ij}^\top \gamma^D}. \tag{4}$$

In this notation, $\alpha_{jb_i}^M = \log \lambda_{ojb_i}^M$ is the logarithm of the baseline hazard of death $(\lambda_{0jb_i}^M)$ and $\alpha_{jb_i}^D = \log \lambda_{ojb_i}^D$ the logarithm of the baseline hazard of disease incidence $(\lambda_{0jb_i}^D)$. Similarly, β^M and β^D are the vectors of regression coefficients for death and disease incidence, respectively, for the covariates X_i, which can be either categorical, continuous or their interactions. Furthermore, Z_{ij} is the vector with length $J \times B + m$ including $J \times B$ indicators of time interval and birth cohort and the covariates corresponding to the regression coefficients $\gamma^M = (\alpha_{11}^M, \ldots, \alpha_{JB}^M, \beta_1^M, \ldots, \beta_m^M)^\top$ and $\gamma^D = (\alpha_{11}^D, \ldots, \alpha_{JB}^D, \beta_1^D, \ldots, \beta_m^D)^\top$. The λ_{ij}^{*M} and λ_{ij}^{*D} follow similarly by replacing X_i by X_i^* in (3) and (4).

3.3. PAF for total mortality

The estimation of PAF for total mortality is described in detail elsewhere (Laaksonen *et al.* 2010b) and is only briefly summarized here.

The probability of death during $(0, a_j]$ for the ith individual given the birth cohort b_i and the risk factors X_i can be calculated as

$$\mathsf{P}\{T_i^M \leq a_j | b_i, X_i\} = 1 - S_{ij}^M,$$

where $S_{ij}^M = e^{-\sum_{k=1}^{j} \lambda_{ik}^M (a_k - a_{k-1})}$ is the survival up to time a_j and $j \in \{1, \ldots, J\}$. The S_{iJ}^{*M} follows similarly by replacing X_i by X_i^* in this formula. Thus, according to Equation 1 the PAF for total mortality during $(0, a_J]$, $\text{PAF}_{(0,a_J]}^M$, can be calculated as

$$\text{PAF}(T^M \leq a_J) = 1 - \frac{\sum_{i=1}^{n} \left(1 - S_{iJ}^{*M}\right)}{\sum_{i=1}^{n} \left(1 - S_{iJ}^{M}\right)}. \tag{5}$$

The PAF for total mortality at any chosen interval $(a_{j-1}, a_j]$, $\text{PAF}_{(a_{j-1},a_j]}^M$, can be calculated similarly by using probabilities $\mathsf{P}\{a_{j-1} < T_i^M \leq a_j | b_i, X_i\} = S_{i,j-1}^M - S_{ij}^M$.

In order to estimate the PAF for total mortality, written briefly PAF in here, we first need to estimate the model parameters γ^M. Estimation of these parameters is based on data of the individual follow-up times until death or censoring, whichever comes first: $T_i = \min(T_i^M, T_i^C)$. In this study, maximum likelihood estimation is used and the SAS procedure LIFEREG is used to compute these maximum likelihood estimates $\hat{\gamma}^M$ and their estimated covariance matrix $\hat{\Sigma}^M$. The point estimate of PAF, $\widehat{\text{PAF}}$, is then obtained by replacing the unknown parameter values γ^M in Equation 5 by their maximum likelihood estimates $\hat{\gamma}^M$. A symmetrizing monotone strictly increasing complementary logarithmic transformation of PAF, $g(\text{PAF}) = \log(1 - \text{PAF})$, is used to obtain an approximate 95% confidence interval of PAF. According to the delta method

$$\sqrt{n}(g(\widehat{\text{PAF}}) - g(\text{PAF})) \xrightarrow{D} N(0, \sigma_{g(\text{PAF})}^2).$$

The confidence interval of the transformation of PAF is then obtained by

$$g(\widehat{\text{PAF}}) \pm 1.96 \times \sqrt{\hat{\sigma}_{g(\text{PAF})}^2}, \tag{6}$$

where the limiting variance of g(PAF) can be consistently estimated by

$$\hat{\sigma}_{g(\text{PAF})}^2 = \left(\frac{\partial g(\text{PAF})}{\partial \gamma^M}\right)^{\top} \hat{\Sigma}^M \left(\frac{\partial g(\text{PAF})}{\partial \gamma^M}\right)\Big|_{\gamma^M = \hat{\gamma}^M}. \tag{7}$$

The confidence interval is finally transformed back to the original scale using the inverse of the complementary logarithmic transformation

$$g^{-1}\left(g(\widehat{\text{PAF}}) \pm 1.96 \times \sqrt{\hat{\sigma}_{g(\text{PAF})}^2}\right). \tag{8}$$

3.4. PAF for disease incidence

The estimation of PAF for disease incidence is described in detail elsewhere (Laaksonen *et al.* 2010a) and is only briefly summarized here.

The probability of disease occurrence, when also the time of death is taken into account, for the ith individual given the birth cohort b_i and the risk factors X_i, can be calculated as

$$\begin{aligned}
& \mathsf{P}\{T_i^D \leq \min(T_i^M, a_j) | b_i, X_i\} \\
= & \sum_{k=1}^{j} \mathsf{P}\{T_i = T_i^D | a_{k-1} < T_i \leq a_k, b_i, X_i\} \mathsf{P}\{a_{k-1} < T_i \leq a_k | b_i, X_i\} \\
= & \sum_{k=1}^{j} \frac{\lambda_{ik}^D}{\lambda_{ik}^D + \lambda_{ik}^M} \left(S_{i,k-1} - S_{ik}\right),
\end{aligned}$$

where $T_i = \min(T_i^D, T_i^M)$ and $S_{ij} = S_{ij}^D S_{ij}^M = e^{-\sum_{k=1}^{j}(\lambda_{ik}^D + \lambda_{ik}^M)(a_k - a_{k-1})}$ is the disease-free survival up to time a_j. Thus, according to Equation 2 the PAF for the incidence of disease during $(0, a_J]$, $\text{PAF}_{(0,a_J]}^D$, can be calculated as

$$\text{PAF}(T^D \leq \min(T^M, a_J)) = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{\lambda_{ij}^{*D}}{\lambda_{ij}^{*D}+\lambda_{ij}^{*M}} \left(S_{i,j-1}^* - S_{ij}^*\right)}{\sum_{i=1}^{n}\sum_{j=1}^{J} \frac{\lambda_{ij}^{D}}{\lambda_{ij}^{D}+\lambda_{ij}^{M}} \left(S_{i,j-1} - S_{ij}\right)}. \tag{9}$$

In order to estimate the PAF for disease incidence, written briefly PAF in here, we first need to estimate the model parameters γ^D and γ^M. Estimation of these parameters is based on data of the individual follow-up times until the occurrence of the disease, death or censoring, whichever comes first: $T_i = \min(T_i^D, T_i^M, T_i^C)$. Similarly to the Section 3.3, the maximum likelihood estimation method is used. The maximum likelihood estimates $\hat{\gamma}^D$ and $\hat{\gamma}^M$ are asymptotically independent (the Fisher information matrix is block-diagonal). The asymptotic distribution of

$$\sqrt{n}\begin{pmatrix} \hat{\gamma}^D - \gamma^D \\ \hat{\gamma}^M - \gamma^M \end{pmatrix} \text{ is then } N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_n^D & 0 \\ 0 & \Sigma_n^M \end{pmatrix}\right).$$

The point estimate of PAF, $\widehat{\text{PAF}}$, is obtained by replacing the unknown parameter values γ^D and γ^M in Equation 9 by their maximum likelihood estimates $\hat{\gamma}^D$ and $\hat{\gamma}^M$. The approximate 95% confidence interval of the transformation $g(\text{PAF}) = \log(1 - \text{PAF})$ is then obtained as in Equation 6, where the limiting variance of g(PAF) can be consistently estimated by

$$\begin{aligned} \hat{\sigma}^2_{g(\text{PAF})} &= \left(\frac{\partial g(\text{PAF})}{\partial \gamma^D}\right)^\top \hat{\Sigma}^D \left(\frac{\partial g(\text{PAF})}{\partial \gamma^D}\right)\Big|_{\gamma^D = \hat{\gamma}^D} \\ &+ \left(\frac{\partial g(\text{PAF})}{\partial \gamma^M}\right)^\top \hat{\Sigma}^M \left(\frac{\partial g(\text{PAF})}{\partial \gamma^M}\right)\Big|_{\gamma^M = \hat{\gamma}^M}. \end{aligned} \tag{10}$$

The confidence interval is finally transformed back to the original scale according to Equation 8.

