graphPAF: An R package to estimate and display population attributable fractions

John Ferguson
University of Galway

Maurice O'Connell
University of Galway

Abstract

graphPAF is intended to facilitate analysis of large real world epidemiological data structures linking risk factors (such as smoking or pollution) to disease. It focuses on estimation and display of different types of population attributable fractions (PAF) and impact fractions which measure the disease burden attributable to risk factors and can subsequently be used to prioritise public health interventions that best prevent disease on a population level.

Features of **graphPAF** demonstrated in this manuscript include inference for standard population attributable fraction and impact fractions, the display of these calculations using attributable fraction fan-plots and nomograms, computation and display of attributable fraction analogues for continuous exposures, inference for attributable fractions appropriate for specific risk-factor → mediator → outcome pathways (pathway-specific attributable fractions) and Bayesian network-based calculations and inference for joint, sequential and average population attributable fractions in multi-risk factor scenarios.

Keywords: PAF, Bayesian network, causal inference, mediation, epidemiology, nomogram.

1. Introduction

Population attributable fractions (PAFs) measure the extent that a particular disease is related to a particular behaviour or exposure (typical examples being smoking, exercise or air pollution), often referred to as a risk factor. The most straightforward examples of attributable fractions pertain to risk-factors that can be eliminated from the population, at least in theory. For instance, one could imagine a population similar to Ireland in almost every way (for instance having similar demographics, culture, a similar health system and so on), except that nobody in this population smoked. How might the rate of heart failure in hypothetical non-smoking Ireland compare to the rate in real Ireland? If the PAF for heart disease attributable for smoking is 12% (as was estimated in Sinha, Ning, Carnethon, Allen, Wilkins, Lloyd-Jones, and Khan (2021)), this means that 12% of the cases of heart failure that occur in the real Ireland, would be avoided in a hypothetical Ireland where nobody smoked. Population attributable fractions are important metrics for determining how pertinent particular risk factors are in determining disease, as well as ranking differing risk factors as targets for health interventions.

There are a number of current R packages for estimating attributable fractions under various study designs, mostly designed for the standard setting that considers population-level elimination of a single binary-valued risk factor. **paf**, implements methods described in Chen, Lin,

and Zeng (2010) and concentrates on estimation under Cohort designs using a Proportional Hazards model. **attribrisk**, Louis Schenck (2014), estimates attributable fractions in matched and unmatched case-control designs. More recently, **AF** and **stdReg** described in Dahlqwist, Zetterqvist, Pawitan, and Sjölander (2016) and Sjölander (2018) enable estimation of PAF in cross-sectional, case-control and cohort settings. **pifpaf**, Camacho-García-Formentí and Zepeda-Tello (2019) specializes on estimation of PAF using cross-sectional summary data over several independent populations. The new R package **graphPAF** described here also estimates PAF for cross sectional, case control and cohort data. However, graphPAF extends these calculations in a number of ways as mentioned below.

In the case that many risk factors are under consideration, differing kinds of analyses may be of interest. **graphPAF** implements fan-plots and nomograms that graphically display the inter-reationships between attributable fractions, relative risk and risk factor prevalence for multiple risk factors, as described in Ferguson, O'Leary, Maturo, Yusuf, and O'Donnell (2019). These analyses can be useful in identify clusters of risk factors that behave similarly, in producing visual rankings of disease burden attributable to differing risk factors, and sometimes in visualizing the effects of interventions.

Joint PAF refers to collective disease burden represented by a group of risk factors (and involves consideration of a hypothetical population where all risk factors in the group were eliminated). Sequential PAFs examine incremental effects on population disease prevalence when each of the risk factors in the group is eliminated in some order. Average PAFs (literally an average of all possible sequential PAF for each risk factor), allow one to partition the joint PAF into contributions for each risk factor. Previous R implementations of average, sequential and joint PAF (for example the R package (**averisk**, Ferguson, Alvarez-Iglesias, Newell, Hinde, and O'Donnell (2018)), have been agnostic to the causal structure linking risk factors to disease, which can result in biased estimation in scenarios where multiple risk factors of interest are on the same causal pathway to disease (for instance if smoking affects blood pressure which affects disease, smoking and blood pressure would be considered to be on the same pathway). In contrast, in these settings **graphPAF** incorporates known risk-factor/risk-factor and risk-factor/disease relationships using a causal Bayesian network model Ferguson, O'Connell, and O'Donnell (2020b).

