AMSI
VACATION
RESEARCH
SCHOLARSHIPS
2018-2019

Estimating population attributable fractions in the presence of competing risks

Ian Powell

Supervised by Dr Maarit Laaksonen and Prof Jake Olivier

February 28, 2019

Vacation Research Scholarships are funded jointly by the Department of Education and Training and the Australian Mathematical Sciences Institute.

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Australian Government
Department of Education and Training

Abstract

The Population Attributable Fraction (PAF) is a highly useful measure in population health. PAFs provide a measure of the impact of a modifiable risk factor on a detrimental outcome at a population level. Sometimes the observation of an outcome of interest may be hindered or otherwise affected by a competing risk, and this needs to be accounted for in calculations of outcome incidence. We present a package for the R programming language which can calculate PAF point estimates and their confidence intervals from cohort study information. This work should make the use of PAFs more widely available for use by policymakers and researchers who desire to quantify the effect of some risk factor at a population level.

1 Introduction

It is vital for policymakers, decision makers and the general public to understand the effect of a particular modifiable risk factor at the population level. This so the policymakers are able to estimate and quantify the effect of changes to current policies on the population of interest. Many studies of a particular risk factor report the relative risk (RR) or odds ratio (OR). These quantities are useful at the individual level. That is, they estimate how much more likely it is for an individual exposed to the deleterious risk factor (for example, smoking) to suffer a negative outcome (for example, lung cancer or early death). However, on their own, relative risks and odds ratios are not always useful in population health because they do not account for the prevalence of a risk factor in a population.

The *Population Attributable Fraction* (PAF) estimates the proportion of outcomes in a population that can be attributed to a particular risk factor. The PAF takes into account both strength of exposure-outcome association and exposure prevalence. A PAF of zero would suggest that none of the outcome burden is attributable to the risk factor of interest, while a PAF of 0.5 would indicate that 50% of the occurrences of the outcome can be attributed to a risk factor. The PAF can be no larger than 1 (which would indicate that all occurrences of the outcome can be attributed to the risk factor), but can also be negative, which would suggest a factor is, in fact, protective.

In this report, we are interested in the estimation of PAFs and their confidence intervals from cohort studies. These PAFs can be calculated for the incidence of death, or the incidence of a particular disease. In either case, we might want to consider and account for potential competing risks, which have been shown to bias PAF estimates (Laaksonen et al., 2010a). For example, death is a competing risk for diabetes, because one cannot contract diabetes postmortem. A more complicated case could be separating out different causes of death – we might be interested in deaths associated with obesity-related causes, in which case death by any other cause (e.g. cancer) is a competing risk.

Presently, there are no packages available in the R programming language (R Core Team, 2018) to calculate PAFs in the presence of these competing risks. We present a preliminary package for R that estimates PAFs and their confidence intervals from cohort studies. Options are available to calculate PAFs over time intervals $(0, t]$ with or without accounting for potential competing risks. Similar macros have been available in SAS for some time (Laaksonen et al., 2011). In Section 2, we present the relevant concepts for the estimation of PAFs in cohort studies. In Section 3, we present the `rpaf` package, which implements the PAF calculations as described. Section 4 contains further avenues of development for this package. Appendix A contains two examples of the use of this package, based on the Mini-Finland Health Survey dataset used by Laaksonen et al. (2011).

2 Estimating PAFs from cohort studies

The PAF is the proportional difference in expected outcome incidence in a population of n individuals. More formally, let X be the matrix with rows given by $\mathbf{x}_i$ for $i = 1, \ldots, n$. That is,

$$X = \left(\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top\right)^\top$$

be a matrix of risk factors and let T_i^M denote the time of death. Then, the probability that the i^{th} individual (with risk factors $\mathbf{x}_i$) dies no later than time t is $\mathbb{P}(T_i^M \leq t \mid \mathbf{x}_i)$, where $i = 1, \ldots, n$.

The outcome incidence is the expected proportion of deaths,

$$I^M(X; t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(T_i^M \leq t \mid \mathbf{x}_i). \tag{1}$$

Now suppose that we are able to modify the risk factors to some target value $X^* = \left(\mathbf{x}_1^{*\top}, \ldots, \mathbf{x}_n^{*\top}\right)^\top$. For example, smokers could become non-smokers, or high BMI patients may modify their behaviour to reduce their BMI to a normal level. Then the modified outcome incidence is the expected proportion of deaths assuming the risk factors have been modified,

$$I^M(X^*;t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(T_i^M \le t \mid \mathbf{x}_i^*).$$

The PAF for mortality is the relative difference in the non-modified and modified outcome incidences,

$$\mathrm{PAF}(T^M \le t) = \frac{I^M(X;t) - I^M(X^*;t)}{I^M(X;t)}. \tag{2}$$

Note that in this formulation, the superscript explicitly affirms that we are interested in an outcome (e.g. mortality) without competing risks.

If we desire to estimate a PAF for disease incidence, then we must also consider potential competing risks, as it is possible that an individual did not contract the disease during follow up simply because they died before they could contract the disease from shared risk factors. The definition for outcome incidence can be reformulated to account for competing risks,

$$I^D(\mathbf{x};t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}(T_i^D \le \min\left\{T_i^M, t\right\} \mid \mathbf{x}_i) \tag{3}$$

where T^D is the time of incidence of disease and TˆM is again time to death. As before, we can replace X by X^* to compute the modified outcome incidence, and similarly define the PAF for disease incidence as

$$\mathrm{PAF}(T^D \le \min(T^M, t)) = \frac{I^D(X;t) - I^D(X^*;t)}{I^D(X;t)}. \tag{4}$$

In both cases with and without competing risks, we also need to consider that right-censoring may still occur. That is, the study ends at some time and it is possible that the outcome of interest has not been observed. This time of censoring is denoted T_i^C, where $i = 1, \ldots, n$. Note that the censoring times are not necessarily equal for all individuals. If the outcome of interest is death, we may either observe $T_i^M = \min\left\{T_i^M, T_i^C\right\}$ (where death occurs before censoring) or $T_i^C = \min\left\{T_i^M, T_i^C\right\}$ (where censoring occurs before death). On the other hand, if the outcome of interest is disease incidence, we can observe $T_i^C = \min\left\{T_i^D, T_i^M, T_i^C\right\}$ (where neither death nor disease occur during follow-up), $T_i^M = \min\left\{T_i^D, T_i^M, T_i^C\right\}$ (where death but not disease occurs during follow-up), $T_i^D < T_i^C = \min\left\{T_i^M, T_i^C\right\}$ (where disease, but not death, occurs before censoring), or $T_i^D < T_i^M = \min\left\{T_i^M, T_i^C\right\}$ (where both disease and death occur before censoring).