3.5. PAF in the presence of potential effect modification

In the calculation of PAF, we may want to consider the potential effect modification, i.e., whether the relationship between the risk factor and the outcome of interest, and thus potentially also PAF, varies according to the values of a potential effect modifying factor. Here, the potential effect modifying factor is assumed to be categorical. To analyze the impact of the potential effect modifying factor, an interaction term between the risk factor and the potential effect modifying factor is included in the model, giving separate estimates for the risk factor in the different categories of the potential effect modifying factor. Separate PAF estimates are then calculated in the subpopulations defined by the categories of the effect modifying factor. The statistical significance of interaction can be determined by calculating the 95% confidence intervals for the differences between these PAF estimates. If the confidence interval does not cover zero the difference between the PAF estimates is considered to be statistically significant. For example, in case of an effect modifying factor with two categories, we calculate two separate PAF estimates PAF_1 and PAF_2 and estimate the difference $\widehat{\text{PAF}}_1 - \widehat{\text{PAF}}_2$ and its 95% confidence interval

$$\left(\widehat{\text{PAF}}_1 - \widehat{\text{PAF}}_2\right) \pm 1.96 \times \sqrt{\hat{\sigma}^2_{(\text{PAF}_1 - \text{PAF}_2)}},$$

where PAF is used to denote either PAF for total mortality or PAF for the incidence of disease. The variance of PAF difference is obtained using the delta method, where the limiting variance of $\mathrm{PAF}_1 - \mathrm{PAF}_2$ can be consistently estimated by

$$\begin{aligned}\hat{\sigma}^2_{(\mathrm{PAF}_1-\mathrm{PAF}_2)} &= \left(\frac{\partial(\mathrm{PAF}_1-\mathrm{PAF}_2)}{\partial\gamma^D}\right)^\top \hat{\Sigma}^D \left(\frac{\partial(\mathrm{PAF}_1-\mathrm{PAF}_2)}{\partial\gamma^D}\right)\Big|_{\gamma^D=\hat{\gamma}^D} \\ &+ \left(\frac{\partial(\mathrm{PAF}_1-\mathrm{PAF}_2)}{\partial\gamma^M}\right)^\top \hat{\Sigma}^M \left(\frac{\partial(\mathrm{PAF}_1-\mathrm{PAF}_2)}{\partial\gamma^M}\right)\Big|_{\gamma^M=\hat{\gamma}^M}.\end{aligned}$$

Similarly, we can also use main effect models and calculate and compare PAF estimates in subpopulations defined by some other factor of interest (such as sex) included in the model concerned.

4. SAS modules

The estimation procedure of PAF for total mortality or disease incidence, in the presence of confounding factors and effect modification, is organized as a sequence of SAS macros. This section outlines the functionality of these macros so that an advanced user can make use of them. The use of these macros requires SAS 9.2 procedures `LIFEREG`, `LOGISTIC`, `TRANSPOSE`, `SQL`, and `IML` (SAS Institute Inc. 2010).

To perform PAF analysis, a data preparation procedure is required to create input data files for the main SAS macro, `PAF_M` for total mortality or `PAF_D` for disease incidence (see Figure 1). The main macros, `PAF_M` and `PAF_D`, are composed of the following steps:

1. The main macro calls the macro `EST_MATRIX` to prepare the design matrices (Z and Z^* in Equation 3 or Equation 4), obtained using the `LOGISTIC` procedure, and to produce the parameter estimates ($\hat{\gamma}^M$ in Equation 3 or $\hat{\gamma}^D$ in Equation 4) and their estimated covariances ($\hat{\Sigma}^M$ in Equation 7 or $\hat{\Sigma}^M$ and $\hat{\Sigma}^D$ in Equation 10), obtained using the `LIFEREG` procedure.

2. The main macro calls the macro `EST_PAF_M` for total mortality or `EST_PAF_D` for disease incidence to calculate the PAF estimates, their standard errors and 95% confidence intervals (the `IML` procedure) using the formulas provided in Sections 3.3 and 3.4.

3. The main macro prints out the relative risks and PAF estimates together with their 95% confidence intervals for the risk factors of interest (a more comprehensive output from the `LIFEREG` procedure is optional).

When subpopulation analyses with respect to a certain factor of interest are made, the design matrices (Z, Z^*) created at Step 1 are divided into separate design matrices (Z_1, Z_1^*, ..., Z_K, Z_K^*) according to the K categories of this factor. Then, the macros at Step 2 are called K times, and the $K(K-1)/2$ differences between the subpopulation-specific PAF estimates are analyzed by calling yet another macro (`EST_PAF_DIFF_M` for total mortality and `EST_PAF_DIFF_D` for disease incidence) in which the formulas given in Section 3.5 are implemented. A more detailed description of the functioning of all these macros is given in Section 4.2 after the description of the data preparation in Section 4.1.

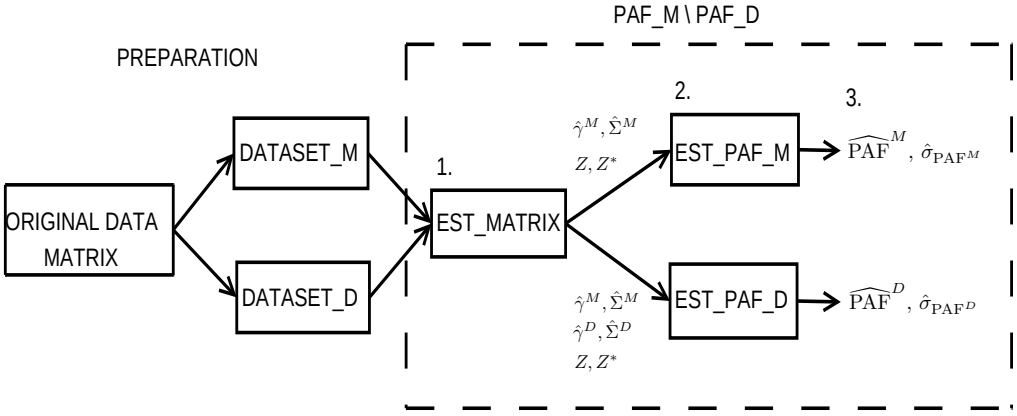

Figure 1: Estimation procedure of PAF.