Referring to the putative pathway: "smoking -> blood pressure -> heart disease" mentioned above, one might wonder about the extent to which this particular pathway contributes to heart disease. This is measured by the pathway-specific population attributable fraction, O'Connell and Ferguson (2022), and can be also calculated by **graphPAF**. Smoking may affect heart disease by mechanisms other than through blood pressure; provided data is available pathway-specific attributable fractions can also be used to determine the most important pathways through which smoking affects disease. The previous R package **causalPAF**, O'Connell and Ferguson (2021), can be used to calculate pathway-specific PAFs, however we have updated the estimation routine to be more efficient and robust in **graphPAF**.

In the case of continuous risk-factors or exposures, zero exposure or alternatively elimination of the risk factor can be nonsensical to consider. Consider body mass index (BMI) as an example; zero BMI is obviously unattainable, and extremely low BMI might be as detrimental to one's health as high BMI. Versions of attributable fractions appropriate in these settings, that allow valid comparisons of disease burden across differing exposures and don't resort to categorization, are described in Ferguson, Maturo, Yusuf, and O'Donnell (2020a). These metrics are also implemented in **graphPAF**.

In summary, **graphPAF** extends and consolidates existing packages for PAF estimation in multiple ways. In this manuscript, we describe each of its features in more depth, interweaving between the theory for PAF estimation and using **graphPAF** in practice.

2. Basic PAF estimation

In this section, we imagine a setting where the risk factor is either binary, or perhaps multi-category with some level indicating 'elimination' for the risk factor. Let Y denote a binary disease outcome (1 indicating disease) for a randomly selected individual from the population, and Y_0 the same binary disease outcome but where the individual is sampled from a hypothetical population with the risk factor eliminated. The population attributable fraction can be defined as:

$$PAF = \frac{P(Y = 1) - P(Y_0 = 1)}{P(Y = 1)} \tag{1}$$

, where $P(Y = 1)$ represents the prevalence of disease in the current population, and $P(Y_0 = 1)$ the prevalence of disease in the hypothetical population with the risk factor eliminated. (1) is an appropriate estimand in case/control and cross sectional studies. Often in longitudinal cohort studies, a cohort of healthy individuals are followed over time with some eventually developing disease. In this setting, a differing kind of population attributable fraction is calculated where the cumulative incidence of disease for that cohort as a function of time is compared to what they incidence would be if the factor had been eliminated from the cohort:

$$PAF(t) = \frac{P(T \leq t) - P(T_0 \leq t)}{P(T \leq t)} \tag{2}$$

Here random sampling is interpreted as random sampling from the cohort of interest and PAF, Y and Y_0 from (1) are replaced by $PAF(t)$, $I\{T \leq t\}$ and $I\{T_0 \leq t\}$, with the random variables T and T_0 representing time to disease in the current population and under hypothetical elimination of the risk factor. In the setting of competing events (such as death) we can interpret T as the time an individual would have developed disease had the competing event not occurred, and the PAF in terms of prevented disease under elimination of the risk factor provided the competing event did not happen. However we can also incorporate competing events directly in the definition of PAF. Suppose Δ represents an indicator for the event that happens first with $\Delta = 1$ indicating that disease occurred before any other event). We can then write:

$$PAF^*(t) = \frac{P(T \leq t \text{ and } \Delta = 1) - P(T_0 \leq t \text{ and } \Delta_0 = 1)}{P(T \leq t \text{ and } \Delta = 1)} \tag{3}$$

Note that as $t \to \infty$, $PAF^*(t)$ converges to the PAF for disease incidence described in Laaksonen, Härkänen, Knekt, Virtala, and Oja (2010). While (3) may at first seem a more sensible estimand than (2) in the presence of competing events, care must be taken in its interpretation. For instance, if the risk factor leads to early mortality due to other mechanisms than the disease of interest (3) may be negative for large t even when the risk factor causes disease. In other words, while (3) is the proportional difference in cumulative incidence in

disease by time t due to removing the risk factor, it can't be interpreted as disease-incidence prevented by eliminating the risk factor. In contrast (2) does have an interpretation in terms of prevented hypothetical incidence in the absence of competing events.