2.1 Model assumptions

In order to calculate the probabilities in Equations 1 and 3, we need to develop a model for the times to event T^D and T^M. We shall assume that T_i^D and T_i^M are conditionally independent given risk factors $\mathbf{x}_i$ for all $i = 1, \ldots, n$. This is equivalent to assuming that our model considers all risk factors that affect both mortality and disease. We shall also assume that the time to censoring T_i^C is independent of both T_i^M and T_i^D.

One method for modelling times to event T^D and T^M is the piecewise constant hazards model. In this model, we partition the study follow up time into J intervals with breaks at $a_0, a_1, \ldots, a_J$, and the hazard of the event occurring for each individual is constant within each interval. We write that $(a_{j-1}, a_j]$ is the j^{th} follow-up period. We can write the hazard function for the i^{th} individual as

$$h(t; \mathbf{x}_i) = \sum_{j=1}^{J} \lambda_{ij} \mathbb{1}_{(a_{j-1}, a_j]}(t) \tag{5}$$

where $\mathbb{1}_{(a_{j-1}, a_j]}$ is the indicator function which is 1 if t is in the j^{th} time period and 0 otherwise and λ_{ij} is the constant hazard for the i^{th} individual in the j^{th} follow-up period. Then, the time to event T_i for the i^{th} individual in the j^{th} follow-up period is partitioned as,

$$T_{ij} = \begin{cases} a_j - a_{j-1} & \text{if } T_i \geq a_j \\ T_i - a_{j-1} & \text{if } a_{j-1} < T_i \leq a_j \\ \text{undefined} & \text{if } T_i < a_{j-1}. \end{cases}$$

Note that, in the first of these cases, censoring for the current time period occurs before the event of interest, though the event may occur in a later time period; in the second, the event occurs during the current time period; and in the third, the event has already occurred.

The main advantage of this model is that it is both parametric and can flexibly accommodate any hazard functions. Parametric models are particularly useful in survival analysis because they allow us to estimate variances of the regression coefficients, which is necessary to report confidence intervals along with our PAF point estimates. Furthermore, we can approximate any hazard function to arbitrary precision if we have enough data for J to be large. Using this approximation, we can estimate the survival function $S(t; \mathbf{x})$, which is the probability that the outcome occurs later than time t. That is, $S(t; \mathbf{x}) = \mathbb{P}(T > t \mid \mathbf{x}_i)$. Then we can calculate the survival from the hazard function using the ordinary differential equation

$$-\frac{S'(t; \mathbf{x})}{S(t; \mathbf{x})} = h(t; \mathbf{x}), \qquad S(0; \mathbf{x}) = 1. \tag{6}$$

We shall also assume that we can write the constant hazards λ_{ij} as a product of a baseline hazard (which depends on age and time) and an individual hazard (which depends only on the individual's risk factors). That is, if we divide the range of individual dates of birth into B cohorts $(v_0, v_1], \ldots, (v_{B-1}, v_B]$. Denote the birth cohort of the i^{th} individual as b_i, and note that these are not necessarily unique. Then, we shall assume that $\lambda_{ij} = \lambda_{0jb_i}\lambda_i$, where λ_{0jb_i} is the baseline hazard for an individual in birth cohort b_i in time period j, and λ_i is the hazard based on risk factors $\mathbf{x}_i$.

2.2 Survival analysis

We fit a survival regression to model our T_{ij}. The rule for this regression is

$$\log T_{ij} = \alpha_{jb_i} + \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_{ij}, \qquad \alpha_{jb_i} = -\log \lambda_{0jb_i}.$$

Here, $\mathbf{x}_i$ is a vector of covariates corresponding to the risk factors of interest with corresponding coefficients $\boldsymbol{\beta}$. The random errors ϵ_{ij} are taken from an extreme value distribution with density $f_\epsilon(t) = \exp(t - e^t)$. It can be shown that $T_{ij} \sim \text{Exp}(\lambda_{ij})$ (Kalbfleisch, 2002), where $\lambda_{ij} = \lambda_{0jb_i} e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}$ and so by our assumptions, $\lambda_i = e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}$. Note that we wish to estimate both α_{jb_i} for all $j = 1, \ldots, J$ and all birth cohorts $b_i = 1, \ldots, B$, and $\boldsymbol{\beta}$. This can be accomplished by reformulating this problem using a single model matrix $Z = \left(\mathbf{z}_{11}^\top, \mathbf{z}_{12}^\top, \ldots, \mathbf{z}_{nJ}^\top\right)^\top$ which is a combinatorial merge of indicators for risk factors and indicators for birth cohort and time period. Then the coefficient vector becomes $\boldsymbol{\gamma} = (\alpha_{11}, \ldots, \alpha_{JB}, \beta_1, \ldots, \beta_p)$ (Laaksonen et al., 2011), for which we can find a maximum likelihood estimate using iterative techniques.

When calculating PAFs for mortality, the time to event is simply T_{ij}^M and we estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the usual manner. However, when calculating PAF for disease, we censor on the first event of interest; that is, if an individual contracts the disease at time T^D, then we censor the individual's time of mortality T^M in regression calculations (Laaksonen et al., 2011). Thus when calculating disease PAFs, we indeed perform two survival regression analyses to estimate $\boldsymbol{\alpha}^M, \boldsymbol{\beta}^M$ and $\boldsymbol{\alpha}^D, \boldsymbol{\beta}^D$ separately. Note that, as T_i^D and T_i^M are conditionally independent given our risk factors, these two sets of regression coefficients are assumed to be independent.