4.1. Preparation of data for PAF analysis

In the original data matrix, the rows usually correspond to individuals. On the other hand, the columns correspond to individual attributes: identification number (`ID`), birth year (`BYEAR`, birth cohorts (`B_COHORT`), and risk factors of interest. The categorical risk factors include: gender (`SEX`), blood pressure status (`BP`), smoking status (`SMOKE`), age group (`AGEGRP`) and body mass index (`BMI_2`, indicating whether BMI is less than (= `1`) or greater than or equal to (= `2`) 25kg/m^2). The risk factors can also be continuous; for example, age (`AGE`) and body mass index (`BMI`). Refer to the SAS data set `example`. If the outcome of interest is death, a binary variable (`0/1`) indicating whether the person died during the follow-up (`DEATH`) and the follow-up time to the occurrence of death or censoring (`DEATH_FT`) should be included in the data matrix (see SAS data set `example`). If the outcome of interest is disease, also a binary variable (`0/1`) indicating whether the disease occurred during the follow-up (`DIAB`) and the follow-up time to the occurrence of the disease or censoring (`DIAB_FT`) should be included in the data matrix (see SAS data set `example`). When the order of the occurrence of the disease and death is followed for a certain time, for each individual we observe one of the four possible combinations demonstrated, theoretically, in Table 1 and, in practice, in Table 2 which shows a sample of four individuals from the SAS data set `example`.

For the PAF analysis, the original data matrix must be duplicated according to the number of follow-up time intervals used in the piecewise constant hazards model, to create input data

Observed event	`DEATH_FT`	`DEATH`	`DIAB_FT`	`DIAB`
1. $T^M < \min(T^D, T^C)$	T^M	1	T^M	0
2. $T^D < T^M < T^C$	T^M	1	T^D	1
3. $T^D < T^C < T^M$	T^C	0	T^D	1
4. $T^C < \min(T^M, T^D)$	T^C	0	T^C	0

Table 1: Four different possible types of observed events and notation related to them.

ID	DEATH	DEATH_FT	DIAB	DIAB_FT	BYEAR	B_COHORT	AGE	AGEGRP	SEX	BMI	BMI_2	BP	SMOKE
3	1	3.84	0	3.84	1905	1890	72	4	2	21.93	1	2	1
19	1	11.05	1	10.30	1916	1910	61	3	2	29.72	2	2	1
83	0	16.92	1	2.83	1921	1910	56	2	1	30.12	2	2	2
87	0	16.92	0	16.92	1920	1910	57	2	1	24.11	1	2	2

Table 2: Selected sample from SAS data `example` (four individuals).

ID	DEATH	F_TIME	F_PERIOD	BYEAR	B_COHORT	AGE	AGEGRP	SEX	BMI	BMI_2	BP	SMOKE
3	1	3.84	0	1905	1890	72	4	2	21.93	1	2	1
19	0	5.00	0	1916	1910	61	3	2	29.72	2	2	1
19	0	5.00	5	1916	1910	61	3	2	29.72	2	2	1
19	1	1.05	10	1916	1910	61	3	2	29.72	2	2	1
83	0	5.00	0	1921	1910	56	2	1	30.12	2	2	2
83	0	5.00	5	1921	1910	56	2	1	30.12	2	2	2
83	0	5.00	10	1921	1910	56	2	1	30.12	2	2	2
83	0	1.92	15	1921	1910	56	2	1	30.12	2	2	2
87	0	5.00	0	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	5	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	10	1920	1910	57	2	1	24.11	1	2	2
87	0	1.92	15	1920	1910	57	2	1	24.11	1	2	2

Table 3: Selected sample from SAS data `example_death` (four individuals).

matrix in which one row corresponds to one follow-up time interval. The total follow-up time must be divided into these intervals (`F_TIME`). The input data matrix must also contain a categorical variable which indicates the cumulative follow-up time by the beginning of each follow-up time interval (`F_PERIOD`). If the baseline hazard in the piecewise constant hazards model is stratified by the birth cohort, a categorical variable, birth cohort indicator (`B_COHORT`), should be included in the input data matrix. Note that it is the user's responsibility to ensure that the choice of the follow-up time intervals and birth cohorts leads to sufficient cases in all strata of the baseline hazard variables (no zero cells) for reliable estimation of PAF.

If the outcome of interest in the PAF analysis is death, the follow-up is continued until the occurrence of death or censoring, and the follow-up time until the first event must be divided into the chosen follow-up time intervals. Thus, if the length of follow-up time intervals is chosen to be 5 years and the length of birth cohorts 20 years, the original data matrix in Table 2 needs to be modified to appear as the data matrix in Table 3 for the estimation of PAF for total mortality.

If the outcome of interest in the PAF analysis is disease incidence, the follow-up is continued until the occurrence of the disease, death or censoring, and if the first event is disease, then the follow-up time until death or censoring (`DEATH_FT`) needs to be set equal to the time of the occurrence of disease (`DIAB_FT`) and the indicator of death (`DEATH`) to zero (if it originally was one indicating that death would have occurred later during the follow-up). Two separate input data files for disease and death, in which the follow-up time intervals and birth cohorts are of the same length, must be created. The original data matrix in Table 2 thus needs to be modified to appear as the data matrices for death in Table 4 and for disease in Table 5 (where the length of follow-up time intervals is chosen to be 5 years and the length of birth cohorts 20 years) for the estimation of PAF for disease incidence.

ID	DEATH	F_TIME	F_PERIOD	BYEAR	B_COHORT	AGE	AGEGRP	SEX	BMI	BMI_2	BP	SMOKE
3	1	3.84	0	1905	1890	72	4	2	21.93	1	2	1
19	0	5.00	0	1916	1910	61	3	2	29.72	2	2	1
19	0	5.00	5	1916	1910	61	3	2	29.72	2	2	1
19	0	0.30	10	1916	1910	61	3	2	29.72	2	2	1
83	0	2.83	0	1921	1910	56	2	1	30.12	2	2	2
87	0	5.00	0	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	5	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	10	1920	1910	57	2	1	24.11	1	2	2
87	0	1.92	15	1920	1910	57	2	1	24.11	1	2	2

Table 4: Selected sample from SAS data `example_death_2` (four individuals).

ID	DIAB	F_TIME	F_PERIOD	BYEAR	B_COHORT	AGE	AGEGRP	SEX	BMI	BMI_2	BP	SMOKE
3	0	3.84	0	1905	1890	72	4	2	21.93	1	2	1
19	0	5.00	0	1916	1910	61	3	2	29.72	2	2	1
19	0	5.00	5	1916	1910	61	3	2	29.72	2	2	1
19	1	0.30	10	1916	1910	61	3	2	29.72	2	2	1
83	1	2.83	0	1921	1910	56	2	1	30.12	2	2	2
87	0	5.00	0	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	5	1920	1910	57	2	1	24.11	1	2	2
87	0	5.00	10	1920	1910	57	2	1	24.11	1	2	2
87	0	1.92	15	1920	1910	57	2	1	24.11	1	2	2

Table 5: Selected sample from SAS data `example_disease` (four individuals).

The original data matrix (`example`) can be transformed into the required input data matrices (`example_death`, `example_death_2`, `example_disease`) by calling `GEN_DATA`:

```
%GEN_DATA(INDATA = , IDVARIABLE = ,
   CENSORVARIABLE_1 = , CENSORVALUE_1 = , FTVARIABLE_1 = ,
   CENSORVARIABLE_2 = , CENSORVALUE_2 = , FTVARIABLE_2 = ,
   FTLENGTH = , PERIODLENGTH = , DROPVARIABLES = , OUTDATA = ,
   PERIODVARIABLE = , TIMEVARIABLE = );
```

Input arguments of `GEN_DATA` are defined as follows:

- `INDATA =` *SAS data set*
  specifies original data set.
- `IDVARIABLE =` *variable*
  identifies each individual (see variable `ID` in data set `example`).
- `CENSORVARIABLE_1 =` *variable*
  indicates whether censoring before the outcome of event 1 occurred. Must be a binary variable (`0`/`1`).
- `CENSORVALUE_1 = 0` or `1`
  specifies value of the `CENSORVARIABLE_1 =` *variable* (default value is `0`).
- `FTVARIABLE_1 =` *variable*
  indicates the individual length of follow-up until the outcome of event 1 or censoring.