Note that (1), (2) and (3) are causal entities, and unbiased estimation with observational data requires relatively strong assumptions. For instance, if the random variable $A \in \{0, 1, ..., n_A\}$ represents the observed risk factor ($A = 0$ coding for elimination), one cannot say that $P(Y = 1|A = 0) = P(Y_0 = 1)$, unless the risk factor A could be considered randomly assigned. Informal sufficient conditions for the possibility of asymptotically unbiased estimation of (1) are:

1. Unambiguous definition and measurement of the potential outcome: Y_0, representing risk factor elimination. (This is essentially the famous Stable Unit treated value assumption (SUTVA), first described in Rubin (1974))
2. The measurement of a collection of covariates C, so that for observed value of C, $P(Y = 1|A = 0, C) = P(Y_0 = 1|C)$ (This will be true if within joint strata of the covariates C, the risk factor A behaves as if it were randomly assigned). The collection C is sometimes refered to as a sufficient adjustment set of covariates.
3. The proposed model for disease, conditional on risk factor and covariates, $P(Y = 1|A = 0, C)$ is correctly specified.

Similar conditions need to be assumed to estimate (2) and (3). The variables C are often, but not always, a set of confounders of the risk factor/outcome relationship (that is they are joint causes of A and Y). We will assume the veracity of these conditions (including the measurement of a sufficient adjustment set C) in what follows, although their validity should be carefully considered in any practical application.

Assuming these conditions, differing estimators are appropriate dependent on the study design. For cross sectional and case control designs, (1) can be estimated by:

$$\frac{\sum_{i \leq N} w_i(\hat{P}(Y_i = 1|A_i, C_i) - \hat{P}(Y_i = 1|A_i = 0, C_i))}{\sum_{i \leq N} w_i \hat{P}(Y_i = 1|A_i, C_i)} \tag{4}$$

where $i \in \{1, ..., N\}$ indexes the sampled individuals, Y_i, A_i, C_i the disease outcome, risk factor and covariates for individual i, and w_i is a weight specific to individual i. Usually w_i would be set to 1 for cross sectional datasets and is specified based on disease prevalence for case-control datasets. In **graphPAF**, $\hat{P}(Y_i = 1|A_i, C_i)$ may be estimated via log-linear, logistic or conditional logistic models (when disease prevalence is known). Note that when disease prevalence is specified as π, **graphPAF** adds a constant to the linear predictor of the estimated model to ensure that $\sum_{i \leq N} \hat{P}(Y_i = 1|A_i, C_i) = N\pi$. If disease prevalence is unknown, (4) can't be used for estimation in case control studies; instead, the formula by Bruzzi, Green, Byar, Brinton, and Schairer (1985) should be used:

$$1 - \frac{1}{N_c} \sum_{i \leq N: Y_i = 1} \frac{\hat{P}(Y = 1|A_i = 0, C_i)}{\hat{P}(Y = 1|A_i, C_i)} = 1 - \frac{1}{N_c} \sum_{i \leq N: Y_i = 1} \hat{RR_i}^{-1} \tag{5}$$

where $RR_i = P(Y = 1|A_i, C_i)/P(Y = 1|A_i = 0, C_i)$ is the estimated relative increase in disease risk encountered by individual i based on their risk factor value A_i and $N_c = \sum_{i \leq N} I\{Y_i =$

1} is the number of cases in the data set. RR_i can be estimated via approximation with the corresponding odds ratio in case-control study designs provided the disease is relatively rare.

In cohort designs, often cox proportional hazard models are used to estimate (2). Under the proportional hazards assumption, suppose $\hat{r}(C_i, A_i)$ is the estimated hazard ratio for an individual with covariates C_i and risk factor A_i compared to their hazard assuming all covariates and risk factors were at reference levels (defined as 0 for continuous covariates). Let $\hat{H}_0(t)$ be an estimate of the cumulative baseline hazard function. (**graphPAF** uses the Kalbfleich-Prentice estimate for the baseline cumulative hazard function, the default in the **survival** package) for $\hat{H}(t)$. $PAF(t)$ is estimated as:

$$\hat{PAF}(t) = \frac{\sum_{i \leq N} e^{-\hat{H}_0(t)\hat{r}(C_i, A_i=0)} - e^{-\hat{H}_0(t)\hat{r}(C_i, A_i)}}{\sum_{i \leq N}(1 - e^{-\hat{H}_0(t)\hat{r}(C_i, A_i)})} \tag{6}$$

To estimate $PAF^*(t)$, $\hat{r}(C_i, A_i)$ can be replaced in (6) by the estimated Fine Gray subdistribution hazard ratio: $\hat{r}^{FG}(C_i, A_i)$ for disease incidence, and $e^{-\hat{H}_0(t)}$ by $e^{-\int_0^t \hat{h_0}^{FG}(u)du}$, where $h_0^{FG}(u)$ is the baseline subdistribution hazard function at time u. As described in the estimation section below, these functions can be estimated by prior weighting of the Cox Model.