2.3 Incidence and gradient calculations

When we are interested in the time to death from any cause T^M in the absence of any competing risks, then the hazard for the i^{th} individual in the j^{th} follow-up period is given by $\lambda_{ij}^M = \exp(-\alpha_{jb_i} - \mathbf{x}_i^\top \boldsymbol{\beta})$. Then, we can substitute the relation in Equation 5 to solve the ODE in Equation 6 to recover the probability of survival of the i^{th} individual at the end of the j^{th} period

$$S_{ij}^M = S^M(a_j; \mathbf{x}_i) = \exp\left(-\sum_{k=1}^{j} \lambda_{ik}^M (a_k - a_{k-1})\right). \tag{7}$$

The mortality incidence given model matrix X is then just the proportion of subjects who have not survived until the end of the j^{th} period:

$$I_j^M(X) = I^M(X; a_j) = \frac{1}{n} \sum_{i=1}^{n} \left(1 - S_{ij}^M\right). \tag{8}$$

Confidence interval calculations require us to calculate the gradients of S^M and I^M with respect to the regression coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. If we consider the formulation briefly introduced in Section 2.2 with hazards $\lambda_{ij} = \exp\left(-\alpha_{jb_i} - \mathbf{x}_i^\top \boldsymbol{\beta}\right) = \exp\left(-\mathbf{z}_{ij}^\top \boldsymbol{\gamma}\right)$, then the derivatives become (with reference to Equations 7 and 8)

$$\begin{aligned} \nabla \lambda_{ij}^M &= \lambda_{ij}^M \mathbf{z}_{ij} \\ \nabla S_{ij}^M &= -S_{ij}^M \sum_{k=1}^{j} (a_k - a_{k-1}) \nabla \lambda_{ik}^M \\ \nabla I_j^M(X) &= -\frac{1}{n} \sum_{i=1}^{n} \nabla S_{ij}^M. \end{aligned} \tag{9}$$

where the gradients denoted by ∇ are with respect to the regression coefficient vector $\boldsymbol{\gamma}$. A more detailed description of this process is presented by Laaksonen et al. (2010b, Appendix 2).

When we are interested in the time to event when there is the possibility of competing risk, we now have two sets of regression coefficients: $\boldsymbol{\alpha}^M, \boldsymbol{\beta}^M$ and $\boldsymbol{\alpha}^D, \boldsymbol{\beta}^D$. As for the case with competing risks, we calculate λ_{ij}^M and λ_{ij}^D as before, and then use Equation 7 to calculate both S_{ij}^M and, with appropriate alteration, S_{ij}^D. From here, we introduce the disease-free survival

$$S_{ij} = S_{ij}^M S_{ij}^D.$$

Now, the probability that disease is the first event to occur (rather than mortality) is

$$\mathbb{P}\left(\min(T_i^D, T_i^M) = T_i^D \mid \min(T_i^D, T_i^M) > t\right) = \frac{h^D(t;\mathbf{x}_i)}{h^D(t;\mathbf{x}_i) + h^M(t;\mathbf{x}_i)}$$

due to conditional independence. Therefore, the incidence of disease in our population at the end of our j^{th} period is expected to be

$$\begin{aligned} I_j^D(X) &= I^D(X; a_j) \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^j \mathbb{P}\left(T_i^{min} = T_i^D \mid a_{k-1} < T_i^{min} \le a_k\right) \mathbb{P}\left(a_{k-1} < T_i^{min} \le a_k\right) \\ &= \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^k \frac{\lambda_{ij}^D}{\lambda_{ij}^D + \lambda_{ij}^M}\left(S_{i,k-1} - S_{ik}\right). \end{aligned}$$

where $T_i^{min} = \min(T_i^D, T_i^M)$.

To calculate the gradients, we note that we now have two coefficient vectors, termed $\boldsymbol{\gamma}^D$ and $\boldsymbol{\gamma}^M$. We use ∇_D and ∇_M to denote the gradients with respect to these vectors respectively. Then, as with Equation 9, we have $\nabla_D \lambda_{ij}^D = \lambda_{ij}^D \mathbf{z}_{ij}$ and $\nabla_M \lambda_{ij}^M = \lambda_{ij}^M \mathbf{z}_{ij}$, and $\nabla_D S_{ij}^D = -S_{ij}^D \sum_{k=1}^j (a_k - a_{k-1}) \nabla \lambda_{ik}^D$ and $\nabla_M S_{ij}^M = -S_{ij}^M \sum_{k=1}^j (a_k - a_{k-1}) \nabla \lambda_{ik}^M$. Noting that S_{ij}^D is independent of $\boldsymbol{\gamma}^M$ and similarly S_{ij}^M is independent of $\boldsymbol{\gamma}^D$, we note that

$$\begin{aligned} \nabla_D S_{ij} &= S_{ij}^M \nabla_D S_{ij}^D \\ \nabla_M S_{ij} &= S_{ij}^D \nabla_M S_{ij}^M. \end{aligned}$$

One can use repeated application of the chain rule from here, as we have done previously, to calculate the gradients of $I_j^D(X)$ (Laaksonen et al., 2010a).

Using the incidence calculations, either with or without competing risks, we can now use Equations 2 and 4 to calculate the PAFs. That is, we follow the basic formula

$$\text{PAF}(T \le a_j) = \frac{I(X; a_j) - I(X^*; a_j)}{I(X; a_j)}$$

where I can be I^M or I^D as required, and has been calculated for our cohort with risk factors X, and the idealised cohort with risk factors X^*.

2.4 PAF confidence intervals

Confidence intervals for PAFs can be derived using the delta method. However, by its definition, the PAF lies in the range $(-\infty, 1]$, and so cannot be normally distributed. Instead, we perform the delta method with a transformation $g(\text{PAF}) = \log(1 - \text{PAF})$. Our estimate for the PAF then approaches a normal distribution asymptotically:

$$\sqrt{n}\left(g(\widehat{\text{PAF}}) - g(PAF)\right) \xrightarrow{d} N\left(0, \sigma^2_{g(PAF)}\right)$$

where we can consistently estimate $\sigma^2_{g(PAF)}$ by

$$\sigma^2_{g(PAF)} = \begin{cases} \left(\nabla g(PAF)\right)^\top \widehat{\text{Var}\boldsymbol{\gamma}} \left(\nabla g(PAF)\right) & \text{without competing risks} \\ \\ \left(\nabla_D g(PAF)\right)^\top \widehat{\text{Var}\boldsymbol{\gamma}^D} \left(\nabla_D g(PAF)\right) & \\ \quad + \left(\nabla_M g(PAF)\right)^\top \widehat{\text{Var}\boldsymbol{\gamma}^M} \left(\nabla_M g(PAF)\right) & \text{with competing risks.} \end{cases}$$

One can show that the gradients of the transformed PAFs are

$$\nabla g(PAF) = \frac{\nabla I(X^*)}{I(X^*)} - \frac{\nabla I(X)}{I(X)}$$

so we can use the gradient calculations from the previous sections to calculate this gradients (where we can replace ∇ by ∇_D or ∇_M and make other substitutions as appropriate).