- CENSORVARIABLE_2 = *variable*
  indicates whether censoring before the outcome of event 2 occurred. Must be a binary variable (0/1).
- CENSORVALUE_2 = 0 or 1
  specifies value of the CENSORVARIABLE_2 = *variable* (default value is 0).
- FTVARIABLE_2 = *variable*
  indicates the individual length of follow-up until the outcome of event 2 or censoring.
- FTLENGTH = *number*
  indicates the maximum length of follow-up time.
- PERIODLENGTH = *number*
  indicates the length of follow-up time intervals. Must be the same in the input data matrices for death (example_death_2) and disease (example_disease) for the estimation of PAF for disease incidence.
- DROPVARIABLES = *variables*
  specifies the variables to be dropped out from the output data set (the variables related to disease when preparing the input data matrix for death, the indicator variable for the disease occurrence during the follow-up (DIAB) and the follow-up time until the disease occurrence (DIAB_FT), and the follow-up time to the occurrence of death (DEATH_FT), and vice versa).
- OUTDATA = *SAS data set*
  specifies output data set.

If PAF for total mortality is to be estimated, the outcome of interest is death and only censoring variable (CENSORVARIABLE_1), censoring value (CENSORVALUE_1) and follow-up time variable (FTVARIABLE_1) related to death need to be given in the GEN_DATA macro call to prepare the required input data set. In the preparation of this data set, a new follow-up period variable (F_PERIOD) indicating the cumulative follow-up time by the beginning of each follow-up time interval is created, the total follow-up time until the occurrence of death or censoring is divided into different follow-up time intervals (F_TIME), the information not needed (related to disease incidence or total follow-up time) is dropped, and the information that remains the same from one interval to another (birth year (BYEAR) and cohort (B_COHORT), and the risk factors of interest measured at baseline) are dropped (see SAS data set example_death in Table 3).

If PAF for disease incidence is to be estimated, the outcome of interest is disease, but also censoring due to death is taken into account, and thus separate input data sets for disease and death are prepared for the PAF analysis. When the input data set for disease is prepared (see SAS data set example_disease in Table 5) only censoring variable (CENSORVARIABLE_1), censoring value (CENSORVALUE_1) and follow-up time variable (FTVARIABLE_1) related to disease need to be given in the GEN_DATA macro call. When the input data set for death is prepared (see SAS data set example_death_2 in Table 4) both censoring variable (CENSORVARIABLE_1), censoring value (CENSORVALUE_1) and follow-up time variable (FTVARIABLE_1) related to death and censoring variable (CENSORVARIABLE_2), censoring value (CENSORVALUE_2) and follow-up time variable (FTVARIABLE_2) related to disease need to be given in the GEN_DATA

macro call. If the first event is disease (CENSORVALUE_2 = 1), then the follow-up time until death or censoring is set equal to the time of the occurrence of disease and the indicator of death (DEATH) is set to zero .

A more detailed description of the function of the GEN_DATA macro is given in the SAS program gen_data.sas.

4.2. PAF analysis

When the input data matrices have been prepared as described in previous Section 4.1, a SAS macro, PAF_M, for the estimation of PAF for total mortality:

```
%PAF_M(DATASET_M = , CENSORVARIABLE_M = , CENSORVALUE_M = ,
   TIMEVARIABLE_M = , PERIODVARIABLE_M = , IDVARIABLE = ,
   COHORTVARIABLE = , DELTALENGTH = , CLASSVARIABLES = ,
   CLASSORDER = , COVARIATE_MODEL = , GROUPVARIABLE = ,
   MODIFICATIONS = , PRINT = );
```

or a SAS macro, PAF_D, for the estimation of PAF for disease incidence:

```
%PAF_D(DATASET_M = , CENSORVARIABLE_M = , CENSORVALUE_M = ,
   TIMEVARIABLE_M = , PERIODVARIABLE_M = , DATASET_D = ,
   CENSORVARIABLE_D = , CENSORVALUE_D = , TIMEVARIABLE_D = ,
   PERIODVARIABLE_D = , IDVARIABLE = , COHORTVARIABLE = ,
   DELTALENGTH = , CLASSVARIABLES = , CLASSORDER = ,
   COVARIATE_MODEL = , GROUPVARIABLE = , MODIFICATIONS = ,
   PRINT = );
```

is called depending on the outcome of interest. Input arguments of PAF_M and PAF_D are defined as follows:

- DATASET_M = *SAS data set*
  specifies input data set for death. Must be of the form explained in the previous Section 4.1 (see data sets example_death and example_death_2 in Tables 3 and 4).
- CENSORVARIABLE_M = *variable*
  indicates whether censoring before death occurred. Must be a binary variable (0/1) (see variable DEATH in data sets example_death and example_death_2 in Tables 3 and 4 in Section 4.1).
- CENSORVALUE_M = 0 or 1
  specifies value of the CENSORVARIABLE_M = *variable* (default value is 0).
- TIMEVARIABLE_M = *variable*
  indicates the individual length of follow-up until death or censoring within each follow-up time interval (see variable F_TIME in data sets example_death and example_death_2 in Tables 3 and 4 in Section 4.1).
- PERIODVARIABLE_M = *variable*
  indicates the cumulative follow-up time by the beginning of each follow-up time interval

in input data set for death (see variable F_PERIOD in data sets example_death and example_death_2 in Tables 3 and 4 in Section 4.1).

- DATASET_D = *SAS data set*
  specifies input data set for disease. Must be of the form explained in the previous Section 4.1 (see data set example_disease in Table 5).
- CENSORVARIABLE_D = *variable*
  indicates whether censoring before incidence of disease occurred. Must be a binary variable (0/1) (see variable DIAB in data set example_disease in Table 5 in Section 4.1).
- CENSORVALUE_D = 0 or 1
  specifies value of the CENSORVARIABLE_D = *variable* (default value is 0).
- TIMEVARIABLE_D = *variable*
  indicates the individual length of follow-up until disease occurrence or censoring (due to death or end of follow-up) within each follow-up time interval (see variable F_TIME in data set example_disease in Table 5 in Section 4.1).
- PERIODVARIABLE_D = *variable*
  indicates the cumulative follow-up time by the beginning of each follow-up time interval in input data set for disease (see variable F_PERIOD in data set example_disease in Table 5 in Section 4.1).
- IDVARIABLE = *variable*
  identifies each individual (see variable ID in data sets example_death, example_death_2 and example_disease in Tables 3, 4, and 5 in Section 4.1).
- COHORTVARIABLE = *variable*
  indicates to which birth cohort the individual belongs (see variable B_COHORT in data sets example_death, example_death_2 and example_disease in Tables 3, 4, and 5 in Section 4.1). If this is omitted the baseline hazard in the piecewise constant hazard model is only stratified according to the follow-up time intervals.
- DELTALENGTH = *variable*
  indicates the maximum length of follow-up time.
- CLASSVARIABLES = *variables*
  specifies categorical variables included in the model, separated by blanks (such as the variables SEX, BMI_2, BP, and SMOKE in data sets example_death, example_death_2 and example_disease in Tables 3, 4, and 5 in Section 4.1).
- CLASSORDER = DESC or ASC
  specifies descending (DESC) or ascending (ASC) order of the categories of the variables in the LIFEREG analysis (affects the reference category).
- COVARIATE_MODEL = *variables*
  specifies all variables (categorical, continuous, and their interactions) included in the model, separated by blanks. Interactions between categorical or continuous variables are denoted by an asterisk between the variables (VARIABLE1*VARIABLE2). The COVARIATE_MODEL = *variables* correspond to the right-hand variables in MODEL statement.