We now illustrate the estimation of (1) and (2) using **graphPAF**.

2.1. Estimation of PAF in cross sectional and case control designs

The function `PAF_calc_discrete` allows PAF estimation for binary and multi-level risk factors for cross-sectional, case control and longitudinal cohort designs. As an example, consider estimating the PAF for the variable `exercise`, a binary indicator for physical inactivity, using the dataframe: `stroke_reduced`. `stroke_reduced` is a simulated matched case-control dataset including 10 stroke risk factors for 6,856 stroke cases and 6,856 stroke controls. The simulations were calibrated accorded to probability distributions estimated using a Bayesian network model fitted to real data from the INTERSTROKE project: O'Donnell, Chin, Sumathy, and et al (2016). Stroke cases and healthy controls in `stroke_reduced` are matched by age-group, gender and region.

First, to deal with the case-control matching, we fit a conditional logistic model to describe the relationships between the prevalence of stroke, exercise and assumed confounders:

```
R> library(splines)
R> library(survival)
R> library(graphPAF)
R> model_exercise <- clogit(formula = case ~ age
+    + education +exercise + ns(diet, df = 3) + smoking + alcohol
+    + stress + ns(lipids,df = 3) + ns(waist_hip_ratio, df = 3)
+    + high_blood_pressure +strata(strata),data=stroke_reduced)
```

The PAF for physical inactivity can then be calculated by applying `PAF_calc_discrete` to `model_exercise`

```
R> PAF_calc_discrete(model_exercise, "exercise", refval=0, data=stroke_reduced,
+    calculation_method="B")
```

```
[1] 0.3322625
```

For case-control datasets such as `stroke_reduced`, the "Bruzzi" method is recommended to calculate PAF as it doesn't require specification of disease prevalence, provided the disease is relatively rare (the approxiation of risk-ratios with odds-ratios might be unacceptably inaccurate otherwise). If prevalence or alternatively average incidence over a period of time is known, the '`direct`' method can alternatively be used. For instance:

```
R> library(graphPAF)
R> PAF_calc_discrete(model_exercise, "exercise", refval=0, data=stroke_reduced,
+    calculation_method="D", prev=0.0035)
```

```
[1] 0.3196773
```

assumes that the yearly incidence (averaged over the cohort) of first stroke is 0.0035, which effectively estimates $PAF(1)$, that is equation (2) at $t = 1$. If disease prevalence (rather than an estimated incidence) is used in the argument `prev`, the estimator will estimate (1). Note that when lifetime disease incidence across the cohort low, relative risks and hazard ratios should correspond and one would expect equation (1) to be approximately equal to (2) at varying t.

For PAF calculations with cross sectional dataset, a `glm` model should first be fit that describes the relationship between risk factor and disease, conditional on covariates. Note in this case that only logistic or log-linear binomial models are permitted in **graphPAF**. Provided the sample is representative of the source population, the argument `prev` doesn't need to be set, and `calculation_method="D"` should instead be used. `PAF_calc_discrete` will then estimate equation (1).

By default the reference level for a binary risk factor (that is a risk factor coded as 0/1) is set to 0. In the example above, 1 codes physical inactivity. The function can also estimate PAF for multi-category risk factors provided `refval` is set correctly.

Bootstrap-calculated confidence intervals for PAF can be requested via `ci=TRUE`. The Bootstrap is implemented via the R-package **boot**, which is loaded by default when **graphPAF** is installed. The **boot** library allows parallelization of the calculation of the Bootstrap iterates. While we utilize `multicore` clusters in the code below (which are only implemented on Linux and MacOS based operating systems) to ensure reproducibility of confidence intervals, in general we recommend using `snow` clusters. `snow` clusters tend to be more memory efficient than **multicore** clusters in addition to being platform independent. The arguments `ci_type`, defaulting to confidence intervals appropriate for normal sampling distributions (with bias corrections) and `ci_level` control the type of confidence interval and confidence level, with `nboot` controlling the number of Bootstrap repetitions. Similar bootstrap generated confidence intervals (requested using the same arguments) are also available for other **graphPAF** functions detailed later such as: `impact_fraction`,`PAF_calc_continuous`,`ps_paf`,`joint_paf`,`average_paf` and `seq_paf`.

```
R> library(parallel)
R> options(boot.parallel='multicore')
R> options(boot.ncpus=5)
```

```
R> RNGkind("L'Ecuyer-CMRG")
R> set.seed(23092002, "L'Ecuyer")
R> PAF_calc_discrete(model_exercise, "exercise", refval=0, data=stroke_reduced,
+    calculation_method="B",ci=TRUE, boot_rep=100,ci_type="norm",ci_level=.95)

         est         bias debiased_est   norm_lower   norm_upper
     0.33200      0.00547      0.32700      0.26000      0.39300
```