3 The `rpaf` package

The `rpaf` is a package written for the R programming language to calculate PAFs. It provides options to estimate PAFs with and without consideration of potential competing risks. It is designed to be simple to use, requiring only two function calls to generate and display an analysis of the data. Each of these function calls performs a different step in PAF analysis. The `gen_data()` function performs data preparation, converting wide-form cohort study data into corresponding time-varying data, as required for the analysis. Next, the user can call `mpaf_summary()` or `dpaf_summary()` to obtain the PAF estimates of interest (without or with consideration of potential competing risks, respectively).

3.1 Data preparation – `gen_data()`

The first step in PAF calculations, whether or not they are calculated in the presence of competing risks, is converting the original data to a time-varying format. In the `rpaf` package, data preparation is performed by the `gen_data()` function, which has the following signature:

```
rpaf::gen_data(indata, id_var, ft_breaks,
               death_ind, death_time,
               disease_ind, disease_time,
               variables,
               na.action,
               period_factor, time_var)
```

- `indata` is a data frame containing the information from the cohort study in wide format.
- `id_var` is the column of `indata` that contains the unique individual ID number. It accepts a character string as an argument. In later versions, we may support automatic creation of such a column.
- `ft_breaks` is a numeric vector containing the breaks between periods of constant hazard. For example, if we wanted time periods of $(0, 2], (2, 4]$ and $(4, 5]$, we would specify `ft_breaks = c(0, 2, 4, 5)`. If many regular time periods are desired, then base R's `seq()` function can be used. For example, to specify time periods of $(0, 0.5], (0.5, 1.0], \ldots, (4.5, 5.0]$, we would use `ft_breaks = seq(0, 5, 0.5)`.
- `death_ind` and `death_time` are the names of the columns for indicator for death, and time until follow up for death, respectively. There are both specified as character strings. The `death_ind` column is assumed to be `TRUE` if death occurs before censoring, and `FALSE` otherwise. These two arguments are optional, and a dataset of only predictors is returned if missing.
- `disease_ind` and `disease_time` are as for `death_ind` and `death_time`, except for incidence of disease instead of death. If these two arguments are not specified, then the dataset returned will be for death in the absence of competing risks.
- `variables` indicates which predictors need to be included in the new data. This takes a character vector that contains the names relevant columns from `indata`.

- `na.action` specifies the treatment of missing values, and is described in more detail in the R help page for `na.action`.
- `period_factor` and `time_var` are the names of the new columns that are produced in the prepared dataset. These columns contain the time period (in factor form), and the time of follow-up for each period respectively.

The `gen_data()` function outputs an R list[1] containing the following elements.

- `data_call` is the function call used to generate this data item.
- `nobs` is the number of observations used from the initial data frame.
- `na.action` stores the ID numbers of individuals removed from the data due to missingness
- `breaks` is a numeric vector identical to the `breaks` argument supplied, used in downstream functions.
- `ID` are the identification numbers corresponding to rows of the data frame. This is necessary as the data for each individual spans more than one row, so we can use `ID` to group the data by subject.
- `PERIOD` is factor vector of periods corresponding to the rows of the data frame. Similarly to `ID`, this can be used to group the data by follow-up time period.
- `data` is the data frame containing all information required to fit the survival regression models.

Examples of the use of the `gen_data()` function can be seen in Appendix A, which demonstrates how this function is used in two cases: if outcome of interest alone is specified along with a competing risk, and if only the outcome of interest is specified. Before the `gen_data()` call, the data would have a form similar to that of Table 1, and the corresponding output data is shown in Table 2.

3.2 Summary functions – `mpaf_summary()` and `dpaf_summary()`

Once we have converted our data to a time-varying format, we can use summary functions to calculate the hazard ratios, PAFs, and groupwise PAF differences. If our model does not account

[1] Contrary to SAS macros, R functions are not designed to modify their arguments, and further can only return a single object. Thus, `rpaf` returns a single list that contains multiple objects that can be passed on to downstream functions.

for any competing risks, we use the `mpaf_summary()` function; otherwise, the `dpaf_summary()` function should be used.

In both cases, the summary functions first call a model-fitting function. This function, `est_matrix()`, fits a survival regression using the `survreg()` function from the `survival` package (Therneau, 2015). It also performs a list of specified modifications to the data and returns data frames of both unmodified and modified prevalences. This function also naively reports hazard ratios as $\exp(-\beta_p)$ for the p^{th} parameter.[2] Unlike the `mpaf_summary()` function, `dpaf_summary()` fits two models using `est_matrix()`; one of these models is for the outcome of interest, while the other is for the competing risk.

Following this, each summary function calls a function to estimate the PAFs and their confidence intervals. This is done via `mpaf_est_paf()` and `dpaf_est_paf()` for `mpaf_summary()` and `dpaf_summary()` respectively. These functions take the fitted models from `est_matrix()` and use the prevalence data to calculate the PAFs, reporting their point estimates, standard errors in transformed space, and their gradients.

Finally, if groupwise estimates are desired, the summary functions split the original data into subgroups and calculate PAFs separately for each subgroup. It then uses the gradients reported by the PAF calculation functions to calculate variances via the delta method and reports groupwise differences and their p-values.

The `mpaf_summary()` function should be used after `gen_data()` is called, and has the following signature:

```
rpaf::mpaf_summary(sr_formula, mpaf_data, modifications, covar_model,
                   prevalence_data, group_vars, level, ...)
```

- `sr_formula` is the model description that is passed to `survival::survreg()`. It takes an R formula as an argument (with response defined on the left hand side and predictors given on the right), though requires that the left hand side is a `survival::Surv()` object to be compatible.