- GROUPVARIABLE = *variable*
  defines a categorical variable of interest that constitutes the subgroups in which the PAF estimates are calculated separately. If the *variable* is considered to be a potential effect modifying factor, an interaction term between the *variable* and the risk factor of interest must be included in the COVARIATE_MODEL statement. If GROUPVARIABLE is omitted, the PAF estimate is computed using the entire sample.

- MODIFICATIONS = *variable=value*
  determines the reference category value for the risk factors of interest to which individuals are hypothetically moved in the calculation of PAF. If individuals from only some of the categories of the risk factor are moved to the reference category, then the restriction statement is denoted using the IF-THEN statement of the SAS language. If several risk factors are modified simultaneously, they are separated by slashes. The variables included in the COVARIATE_MODEL statement but not in the MODIFICATIONS statement are treated as confounding factors.

- PRINT = YES or NO
  defines what is printed out from the LIFEREG procedure: PRINT = YES requests all output results while the PRINT = NO option prints only the relative risks (RR) and their 95% confidence intervals (CI; formed with the help of the PROC LIFEREG output). The default is PRINT = NO.

4.3. PAF_M macro

If the outcome of interest is death, input data matrix (DATASET_M), censoring variable (CENSORVARIABLE_M) and censoring value (CENSORVALUE_M) related to death are given in the PAF_M macro call. A more detailed description of the function of this macro is given in the SAS program paf_m.sas. First, the PAF_M macro calls another macro, EST_MATRIX, which prepares the design matrices, vector of parameter estimates and covariance matrix of parameter estimates related to death needed for the calculation of PAF and its confidence interval for total mortality. In addition, the macro produces the relative risks (RR) of death and their 95% confidence intervals for the risk factors of interest given in the COVARIATE_MODEL= option of the PAF_M macro call (a more comprehensive output from the LIFEREG procedure can also be obtained using the PRINT option of the PAF_M macro). Also the convergence status of the LIFEREG analysis is provided. Outputs of PAF_M are listed as follows:

- the data set DESIGN_1 which indicates to which categories of the baseline hazard variables (follow-up time intervals, birth cohorts, and their interactions) each individual belongs on each follow-up time interval and which values of the risk factors each individual has (matrix Z in (3)).

- the data set DESIGN_2 which indicates to which categories of the baseline hazard variables (follow-up time intervals, birth cohorts, and their interactions) each individual belongs on each follow-up time interval and which values of the risk factors each individual has after their hypothetical change to the chosen reference categories indicated in the MODIFICATIONS statement of the PAF_M macro call (matrix Z^*).

- the data set &CENSORVARIABLE_M._EST which contains the column vector of parameter estimates for the baseline hazard variables ($\hat{\alpha}^M$) and the risk factors ($\hat{\beta}^M$) related to death, where &CENSORVARIABLE_M is substituted with the name of the censoring variable given in the PAF_M macro call ($\hat{\gamma}^M$ in (3)).
- the data set &CENSORVARIABLE_M._COVB which contains the covariance matrix of the parameter estimates for the baseline hazard variables and the risk factors related to death, where &CENSORVARIABLE_M is substituted with the name of the censoring variable given in the PAF_M macro call ($\hat{\Sigma}^M$ in (7)).
- the data set &CENSORVARIABLE_M._RR which contains relative risks of death and their 95% confidence intervals for the risk factors given in the COVARIATE_MODEL statement of the PAF_M macro call, where &CENSORVARIABLE_M is substituted with the name of the censoring variable given in the PAF_M macro call.
- the data set &CENSORVARIABLE_M._CONV which indicates the convergence status of the model estimation of PROC LIFEREG.

The function of this macro is described in more detail in est_matrix.sas. Second, the PAF_M macro calls a macro EST_PAF_M which calculates the point estimate, standard error and 95% confidence interval of PAF for total mortality using the formulas described in Section 3.3. A more detailed description of these calculations in the SAS/IML language is provided in est_paf_m.sas. Finally, the PAF_M macro prints out both the relative risks of death and their 95% confidence intervals for the risk factors specified in the COVARIATE_MODEL= option of the macro call as well as the piecewise and cumulative point estimates, standard errors and 95% confidence intervals of PAF for total mortality. If there is a problem in the convergence of the model chosen, an error message is generated.

4.4. PAF_D macro

If the outcome of interest is disease, input data matrices (DATASET_M, DATASET_D), censoring variables (CENSORVARIABLE_M, CENSORVARIABLE_D) and censoring values (CENSORVALUE_M, CENSORVALUE_D) related to both death and disease are given in the PAF_D macro call. In this case, the follow-up time must be divided into time intervals of the same length (TIMEVARIABLE_M, TIMEVARIABLE_D) and birth cohorts (B_COHORT) must be of the same length in both input data matrices (see SAS data sets example_death_2 and example_disease in Tables 4 and 5 in Section 4.1). Furthermore, the variables specified in the IDVARIABLE=, CLASSVARIABLES=, and COVARIATE_MODEL= options must be identical in both input data matrices. A more detailed description of the function of this macro is given in paf_d.sas. First, the PAF_D macro calls the macro EST_MATRIX separately for death and disease to prepare the design matrices (Z and Z^*), vectors of parameter estimates ($\hat{\gamma}^M$ and $\hat{\gamma}^D$) and their covariance matrices ($\hat{\Sigma}^M$ and $\hat{\Sigma}^D$) in (3), (4) and (10) needed for the calculation of PAF and its confidence interval for disease incidence (see est_matrix.sas). Second, the PAF_D macro calls the macro EST_PAF_D which using these outputs as its inputs calculates the point estimate, standard error and 95% confidence interval of PAF for disease incidence using the formulas described in Section 3.4. A more detailed description of these calculations in the SAS/IML language is provided in est_paf_d.sas. Finally, the PAF_D macro prints out both the relative risks of both death and incidence of disease and their 95% confidence intervals for the

risk factors specified in the COVARIATE_MODEL= option of the macro call as well as the point estimate, standard error and 95% confidence interval of PAF for disease incidence. If there is a problem in the convergence of the model chosen, an error message is delivered.

If the GROUPVARIABLE is given in the PAF_M or PAF_D macro call, the point estimate, standard error and 95% confidence interval of PAF either for total mortality or disease incidence are calculated separately in the subgroups defined by this variable. To do this the design matrices DESIGN_1 and DESIGN_2 prepared by the EST_MATRIX macro are divided into separate design matrices according to the categories of the GROUPVARIABLE. Then, depending on the outcome of interest, either the PAF_M macro calls the macro EST_PAF_M (death) or the PAF_D macro calls the macro EST_PAF_D (disease) as many times as there are categories of the GROUPVARIABLE to calculate the separate point estimates, standard errors and 95% confidence intervals of PAF (see paf_m.sas and paf_d.sas for a more detailed description). After this, either the PAF_M macro calls the macro EST_PAF_DIFF_M (death) or the PAF_D macro calls the macro EST_PAF_DIFF_D (disease) which calculates the differences between these PAF estimates, their 95% confidence intervals as well as a p value to determine the statistical significance of these differences using the formulas described in Section 3.5. More detailed descriptions of these calculations in the SAS/IML language are given in est_paf_diff_m.sas and est_paf_diff_d.sas. Finally, the PAF_M and PAF_D macros print out the groupwise point estimates, their standard errors and 95% confidence intervals of PAF either for total mortality or disease incidence as well as the differences between these groupwise PAF estimates, their standard errors and 95% confidence intervals, and a p value for the statistical significance of these differences.