In Cohort datasets, often survival regression methods are used and estimation focuses on (2). Often the initial model considered is a proportional hazards regression model, fit via the R-function `phreg`, from the **survival** library. As an example in the dataframe `stroke_reduced` `time` denotes a simulated survival time to some event in the stroke controls (individuals with `event=0` are considered to not have experienced the event at study completion or when they left the study, and are censored). The following model might be fit:

```
R> model_high_blood_pressure_coxph <- coxph(formula = Surv(time,event) ~
+    ns(age,df=5) + education + exercise + ns(diet, df = 3) + smoking +
+      alcohol + stress + ns(lipids,df = 3) + ns(waist_hip_ratio, df = 3)
+    + high_blood_pressure,data=stroke_reduced[stroke_reduced$case==0,])
```

Suppose we are interested in the proportion of the events in the sub-cohort that might have been avoided if nobody in the cohort was hypertensive (hypertension being represented by the binary variable `high_blood_pressure`. At time 0, nobody had experienced an event, and over time the cumulative number of events (and also the proportion of events that might be avoided) will change. The user can specify the times, t, at which to calculate $PAF(t)$ using the argument `t_vector`:

```
R> PAF_calc_discrete(model_high_blood_pressure_coxph, "high_blood_pressure", refval=0,
+    data = stroke_reduced[stroke_reduced$case==0,], calculation_method="D", ci=TRUE,
+    boot_rep=50, ci_type=c('norm'),t_vector=c(1,2,3,4,5,6,7,8,9))

    est    bias debiased_est norm_lower norm_upper
1 0.397 0.00341        0.394     0.3640      0.424
2 0.391 0.00317        0.388     0.3590      0.417
3 0.379 0.00369        0.376     0.3450      0.406
4 0.361 0.00402        0.357     0.3210      0.393
5 0.327 0.00483        0.322     0.2740      0.370
6 0.293 0.00585        0.287     0.2270      0.347
7 0.256 0.00653        0.249     0.1800      0.319
8 0.220 0.00804        0.212     0.1350      0.288
9 0.176 0.00869        0.168     0.0876      0.248
```

The results indicating that while 39.7% of events that happen with a year might have been avoided in a hypertension-free population, only 17.6% of events that happen within 9 years would be avoided. This is the typical pattern one expects for an event such as death which (unfortunately) can only be delayed but not prevented by the risk factor's absense.

If it is prefered to estimate (3) rather than (2), and data on competing events exists a weighted Cox model should instead be used with weights calculated using the function `finegray` from the **survival** package. Sending the weighted cox model to `PAF_calc_discrete` will utilize the

Fine Gray modification of (6) described earlier. See Therneau, Crowson, and Atkinson (2021) for more details.

2.2. Estimation of Imact fractions

While population attributable fractions (PAF) can summarize the overall impact or importance of a risk factor on disease burden, they tend to give an overly optimistic impression of what an intervention on that risk factor might achieve. The predominant reasons for this are first that it may be difficult if not impossible to eliminate the risk factor from the population (think of the difficulties in preventing all forms of smoking or alcohol-use or enticing an entire population to change their dietary habits) and second that even if one could eliminate the risk factor, disease risk in individuals who previously were exposed might not equal the disease risk if they were never exposed (for instance, former smokers may have higher disease risk than comparable individuals who never smoked) Bulterys, Morgenstern, and Weed (1997).