- `mpaf_data` is the output data from `gen_data()`. In this case, `gen_data()` must have been

[2] This is naive in the sense that hazard ratios are not so simply defined when there are interaction terms present, but the `rpaf` package does not yet account for this.

called with the death_ind and death_time arguments specified, else there is no response to be used.

- modifications is a named list which describes the modifications $X \to X^*$ to be made. The names of each list element represent the columns to be modified. Each list element is a vector, often (but not necessarily) a character vector. The first item in this vector is the preferred risk factor level. If no subsequent items are given (i.e. the vector has length one), then all entries will be modified to take the first element. On the other hand, if further items are present in the vector, then all items *after the first* represent the items that should be changed. So, suppose we want our modified population to all have a BMI under 25 kg/m^2. Then, we might see that the corresponding column for BMI is BMI_2 and the desired level is "<25.0". Now suppose further that we want our modified population to have no current smokers. The column for smoking status might be called SMOKE, and it might have levels Never, Former, <30/day and >=30/day. Then in the SMOKE column we want to replace every instance of <30/day or >=30/day with Never. The modifications argument to perform both of these modifications simultaneously is

```
...
  modifications = list(BMI_2 = "<25.0",
                       SMOKE = c("Never", "<30/day", ">=30/day"))
...
```

- covar_model specifies the factors for which hazard ratios are reported. This is a character vector which may contain any of the predictors supplied in sr_formula, with interactions specified with the : separator.

- prevalence_data is an optional new data object as output by gen_data() (though optionally with no response columns), compatible with mpaf_data, that contains different information on the same predictors with with we calculate the PAFs. This allows for a different data frames to be used for model fitting and PAF calculation.

- group_vars is an optional character vector of column names by which to group PAF estimates. That is, for each combination of levels that occur in the specified group variables, a separate PAF estimate will be calculated, and then differences between groupwise PAF estimates will be estimated and tested.

- `level` simply represents the level of interest desired for the confidence intervals of the hazard ratios reported.
- The final argument, `...`, indicates that further arguments can be passed; these arguments will be passed on to `survival::survreg()` for further control of the survival regression (e.g. decreasing tolerance or increasing maximum number of iterations). These options are detailed in the help pages for the `survival` package (Therneau, 2015).

The resulting output can then be passed to the R's generic `print()` function, where it will provide a summary of the calculations to the console. Interrogating the structure of the summary object reveals that it contains a significant amount of information. The main information of interest is the hazard ratios, total PAF and groupwise PAF estimates. However, it contains diagnostic information on the fitted survival regression, missing data, number of observations and the function calls used to produce this output.

The function signature if competing risks are present is similar:

```
rpaf::dpaf_summary(disease_resp, death_resp, predictors,
                   dpaf_data, modifications, covar_model,
                   prevalence_data, group_vars, level, ...)
```

The only difference in this signature is the formula arguments. Because disease PAFs require two models to be fitted, a formula for each has to be specified. This is achieved using three R formula objects:

- `disease_resp`, the response for the disease (i.e. primary outcome of interest) model. This should take the form `survival::Surv(disease_time, disease_ind) ~ .` (note the period in the RHS). The form for `death_resp` is the same, except using time and indicator for death instead of disease.
- `predictors` is a formula of the predictors which is common to both models. This takes the form `~ pred1 + pred2 + ... + predn` (note the empty LHS).

All other variables are the same as for `mpaf_summary()`.

This function returns a similar summary list to `mpaf_summary()`, however as the model fitting function has been called twice, some of the names are duplicated. To prevent naming collisions, information from the `est_matrix()` called are appended with `_d` for the outcome of interest (disease) and `_m` for the competing risk (mortality).

Examples of the use of both of these functions are presented in the second and third examples in Appendix A.

4 Conclusions and further work

There is a single program currently to estimate PAFs in the presence of competing risks, which is implemented using SAS (Laaksonen et al., 2011). However, no packages are available in R, which is free and open source. In this report, we present the first R package that is capable of calculating PAF estimates and their approximate 95% confidence intervals from cohort study data. This package should, in due course, make the calculation of this complex statistic more accessible to researchers and policymakers from a variety of fields.

However, extra work should be conducted both on this package, and in other areas of PAF theory. In particular, we are interested in using external prevalence estimates for the predictive PAF calculations. This would allow us to expand our predictions beyond our original cohort population by assuming that the strength of association remains constant. Then taking the prevalence estimates from another country, or from a later date (both of which could be significantly different to our original cohort) would allow us to estimate the PAFs for these new populations. For example, with recent public health inititatives to reduce smoking it would make sense for Finland in the 1980s and Australia in the 2010s to have different prevalences of smoking.

One other limitation of PAFs is that they make an assumption of immediate risk reduction. With the current state of the art, PAFs cannot be used to determine the reduction in burden if the entire population ceased a detrimental behaviour; instead they can only estimate the impact if that behaviour was never adopted. Thus PAFs tend to overestimate the reduction in burden unless gradual risk reduction is considered. (The British Doctors' Study has shown, for example, that ceasing smoking leads to a rapid but not immediate reduction in mortality hazard (Doll et al., 2004).) Therefore one other avenue for research is developing estimation methods for PAFs which account for gradual risk reduction.