5. Data example

This data example demonstrates the importance of certain modifiable risk factors, alone and in interaction, on type 2 diabetes incidence through calculation of PAF for disease incidence using the SAS modules presented in Section 4 previously. It is based on data from the Mini-Finland Health Survey cohort study carried out in 1978–1980 (Aromaa, Heliövaara, Impivaara, Knekt, and Maatela 1989). Altogether 4,517 men and women aged 40–79, who participated in a health examination and were free of type 2 diabetes and cardiovascular diseases at baseline, were included in this study. Their height and weight were measured at the health examination, and their body mass index (BMI) was calculated. Casual blood pressure was measured and hypertensive medication self-reported. Also smoking was self-reported. The follow-up time was defined as days from the baseline examination to the date of type 2 diabetes occurrence, death, or end of follow-up, whichever came first. During a 17-year follow-up, a total of 227 individuals developed type 2 diabetes. The categorisation of the risk factors of interest (BMI, blood pressure, and smoking) and the distribution of the individuals as well as the type 2 diabetes cases across these categories is presented in Table 8.

The original data matrix, in which the rows correspond to individuals and the columns to information related to them (such as the total follow-up time), is shown in example. In this example, the follow-up time is divided into 5-year intervals. The original data matrix is then modified into two separate input data matrices (example_death_2 and example_disease), in which there is one row for each follow-up time interval for each individual, by making the following GEN_DATA macro calls (see data preparation in Section 4.1):

```
%GEN_DATA(INDATA = X.EXAMPLE, IDVARIABLE = ID,
   CENSORVARIABLE_1 = DEATH, CENSORVALUE_1 = 0, FTVARIABLE_1 = DEATH_FT,
   CENSORVARIABLE_2 = DIAB, CENSORVALUE_2 = 0, FTVARIABLE_2 = DIAB_FT,
   FTLENGTH = 20, PERIODLENGTH = 5, DROPVARIABLES = DIAB DIAB_FT DEATH_FT,
   OUTDATA = X.EXAMPLE_DEATH_2, PERIODVARIABLE = F_PERIOD,
   TIMEVARIABLE = F_TIME);
```

and

```
%GEN_DATA(INDATA = X.EXAMPLE, IDVARIABLE = ID,
   CENSORVARIABLE_1 = DIAB, CENSORVALUE_1 = 0, FTVARIABLE_1 = DIAB_FT,
   CENSORVARIABLE_2 = , CENSORVALUE_2 = , FTVARIABLE_2 = ,
   FTLENGTH = 20, PERIODLENGTH = 5, DROPVARIABLES = DEATH DEATH_FT DIAB_FT,
   OUTDATA = X.EXAMPLE_DISEASE, PERIODVARIABLE = F_PERIOD,
   TIMEVARIABLE = F_TIME);
```

After the preparation of input data matrices, the `PAF_D` macro for the analysis of PAF for disease incidence described in `paf_d.sas` is called. To calculate PAF, for example, for smoking (`SMOKE`) so that the current smokers (categories 3 and 4 in Table 8) are hypothetically moved to the reference category of never smokers (category 1 in Table 8), the following `PAF_D` macro call is made (see `paf_example_d.sas`):

```
%PAF_D(DATASET_M = X.EXAMPLE_DEATH, CENSORVARIABLE_M = DEATH,
   CENSORVALUE_M = 0, TIMEVARIABLE_M = F_TIME, PERIODVARIABLE_M = F_PERIOD,
   DATASET_D = X.EXAMPLE_DISEASE, CENSORVARIABLE_D = DIAB,
   CENSORVALUE_D = 0, TIMEVARIABLE_D = F_TIME, PERIODVARIABLE_D = F_PERIOD,
   IDVARIABLE = ID, COHORTVARIABLE = B_COHORT, DELTALENGTH = 17,
   CLASSVARIABLES = SEX SMOKE, CLASSORDER = DESC,
   COVARIATE_MODEL = SEX SMOKE, GROUPVARIABLE = ,
   MODIFICATIONS = IF SMOKE IN(3,4) THEN SMOKE=1,
   PRINT = NO);
```

where `X` is the path to the current working directory. Note that the results are sex- and age-adjusted as sex is included in the model (variable `SEX` in the `CLASSVARIABLES=` and `COVARIATE_MODEL=` option), and age is taken into account through birth cohort (variable `B_COHORT` in the `COHORTVARIABLE=` option), according to which the baseline hazard of the piecewise constant hazards model is stratified. In this example, 20-year birth cohorts are used (see `example_death_2` and `example_disease`) since there are no diabetes cases among the youngest and the oldest, and thus shorter birth cohort intervals of, for example, 10 years would lead to zero cells and make the convergence of the model questionable. The results of this analysis are shown in Table 6 (and are also presented in Table 8):

Note that although the relative risk (RR) of type 2 diabetes incidence for the individuals belonging to the highest smoking category (category 4: $\geq$ 30 cigarettes/day) is quite high (RR = 2.79, 95% CI = 1.25, 6.20) (Table 6b), the PAF for the hypothetical modification of the individuals, belonging to this category or to the previous category of the second highest smoking, to the reference category is very low (PAF = 0.02, 95% CI = -0.05, 0.09) (Table 6c). This is due to the low prevalence (1.7%) of the individuals belonging to the highest smoking

Relative risks of death

CENSOR	Parameter	LEVEL1	LEVEL2	RR	RR LOWERCL	RR UPPERCL
DEATH	SEX	2		0.6367	0.5514	0.7352
DEATH	SEX	1		1.0000	.	.
DEATH	SMOKE	4		3.0047	2.0331	4.4406
DEATH	SMOKE	3		2.0411	1.7300	2.4082
DEATH	SMOKE	2		1.2418	1.0441	1.4770
DEATH	SMOKE	1		1.0000	.	.

! ! !
Algorithm converged.

(a) Relative risks of death and convergence status of the estimation algorithm.

Relative risks of incidence of disease

CENSOR	Parameter	LEVEL1	LEVEL2	RR	RR LOWERCL	RR UPPERCL
DIAB	SEX	2		1.0302	0.7504	1.4144
DIAB	SEX	1		1.0000	.	.
DIAB	SMOKE	4		2.7868	1.2521	6.2023
DIAB	SMOKE	3		1.1446	0.7724	1.6962
DIAB	SMOKE	2		1.4179	0.9863	2.0385
DIAB	SMOKE	1		1.0000	.	.

! ! !
Algorithm converged.

(b) Relative risks of incidence of disease and convergence status of the estimation algorithm.

PAF for disease incidence

CENSOR	PERIOD	PAF	SE LPAF	LPAF LOWERCL	LPAF UPPERCL
DIAB (DEATH censoring)	0-<17	0.019113	0.036864	-0.054381	0.087485

(c) PAF for disease incidence.

Table 6: PAF for disease, modification of smoking.

category (Table 8), as the PAF measure accounts for both the strength of the association between the risk factor and the outcome (RR) and the prevalence of the risk factor. The corresponding `PAF_D` macro calls for calculating PAF for BMI and blood pressure are demonstrated in `paf_example_d` and the results related to them in Table 8. Here, the PAF estimates for the hypothetical modification of BMI (`BMI_2`) or blood pressure (`BP`), from category with the higher risk (category 2) to reference category (category 1), are much higher (PAF = 0.68, 95% CI = 0.55, 0.77, and PAF = 0.31, 95% CI = −0.06, 0.55, respectively) than that for smoking as the prevalence of the individuals belonging to the modified categories is much higher (59.9% and 85.5%, respectively).