In contrast, population impact fractions purport to measure the proportional reduction in disease risk from a realistic health intervention that may reduce the prevalence of a risk factor (rather than eliminate the risk factor), or perhaps favorably change the collective statistical distribution of many risk factors. The function `impact_fraction` in **graphPAF** can estimate impact fractions under the study designs considered above (cross-sectional, cohort and case-control). We first need to specify how the health intervention changes the distribution of risk factors that might affect disease, through the `new_data` argument. For instance, imagine a health-intervention (perhaps a national campaign to encourage walking) reduces the prevalence of inactivity by 20%. Assuming the intervention has no effect on any other risk factor, the following code shows how such an intervention might be specified using the `new_data` argument

```
R> new_data <- stroke_reduced
R> N <- nrow(new_data)
R> inactive_patients <- (1:N)[stroke_reduced$exercise==1]
R> N_inactive <- sum(stroke_reduced$exercise)
R> newly_active_patients <- inactive_patients[sample(1:N_inactive,0.2*N_inactive)]
R> new_data$exercise[newly_active_patients] <- 0
```

The impact fraction for such an intervention is then calculated using:

```
R> impact_fraction(model=model_exercise,stroke_reduced,new_data,
+    calculation_method = "B")
```

```
[1] 0.06707932
```

indicating that the health intervention might result in a 6.5% reduction in the rate of strokes. Note that this calculation really refers to the difference in disease risk in two comparable populations, one with a reduced rate of inactivity. Since changing one's behaviour may not completely eliminate cumulative damage due to prior unhealthy lifestyle, this estimated 6.5% might overestimate the impact of the intervention at least in the short term. If the 20% reduction in 'inactivity' is sustained over many years, it is reasonable that this estimate represents a long run effect of the health intervention.

2.3. PAF nomograms

graphPAF facilitates plotting of the inter-relationships between prevalence, odds ratios and attributable fractions over multiple risk factors using methods described in detail in Ferguson *et al.* (2019). These plots utilize the concept of 'approximate-PAF', derived in the same paper:

$$PAF_{approx} = log(OR)\pi_{control} \approx PAF \tag{7}$$

where OR is the causal odds ratio between a risk factor and disease, and $\pi_{control}$ is the prevalence of the risk factor in controls. This approximation stems from a Taylor expansion of the PAF around a relative-risk of 1, and will be most accurate for risk factors that have relatively small effects on a relatively rare outcome. One interesting observation regarding approximate PAF is the symmetric roles that risk factor prevalence and log-odds ratio play in its definition; indicating that similar changes in either lead to a similar impact on disease on a population level. To create a fan plot, risk factor data (names, prevalences and log-odds ratios) must be first summarized into an `rf_summary` object before plotting. For instance:

```
R> rfs <- rf_summary(rf_names=c('Hypertension','Inactivity','ApoB/ApoA','Diet',
+        'waist_hip_ratio','Smoking','Cardiac causes','Alcohol','Global Stress',
+        'Diabetes'),rf_prev=c(.474,.837,.669,.67,.67,.224,.049,
+        .277,.144,.129),risk=c(1.093,0.501,0.428,0.378,0.294,
+         0.513,1.156,0.186,0.301,0.148),log=TRUE)
```

creates such an object for 10 risk factors from the INTERSTROKE database. `rf_prev` represents the prevalence of the risk factor in controls. For risk factors with more than 2-levels (here `ApoB/ApoA`, `waist_hip_ratio` and `alcohol` have 3 levels), the prevalence of the non-reference levels of the risk factor should be used as `rf_prev`. While technically, `rf_prev` should be the prevalence of the risk factor in controls, this can be substituted with population-prevalence when prevalence in controls is unavailable if the disease is rare. By default, the argument `risk` should specify confounder-adjusted log-odds ratios for association between risk factor and outcome , although odds ratios can be used if `log` is set to `FALSE`. Such log-odds ratios can be conveniently estimated via logistic regression models. Plotting this `rf_summary` object, using default settings, produces Figure 1 below.

Approximate PAF is represented on a fan plot as both the slope of the line adjoining a point to the y-axis, and also the y-axis intercept of that adjoining line. Fan plots are read clockwise from the upper left corner along the rays of decreasing approximate PAF (which is again the slope of the ray), and display risk factor prevalence and odds ratio (based on the x-axis and y-axis intercept of a particular point) for the risk factors under comparison, in addition to the approximate PAF.