Appendix A – Data examples

The data used in this example are an excerpt from the Mini-Finland Health Survey as presented by Laaksonen et al. (2011) to showcase their SAS macros for PAF calculation. The dataset, available in `rpaf::minifhs`, consists of 4517 observations across 14 variables:

- `ID` is the unique individual identification number.
- `DEATH` is a logical which is `TRUE` if the subject died during follow up.
- `DEATH_FT` is the length of time in years until death (if `DEATH` is `TRUE`) or censoring (if `DEATH` is `FALSE`) occurs.
- `DIAB` and `DIAB_FT` are as for `DEATH` and `DEATH_FT`, except for the incidence of type II diabetes instead of death.
- `BYEAR` is the year of birth of the subject
- `B_COHORT` is a categorical variable identifying the "birth cohort" by separating `BYEAR` into 3 blocks of 20 years (i.e. 1890-1909, 1910-1929, 1930-1949).
- `AGE` is the age in years at baseline.
- `AGEGRP` is a categorical variable identifying age group in blocks of 10 years at baseline.
- `SEX` is the sex of the participant.
- `BMI` is the body mass index in kg/m^2.
- `BMI_2` is an indicator for high BMI ($\geq 25.0\text{kg/m}^2$) or lower BMI ($< 25.0\text{kg/m}^2$).
- `BP` is an indicator for normal or elevated blood pressure at baseline.
- `SMOKE` is smoking status at baseline, separated into four groups: never smoked; formerly smoked; smoke fewer than 30 cigarettes per day or pipe/cigar smokers only; or smoking at least 30 cigarettes per day.

Of the 4517 rows, 4507 have no missing data. Missing data only occur in columns relating to BMI, blood pressure, and smoking status.

Effect of smoking on mortality

As a basic example of a PAF calculation, we are interested in the effect of smoking on total mortality. Our risk factors of interest are sex (which is non-modifiable) and smoking (which is modifiable). Our idealised population X^* will have no current smokers; instead, current smokers will be treated as if they had never smoked, with former smokers unaffected. We will partition our follow-up time of 17 years into partitions of 5 years $(0, 5], (5, 10], (10, 15]$ and one remaining two-year period $(15, 17]$. The baseline hazard will therefore be calculated from the interaction of twenty-year birth cohort and follow-up period.

Our first step in calculating PAFs is to prepare the data. We use the `gen_data()` function to convert our data frame into time-varying format. A sample of the original `minifhs` data frame is shown in Table 1. To convert it to the time-varying format, we use the following function call:[3]

```
smoke_data <- gen_data(
  minifhs, id_var = "ID", ft_breaks = c(0,5,10,15,17),
  death_ind = "DEATH", death_time = "DEATH_FT",
  variables = c("B_COHORT", "SEX", "SMOKE"),
  period_factor = "F_PERIOD", time_var = "FT_END"
)
```

The new data frame that this function produces can be seen in Table 2. From here, we can fit a survival model using `est_matrix()`. We recall we want our baseline to be dependent upon the interaction of birth cohort and current period, so in the formula interface we include a cross term `B_COHORT * F_PERIOD`. The model fitting call then becomes:

```
smoke_fit <- est_matrix(
  survival::Surv(FT_END, DEATH) ~ B_COHORT * F_PERIOD + SEX + SMOKE,
  paf_response = smoke_data,
  modifications = list(SMOKE = c("Never", "<30/day", ">=30/day")),
```

[3] In this example, and following examples, we have loaded the `rpaf` library and so are not required to specify the `rpaf` namespace using the `::` operator.

ID	B_COHORT	SEX	SMOKE	DEATH	DEATH_FT
9	1910	Female	<30/day	FALSE	16.95
88	1930	Male	Never	FALSE	16.92
100	1890	Female	Never	TRUE	14.92

Table 1: Sample of original data passed to `gen_data()`. This shows some risk factors and outcomes associated with three individuals: birth cohort (B_COHORT), indicating the start of twenty-year period in which the individual was born; sex (SEX), male or female; smoking status (SMOKE), one of never, former, less than 30 cigarettes per day (pipe or cigar smokers are included in this category) or at least 30 cigarettes per day. The DEATH column indicates if the individual died before the end of follow-up and the DEATH_FT column indicates the length of follow-up in years until death or censoring. Note that some columns from the original dataset columns have been omitted for brevity.

```
  covar_model = c("SEX", "SMOKE")
)
```

Note that we are satisfied with the default confidence level of 95%, so we omit the `level` argument. The `modifications` argument follows the structure outlined in Section ??, which specifies that any value in `SMOKE` taking values `<30/day` or `>=30/day` should be modified to `Never` in the idealised data frame. From this output, we are able to view the hazard ratios, which are presented in Table 3.

Having now fitted the model, we can calculate PAF estimates using `mpaf_est_paf()` as follows:

```
smoke_paf <- mpaf_est_paf(smoke_fit, smoke_data)
```

We are not interested in any new prevalence data, so the `newdata` argument is omitted. Again, we are again satisfied with the default 95% confidence interval, and so omit the `level` argument. The PAFs and their confidence intervals over each time period can then be viewed from `smoke_paf$paf0` (Table 4). This shows that, according to our model, approximately 10.7% (95% CI: 0.082, 0.130) of deaths in this cohort over the course of 17 years could have been avoided if the current smokers in the population never took up the habit.

Note that these calculations, after data preparation, could have been conducted using the `mpaf_summary()` function. The similar `dpaf_summary()` function is exemplified in the next example.

Effect of BMI on diabetes, stratified by sex

As a complete example of all the work that `rpaf` can do, we now want to estimate the PAF for the effect of BMI on diabetes. We also want to calculate the PAFs for each subgroup of the cohort by sex, and test if there is a significant difference between these PAFs. Our risk factors of interest are sex (which is non-modifiable) and a high/low BMI indicator (which is modifiable). We keep the same time periods as in our previous example. Therefore the data preparation call becomes:

```r
bmi_data <- gen_data(
  minifhs, id_var = "ID", ft_breaks = c(0,5,10,15,17),
  death_ind = "DEATH", death_time = "DEATH_FT",
  disease_ind = "DIAB", disease_time = "DIAB_FT",
  variables = c("B_COHORT", "SEX", "BMI_2"),
  period_factor = "F_PERIOD", time_var = "FT_END"
)
```

Now we can do all the calculations we desire using the `dpaf_summary()` function as below. Note that because we are open to the possibility that sex is a confounding factor, we must include the interaction between BMI and sex in our predictor model.

```r
bmi_summ <- dpaf_summary(
  disease_resp = survival::Surv(FT_END, DIAB) ~ .,
  death_resp = survival::Surv(FT_END, DEATH) ~ .,
  predictors = ~ B_COHORT * F_PERIOD + SEX + BMI_2:SEX,
  dpaf_data = bmi_data,
  modifications = list(BMI_2 = "<25.0"),
  covar_model = c("SEX", "SEX:BMI_2"), # weird string black magic requirement
  group_vars = "SEX"
)
```

The raw printed output for `bmi_summ` is shown below. In particular we can see that none of the differences between groupwise PAFs are significant in any time period. Thus we can conclude that BMI affects the incidence of diabetes in males and females equally much on a population scale ($p \gg 0.05$ in all cases; indeed that smallest is roughly $p = 0.84$). In total, we can estimate that 68% (95% CI: 0.550, 0.773) of diabetes cases in this cohort over 17 years would have been avoided if the cohort always had a BMI lower than 25.0 kg/m^2. Indeed, this number changes very little over intervening time periods.