If we have reason to believe that some factors (e.g., blood pressure `BP`) modify the relationship between a risk factor, such as BMI (`BMI_2`), and the outcome, we can analyze this by comparing the PAF results obtained in the subgroups defined by the categories of this potential

Relative risks of death

CENSOR	Parameter	LEVEL1	LEVEL2	RR	RR LOWERCL	RR UPPERCL
DEATH	SEX	2		0.5001	0.4435	0.5639
DEATH	SEX	1		1.0000	.	.
DEATH	BP	2		1.6455	1.2182	2.2227
DEATH	BP	1		1.0000	.	.
DEATH	BP*BMI_2	2	2	0.7864	0.6945	0.8905
DEATH	BP*BMI_2	2	1	1.0000	.	.
DEATH	BP*BMI_2	1	2	0.8236	0.5324	1.2742
DEATH	BP*BMI_2	1	1	1.0000	.	.

! ! !
Algorithm converged.

(a) Relative risks of death and convergence status of the estimation algorithm.

Relative risks of incidence of disease

CENSOR	Parameter	LEVEL1	LEVEL2	RR	RR LOWERCL	RR UPPERCL
DIAB	SEX	2		0.8417	0.6446	1.0990
DIAB	SEX	1		1.0000	.	.
DIAB	BP	2		0.5887	0.2577	1.3446
DIAB	BP	1		1.0000	.	.
DIAB	BP*BMI_2	2	2	5.3018	3.3424	8.4100
DIAB	BP*BMI_2	2	1	1.0000	.	.
DIAB	BP*BMI_2	1	2	1.7353	0.6965	4.3235
DIAB	BP*BMI_2	1	1	1.0000	.	.

! ! !
Algorithm converged.

(b) Relative risks of incidence of disease and convergence status of the estimation algorithm.

PAF for disease incidence

CENSOR	PERIOD	PAF	SE LPAF	LPAF LOWERCL	LPAF UPPERCL
DIAB (DEATH censoring)	0-<17	0.69216	0.17766	0.56394	0.78268

PAF for disease incidence in subgroups of BP

CENSOR	GROUP	PERIOD	PAF	SE LPAF	LPAF LOWERCL	LPAF UPPERCL
DIAB (DEATH censoring)	BP = 1	0-<17	0.2496	0.2665	-0.2650	0.5549
DIAB (DEATH censoring)	BP = 2	0-<17	0.7333	0.2101	0.5975	0.8233

(c) PAF for disease incidence.

Statistical significance of differences between groupwise PAF estimates

CENSOR	GROUP	VALUES	PERIOD	PAF GROUP A	PAF GROUP B	PAF DIFF	SE PAF DIFF	PAF DIFF LOWERCL	PAF DIFF UPPERCL	PAF DIFF Z	PAF DIFF PVALUE
DIAB (DEATH censoring)	BP	1 - 2	0-<17	0.2496	0.7333	-0.4837	0.2076	-0.8906	-0.0769	2.3302	0.0198

(d) PAF for disease incidence in subgroups of BP.

Table 7: PAF for Disease, modification of BMI in subgroups defined by blood pressure.

effect modifying factor. To perform this subgroup PAF analysis, the following PAF_D macro call can be made:

```
%PAF_D(DATASET_M = X.EXAMPLE_DEATH, CENSORVARIABLE_M = DEATH,
   CENSORVALUE_M = 0, TIMEVARIABLE_M = F_TIME, PERIODVARIABLE_M = F_PERIOD,
   DATASET_D = X.EXAMPLE_DISEASE, CENSORVARIABLE_D = DIAB,
   CENSORVALUE_D = 0, TIMEVARIABLE_D = F_TIME, PERIODVARIABLE_D = F_PERIOD,
   IDVARIABLE = ID, COHORTVARIABLE = B_COHORT, DELTALENGTH = 17,
   CLASSVARIABLES = SEX BP BMI_2, CLASSORDER = DESC,
   COVARIATE_MODEL = SEX BP BP*BMI_2,
   GROUPVARIABLE = BP, MODIFICATIONS = BMI_2=1, PRINT = NO);
```

where X is the path to the current working directory. Thus, an interaction term between the risk factor and potential effect modifying factor must be included in the COVARIATE_MODEL statement and the potential effect modifying factor in the GROUPVARIABLE statement. The results of this analysis are shown in Table 7 (and are also presented in Table 8):

In this case, the effect modification by blood pressure turned out to be statistically significant (p value = 0.02) (Table 7d) and the PAF for the hypothetical modification of BMI was much higher for those with elevated blood pressure (PAF = 0.73, 95% CI = 0.60, 0.82) than for those with normal blood pressure (PAF = 0.25, 95% CI = −0.27, 0.55) (Table 7c).

The estimation and statistical inference of total mortality PAF with the GROUPVARIABLE = BP option are carried out in a similar fashion to the other examples of PAF for total mortality in paf_example_m.sas. The only difference is that in case of the estimation of PAF for total mortality both piecewise and cumulative PAF estimates and their standard errors and 95% confidence intervals are produced.

6. Discussion

The methods for the estimation of PAF in a cohort study design, taking adequately into account the follow-up time, have been developed during recent years (Chen *et al.* 2006; Samuelsen and Eide 2008; Cox *et al.* 2009; Laaksonen *et al.* 2010a,b). As far as these authors know, no publicly available program to implement these methods has, however, yet been provided, apparently hindering their wider application. The SAS macros based on the recently developed methods (Laaksonen *et al.* 2010a,b) presented in this paper close the gap between theory and application. Using these macros it is now possible to estimate PAF and its confidence interval in a cohort study design both for total mortality (Laaksonen *et al.* 2010b) and for disease incidence (Laaksonen *et al.* 2010a), taking into account the different sources of censoring. The PAF macros are very flexible in that both categorical and continuous risk factors and confounding factors as well as their interactions can be included in the model, as long as the estimation algorithm still converges. In addition, the estimation of PAF in the presence of potential effect modification and analysis of its statistical significance are possible. However, as both the prevalence of the risk factors and the strength of the association between the risk factors and the outcome affect PAF, both of these components should be taken into account when analyzing the potential effect modification to make sure that it is due to the difference in strength of the association between the risk factor and outcome and not just due to different prevalences. Different prevalences can also prevent a difference in strength

Variable	Category	Modification[1]	Disease cases	Total n	Prevalence (%)	RR (95% CI)	PAF (95% CI)
Non-modifiable risk factors							
Sex	1: Men		99	2004	44.4		
(`SEX`)[3]	2: Women		128	2513	55.6		
Age	1: 40–49		45	1576	34.9		
(`AGE_4`)[3]	2: 50–59		77	1431	31.7		
	3: 60–69		75	952	21.1		
	4: 70–79		30	558	12.3		
Modifiable risk factors							
Smoking	1: Never smoker		124	2598	57.6	1	
(`SMOKE`)[3]	2: Former smoker		57	942	20.9	1.42 (0.99, 2.04)	
	3: Pipe/cigar only or $<$ 30 cigarettes/day		39	893	19.8	1.14 (0.77, 1.70)	
	4: $\geq$ 30 cigarettes/day	3,4 $\rightarrow$ 1	7	79	1.7	2.79 (1.25, 6.20)*	0.02 ($-$0.05, 0.09)
Body mass index (`BMI`)							
(kg/m^2)	1: $<$ 25.0		28	1809	40.1	1	
(`BMI_2`)[3]	2: $\geq$ 25.0	2 $\rightarrow$ 1	199	2705	59.9	4.48 (3.01, 6.66)*	0.68 (0.55, 0.77)*
Blood pressure[4]	1: Normal		19	653	14.5	1	
(BP)[3]	2: Elevated	2 $\rightarrow$ 1	208	3862	85.5	1.63 (1.01, 2.63)*	0.31 ($-$0.06, 0.55)
Blood pressure$*$BMI[2]	1: Normal, $<$ 25.0		8	378	8.4	1	
(`BP*BMI_2`)[3]	2: Normal, $\geq$ 25.0	2 $\rightarrow$ 1	11	274	6.1	1.74 (0.70, 4.32)	0.25 ($-$0.27, 0.55)
	3: Elevated, $<$ 25.0		20	1430	31.7	1	
	4: Elevated, $\geq$ 25.0	4 $\rightarrow$ 3	188	2430	53.8	5.30 (3.34, 8.41)*	0.73 (0.60, 0.82)*

* Statistically significant association

[1] Hypothetical modification (categories modified $\rightarrow$ reference category) in the calculation of PAF.