Imagine now a successful health intervention that reduces the prevalence of smoking by about 50%. This information might be displayed in a `rf_summary` object as follows:

```
R> rfs <- rf_summary(rf_names=c('Hypertension','Smoking',
+          'Smoking (after health intervention)'),rf_prev=c(.474,.224,.11)
+          ,risk=c(1.093,0.513,0.513),log=TRUE)
```

Like a fan plot, attributable fraction nomograms display joint information on prevalence, odds ratio and approximate PAF, but this time on three vertical axes, with a risk factor

```
R> plot(rfs)
```

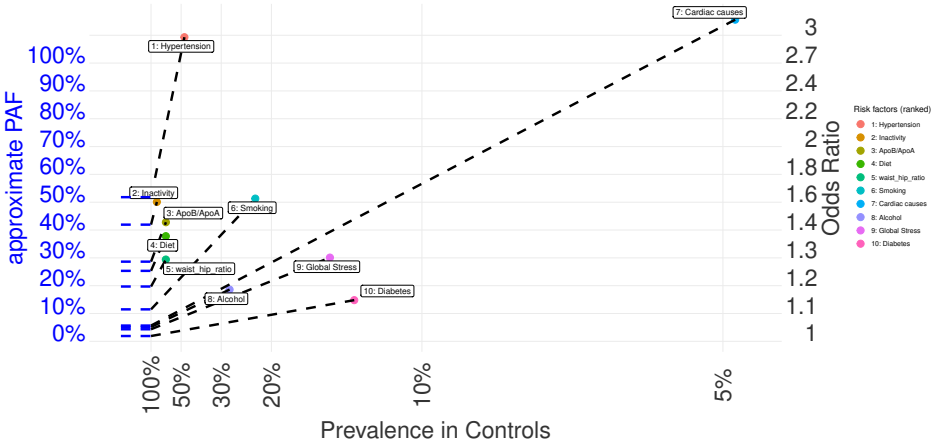


Figure 1: Fan Plot displaying Prevalences, odds ratios and approximate PAF for INTERSTROKE risk factors. Approximate PAF is represented as both the slope of the line adjoining a point to the y-axis, and also the y-axis intercept of that adjoining line. The fan plot indicates that hypertension and inactivity are the two most prominent risk factors in stroke pathogenesis. Cardiac disease is an outlier on the plot. While it has the highest estimated relative risk, it has low prevalence (less than 5%) in comparison with the other risk factors and is only ranked 7th in terms of disease burden

represented by the line connecting these three data-points. An intervention will usually work by changing the population prevalence of the risk factor, without affecting the odds ratio. This can be graphically represented by rotating the line for the risk factor, using the (unaffected) odds ratio as a pivot, from the old prevalence through the new prevalence, as Figure 2 represents. Of course, other risk factors can also be represented on this plot (as is hypertension here). Plotting a `rf_summary` object with argument `type='rn'` produces the Figure below. If preferred, using `type = 'n'`, uses the odds ratio, rather than prevalence as the center-axis, with risk factor prevalence being the left-hand axis, but is otherwise interpreted similarly.

3. Estimation with Continuous Exposures

Frequently, a discrete risk factor such as hypertension is generated by the truncating an underlying continuous exposure, such as blood pressure. Not accounting for this underlying continuity may result in underestimation of disease burden attributable to the exposure as some individuals with the reference value of the discretized risk factor (for instance individuals not regarded as hypertensive) may still be at elevated risk of disease due to their values of the exposure not being optimal (for instance if hypertensive is defined as systolic blood pressure above 140mm/Hg, an individual with systolic blood pressure of 139 would fall into the reference group, but might have some increased risk of cardiovascular disease compared to if their blood pressure was 120). Ferguson *et al.* (2020a) discusses these issues and suggests a variety of appropriate estimands for continuous exposures.

Using the notation from Ferguson *et al.* (2020a), we consider the exposure for a randomly individual from the population as a continuous random variable, X, with Y representing a

```
R> plot(rfs,type='rn')
```