Summary output

```
##
## Call:
## dpaf_summary(disease_resp = survival::Surv(FT_END, DIAB) ~ .,
##     death_resp = survival::Surv(FT_END, DEATH) ~ ., predictors = ~B_COHORT *
##         F_PERIOD + SEX + BMI_2:SEX, dpaf_data = bmi_data, modifications = list(BMI_2
##     covar_model = c("SEX", "SEX:BMI_2"), group_vars = "SEX")
## --------------------------------------------------------
## Survival regression summary:
##
##     Disease:
##
## Call:
## survival::survreg(formula = sr_formula, data = paf_response$data,
##     dist = "exponential", y = FALSE)
##                                    Value Std. Error      z        p
## (Intercept)                      6.71720    0.42893 15.660  < 2e-16
## B_COHORT1910                     0.35254    0.38177  0.923   0.3558
## B_COHORT1930                     1.06174    0.48702  2.180   0.0293
## F_PERIOD(5,10]               -0.75694    0.42732 -1.771   0.0765
## F_PERIOD(10,15]              -0.82104    0.47165 -1.741   0.0817
## F_PERIOD(15,17]              -1.53330    0.78196 -1.961   0.0499
## SEXFemale                     0.19685    0.37905  0.519   0.6035
## B_COHORT1910:F_PERIOD(5,10]  -0.13777    0.48213 -0.286   0.7751
## B_COHORT1930:F_PERIOD(5,10]  -0.19083    0.59799 -0.319   0.7496
## B_COHORT1910:F_PERIOD(10,15]  0.09179    0.52695  0.174   0.8617
## B_COHORT1930:F_PERIOD(10,15]  0.10285    0.64350  0.160   0.8730
## B_COHORT1910:F_PERIOD(15,17]  1.16658    0.91966  1.268   0.2046
## B_COHORT1930:F_PERIOD(15,17]  0.56198    1.03432  0.543   0.5869
## SEXMale:BMI_2>=25.0          -1.47591    0.29772 -4.957 7.15e-07
## SEXFemale:BMI_2>=25.0        -1.51875    0.27569 -5.509 3.61e-08
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -1430.4   Loglik(intercept only)= -1500.3
##  Chisq= 139.73 on 14 degrees of freedom, p= 8e-23
## Number of Newton-Raphson Iterations: 9
## n=15277 (2779 observations deleted due to missingness)
##
##
##     Mortality:
##
## Call:
## survival::survreg(formula = sr_formula, data = paf_response$data,
##     dist = "exponential", y = FALSE)
##                                  Value Std. Error      z        p
## (Intercept)                    2.69338    0.10887 24.740  < 2e-16
## B_COHORT1910                   1.35934    0.13149 10.338  < 2e-16
## B_COHORT1930                   2.59545    0.22220 11.680  < 2e-16
## F_PERIOD(5,10]                -0.65275    0.12678 -5.149 2.63e-07
## F_PERIOD(10,15]               -1.23277    0.12698 -9.709  < 2e-16
## F_PERIOD(15,17]               -1.39069    0.24895 -5.586 2.32e-08
## SEXFemale                      0.72092    0.09214  7.824 5.10e-15
## B_COHORT1910:F_PERIOD(5,10]    0.31983    0.17435  1.834  0.06659
## B_COHORT1930:F_PERIOD(5,10]    0.17469    0.28586  0.611  0.54112
## B_COHORT1910:F_PERIOD(10,15]  0.34597    0.16822  2.057  0.03972
## B_COHORT1930:F_PERIOD(10,15]  0.55894    0.27989  1.997  0.04582
## B_COHORT1910:F_PERIOD(15,17]  0.33751    0.30526  1.106  0.26888
## B_COHORT1930:F_PERIOD(15,17]  0.05800    0.41641  0.139  0.88922
## SEXMale:BMI_2>=25.0            0.24503    0.08294  2.955  0.00313
## SEXFemale:BMI_2>=25.0          0.17003    0.08970  1.896  0.05801
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -5003.8   Loglik(intercept only)= -5566.4
##  Chisq= 1125.18 on 14 degrees of freedom, p= 2.1e-231
## Number of Newton-Raphson Iterations: 8
## n=15277 (2779 observations deleted due to missingness)
##
## --------------------------------------------------------
## Cohort size: 4514 (3 observations deleted due to missingness)
## --------------------------------------------------------
## Hazard ratios:
##
##     Disease:
##                         Hazard ratio  2.5 % 97.5 %
## SEXMale                      1.0000     --     --
## SEXFemale                   0.8213 0.3907  1.726
## SEXMale:BMI_2<25.0          1.0000     --     --
## SEXFemale:BMI_2<25.0        1.0000     --     --
## SEXMale:BMI_2>=25.0         4.3750 2.4410  7.842
## SEXFemale:BMI_2>=25.0       4.5665 2.6602  7.839
##
##     Mortality:
##                       Hazard ratio  2.5 % 97.5 %
## SEXMale                     1.0000     --     --
## SEXFemale                   0.4863 0.4060 0.5826
## SEXMale:BMI_2<25.0          1.0000     --     --
## SEXFemale:BMI_2<25.0        1.0000     --     --
## SEXMale:BMI_2>=25.0         0.7827 0.6653 0.9208
## SEXFemale:BMI_2>=25.0       0.8436 0.7076 1.0058
## --------------------------------------------------------
## Modifications:
## BMI_2: . -> <25.0
## --------------------------------------------------------
## PAFs for disease:
##            PAF  2.5 % 97.5 %
## (0,5]   0.6811 0.5490 0.7745
## (5,10]  0.6803 0.5493 0.7732
## (10,15] 0.6803 0.5496 0.7731
## (15,17] 0.6791 0.5463 0.7730
## (0,5]   0.6811 0.5490 0.7745
## (0,10]  0.6805 0.5493 0.7735
## (0,15]  0.6805 0.5496 0.7733
## (0,17]  0.6803 0.5496 0.7731
## --------------------------------------------------------
## Groupwise PAF estimates:
```

```
##
##      SEX: Male
##      ---------
##             PAF  2.5 % 97.5 %
## (0,5]    0.6685 0.4516 0.7997
## (5,10]   0.6708 0.4570 0.8004
## (10,15]  0.6735 0.4618 0.8019
## (15,17]  0.6757 0.4650 0.8034
## (0,5]    0.6685 0.4516 0.7997
## (0,10]   0.6701 0.4554 0.8002
## (0,15]   0.6712 0.4577 0.8006
## (0,17]   0.6716 0.4586 0.8008
##
##      SEX: Female
##      -----------
##             PAF  2.5 % 97.5 %
## (0,5]    0.6914 0.5028 0.8085
## (5,10]   0.6878 0.5000 0.8051
## (10,15]  0.6854 0.4974 0.8032
## (15,17]  0.6814 0.4904 0.8008
## (0,5]    0.6914 0.5028 0.8085
## (0,10]   0.6890 0.5010 0.8061
## (0,15]   0.6878 0.5000 0.8050
## (0,17]   0.6871 0.4994 0.8044
## --------------------------------------------------------
## Analysis of differences between groupwise PAFs:
##
##      SEX: Male - SEX: Female
##      -----------------------
##            PAF Diff SE(PAF Diff) Z value Pr(>|Z|)
## (0,5]   -0.022870     0.113464  -0.202    0.840
## (5,10]  -0.016998     0.112604  -0.151    0.880
## (10,15] -0.011989     0.112123  -0.107    0.915
## (15,17] -0.005724     0.112086  -0.051    0.959
## (0,5]   -0.022870     0.113464  -0.202    0.840
## (0,10]  -0.018856     0.112830  -0.167    0.867
## (0,15]  -0.016574     0.112510  -0.147    0.883
## (0,17]  -0.015492     0.112415  -0.138    0.890
## --------------------------------------------------------
```