[2] p value = 0.02 for the statistical significance of effect modification (i.e., interaction between the potential effect modifier and the risk factor).

[3] Variable name in the data (see data matrices `example_death` or `example_disease` in Section (4.1).

[4] Elevated: Systolic blood pressure $\geq$ 130mmHg or diastolic blood pressure $\geq$85 mmHg or antihypertensive medication. Normal: Not elevated.

Table 8: Prevalences and sex- and age-adjusted relative risks (RR) and population attributable fractions (PAF) together with their 95% confidence intervals (CI) for non-modifiable and modifiable risk factors of type 2 diabetes in a 17-year follow-up of Mini-Finland Health Survey.

of association from becoming significant in the analysis of effect modification. Furthermore, although in this paper it was assumed that the parameter estimates and PAF estimates were calculated based on the same data, it is also possible to calculate the parameter estimates from external data and apply them to the data of interest by using the macros: for the total mortality case, `EST_MATRIX` and `EST_PAF_M`; for the disease case, `EST_MATRIX` and `EST_PAF_D`.

The time-to-event data is modeled based on a proportional hazards model with a piecewise constant baseline hazard function. The baseline hazard in the piecewise constant hazard model can be stratified with respect to both follow-up time and birth cohort. With a judicious choice of the cut-points in the piecewise constant hazard model almost any baseline hazard can be well approximated in large data sets. It should be noticed, however, that it is the user's responsibility to ensure that the choice of the cut-points for follow-up time intervals and birth cohorts results in at least one case at each time interval and birth cohort to guarantee a reliable estimation of PAF. The macro provides information on the convergence status of the model chosen. In addition to the number of the cut-points, the computation time of the macro depends also on the number and type of variables included in the model. In general, however, the computation time is very fast, even with quite closely-spaced cut-points, partly due to the analytic variance estimation.

Acknowledgments

The financial support of the postgraduate school Doctoral Programs in Public Health (DPPH) for the first author is gratefully acknowledged.

References

Aromaa A, Heliövaara M, Impivaara O, Knekt P, Maatela J (1989). "Aims, Methods and Study Population. Part 1." In A Aromaa, M Heliövaara, O Impivaara, P Knekt, J Maatela (eds.), *The Execution of the Mini-Finland Health Survey. (In Finnish, English summary)*. Publications of the Social Insurance Institution, Finland, Helsinki and Turku, ML:88.

Benichou J (1991). "Methods of Adjustment for Estimating the Attributable Risk in Case-Control Studies: A Review." *Statistics in Medicine*, **10**, 1753–1773.

Benichou J (2001). "A Review of Adjusted Estimators of Attributable Risk." *Statistical Methods in Medical Research*, **10**(3), 195–216.

Brady AR (1998). "Adjusted Population Attributable Fractions from Logistic Regression." *Stata Technical Bulletin*, **42**, 8–12.

Chen YQ, Hu C, Wang Y (2006). "Attributable Risk Function in the Proportional Hazards Model for Censored Time-to-Event." *Biostatistics*, **7**(4), 515–29.

Coughlin SS, Benichou J, Weed DL (1994). "Attributable Risk Estimation in Case-Control Studies." *Epidemiologic Reviews*, **16**, 51–64.

Cox C, Chu H, Muñoz A (2009). "Survival Attributable to an Exposure." *Statistics in Medicine*, **28**, 3276–3293.

Eide GE (2006). *How to Estimate Attributable Fractions in Stata: A Simple Introduction.* Center for Clinical Research Report, Haukeland University Hospital, Bergen.

Friedman M (1982). "Piecewise Exponential Models for Survival Data with Covariates." *The Annals of Statistics*, **10**, 101–113.

Grömping U, Weimann U (2004). "The Asymptotic Distribution of the Partial Attributable Risk in Cross-Sectional Studies." *Statistics*, **38**, 427–438.

Kahn MJ, O'Fallon WM, Sicks JD (1998). *Generalized Population Attributable Risk Estimation.* Technical Report 54, Mayo Foundation, Rochester, Minnesota.

Korn EL, Graubard BI, Midthune D (1997). "Time-to-Event Analysis of Longitudinal Follow-Up of a Survey: Choice of the Time-Scale." *American Journal of Epidemiology*, **145**(1), 72–80.

Laaksonen MA, Härkänen T, Knekt P, Virtala E, Oja H (2010a). "Estimation of Population Attributable Fraction (PAF) for Disease Occurrence in a Cohort Study Design." *Statistics in Medicine*, **29**(7-8), 860–874.

Laaksonen MA, Knekt P, Härkänen T, Virtala E, Oja H (2010b). "Estimation of Population Attributable Fraction for Mortality in a Cohort Study Design Using a Piecewise Constant Hazards Model." *American Journal of Epidemiology*, **171**(7), 837–847.

Lehnert-Batar A (2006). ***pARtial: pARtial** Package.* R package version 0.1. URL http://CRAN.R-project.org/src/contrib/Archive/pARtial/.

Mezzetti M, Ferraroni M, Decarli A, La Vecchia C, Benichou J (1996). "Software for Attributable Risk and Confidence Interval Estimation in Case-Control Studies." *Computers and Biomedical Research*, **29**, 63–75.

Rämsch C, Pfahlberg AB, Gefeller O (2009). "Point and Interval Estimation of Partial Attributable Risks from Case-Control Data Using the R-Package '**pARccs**'." *Computer Methods and Programs in Biomedicine*, **94**, 88–95.

Rückinger S, von Kries R, Toschke AM (2009). "An Illustration of and Programs Estimating Attributable Fraction in Large Scale Surveys Considering Multiple Risk Factors." *BMC Medical Research Methodology*, **9**, 7–12.

Samuelsen SO, Eide GE (2008). "Attributable Fractions with Survival Data." *Statistics in Medicine*, **27**(9), 1447–67.

SAS Institute Inc (2010). *SAS OnlineDoc, Version 9.2.* SAS Institute Inc., Cary, NC. URL http://www.sas.com/.

Spiegelman D, Hertzmark E, Wand HC (2007). "Point and Interval Estimates of Partial Population Attributable Risks in Cohort Studies: Examples and Software." *Cancer Causes Control*, **18**(5), 571–9.

Affiliation:

Maarit A. Laaksonen
National Institute for Health and Welfare
Mannerheimintie 166
00300 Helsinki, Finland
E-mail: maarit.laaksonen@thl.fi

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 43, Issue 7 *Submitted:* 2010-02-22
July 2011 *Accepted:* 2011-06-30