References

Richard Doll, Richard Peto, Jillian Boreham, and Isabelle Sutherland. Mortality in relation to smoking: 50 years' observations on male british doctors. *BMJ*, 328(7455):1519, 2004. ISSN 0959-8138. doi: 10.1136/bmj.38142.554479.AE. URL https://www.bmj.com/content/328/7455/1519.

J. D Kalbfleisch. *The statistical analysis of failure time data*. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, N.J., 2nd ed. edition, 2002. ISBN 1118031237.

Maarit A Laaksonen, Tommi Härkänen, Paul Knekt, Esa Virtala, and Hannu Oja. Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design. *Statistics in medicine*, 29(7-8):860–874, 2010a.

Maarit A Laaksonen, Paul Knekt, Tommi Härkänen, Esa Virtala, and Hannu Oja. Estimation of the population attributable fraction for mortality in a cohort study using a piecewise constant hazards model. *American journal of epidemiology*, 171(7):837–847, 2010b.

Maarit A Laaksonen, Esa Virtala, Paul Knekt, Hannu Oja, and Tommi Härkänen. SAS macros for calculation of population attributable fraction in a cohort study design. *Journal of Statistical Software, Articles*, 43(7):1–25, 2011. ISSN 1548-7660. doi: 10.18637/jss.v043.i07.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL https://www.R-project.org/.

Terry M Therneau. *A Package for Survival Analysis in S*, 2015. URL https://CRAN.R-project.org/package=survival. version 2.38.

ID	F_PERIOD	B_COHORT	SEX	SMOKE	DEATH	FT_END
9	(0,5]	1910	Female	<30/day	FALSE	5.00
9	(5,10]	1910	Female	<30/day	FALSE	5.00
9	(10,15]	1910	Female	<30/day	FALSE	5.00
9	(15,17]	1910	Female	<30/day	FALSE	1.95
88	(0,5]	1930	Male	Never	FALSE	5.00
88	(5,10]	1930	Male	Never	FALSE	5.00
88	(10,15]	1930	Male	Never	FALSE	5.00
88	(15,17]	1930	Male	Never	FALSE	1.92
100	(0,5]	1890	Female	Never	FALSE	5.00
100	(5,10]	1890	Female	Never	FALSE	5.00
100	(10,15]	1890	Female	Never	TRUE	4.92
100	(15,17]	1890	Female	Never	—	—

Table 2: Sample of modified data returned by `gen_data()`. This shows risk factors and outcomes associated with three individuals in a time-varying format: follow-up period (F_PERIOD), the time period in years of interest; birth cohort (B_COHORT), indicating the start of twenty-year period in which the individual was born; sex (SEX), male or female; smoking status (SMOKE), one of never, former, less than 30 cigarettes per day (pipe or cigar smokers are included in this category) or at least 30 cigarettes per day. The DEATH column indicates if the individual died before the end of follow-up and the FT_END column indicates the length of follow-up in years in the current time period until death, censoring, or the start of the next period. Note that missing values (—) are used in rows after follow-up has ended.

	Hazard ratio	2.5 %	97.5 %
SEXMale	—	—	—
SEXFemale	0.64	0.56	0.74
SMOKENever	—	—	—
SMOKEFormer	1.22	1.03	1.45
SMOKE<30/day	1.97	1.67	2.31
SMOKE>=30/day	2.90	1.99	4.24

Table 3: Hazard ratios of death with respect to sex and smoking status. Note that row names consist of the factor name (respectively SEX and SMOKE) followed by the category. Missing values (—) indicate the category is a reference level.

	PAF	2.5 %	97.5 %
(0,5]	0.149	0.113	0.182
(0,10]	0.130	0.100	0.159
(0,15]	0.112	0.086	0.136
(0,17]	0.107	0.082	0.130

Table 4: PAFs and confidence intervals for death in the Mini-Finland Health Survey cohort if current smokers had never smoked. The time interval of interest is shown in the left column, with times measured in years.