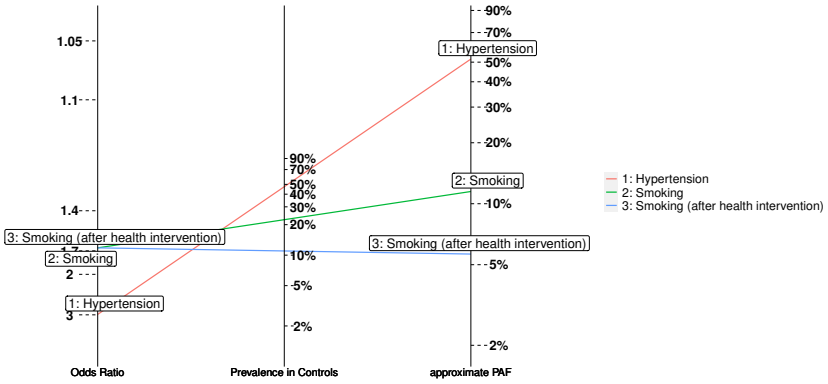


Figure 2: Attributable fraction nomogram for INTERSTROKE risk factors. Nomograms like the above give an alterantive way to display joint intervention regarding odds ratios , prevalences and approximate PAF. They can also be used to visualize interventions. For instance, the green and blue lines represent smoking in a population pre and post intervention. The odds ratio for smoking isn't affected by the intervention, but the prevalence is. The effect of the intervention for smoking PAF can be visualized by rotating the line for smoking (using the left axis odds ratio as a pivot) through the new prevalence post intervention

binary disease outcome. We let Y_x represent the potential outcome if $X = x$, which we assume is well defined. Assuming that $P(Y_x = 1)$, considered as a function of x, has some minimum value x_{min} within the physiological limits of the exposure X, we define PAF as:

$$PAF = \frac{P(Y = 1) - P(Y_{x_{min}} = 1)}{P(Y = 1)}, \tag{8}$$

In the circumstance that $P(Y_x = 1)$ is strictly decreasing or strictly increasing as a function of x, the minimum value x_{min} may be undefined, in which case we define PAF as:

$$PAF = \frac{P(Y = 1) - \text{Inf}\{P(Y_x = 1)\}}{P(Y = 1)}, \tag{9}$$

defining $\text{Inf}\{P(Y_x = 1)\}$ as the infimum of the set of probabilities $P(Y^x = 1)$ with x ranging over the possible range of exposure values.

As explained in Ferguson *et al.* (2020a), the estimands (8) and (9) may be difficult to estimate when x_{min} falls in the extremes of the exposure distribution. As an alternative, the family of estimands: PAF_q for $q \in (0, 1)$ are suggested as alternative metrics. Intuitively PAF_q is the impact fraction for an intervention that shifts the exposure value for individuals in the population where it is beneficial to do so, with the intervention not effecting individuals where such a shift is not necessary. $1-q$ indicates the proportion of individuals affected by the intervention, in addition to how large the shift in exposure values is for those affected (larger values of $1 - q$ indicating larger shifts). More technically, exposure values X for individuals whose disease risk (based purely on the exposure value and not on other covariates) is above

the $100q^{th}$ percentile of disease risk are moved to the closest possible value $f_q(X)$, where the disease risk of $f_q(X)$ is at the $100q^{th}$ percentile. Individuals with good exposure values (corresponding to risk values below the $100q^{th}$ percentile), are unaffected by this intervention. PAF_q when $q > 0$ tends to be easier to estimate than PAF (effectively estimating PAF_q often involves less extrapolation than estimating PAF). It also has a more concrete real world interpretation as the impact fraction for an acheiveable intervention. PAF_q is defined more precisely as:

$$PAF_q = \frac{P(Y=1) - P(I\{X \in R_q\}Y + I\{X \notin R_q\}Y^{f_q(X)} = 1\}}{P(Y=1)} \tag{10}$$

where R_q is the interval of exposure values corresponding to the bottom $100q\%$ of risk and $f_q(X)$ is the closest point in the closure of R_q to X. Note that as $q \downarrow 0$, $PAF_q \uparrow PAF$.

Under continuous analogs of the conditions 1), 2) and 3) listed on pages 3 and 4, (10) can be estimated as

$$PAF_q = \frac{E_C(I\{X \notin R_q\}(\hat{P}(Y=1|X,C) - \hat{P}(Y=1|\,\hat{f}_q(X),C)))}{P(Y=1)} \tag{11}$$

and

$$PAF_q = E_{X,C|Y=1} I\{X \notin R_q\}[1 - \frac{\hat{P}(Y=1|\,\hat{f}_q(X),C)}{\hat{P}(Y=1|X,C)}] \tag{12}$$

and

$$\hat{PAF}_q(t) = \frac{\sum_{i \leq N}(e^{-\hat{H}_0(t)\hat{h}(C_i,A_i)} - e^{-\hat{H}_0(t)\hat{h}(C_i,\hat{f}_q(X))})}{\sum_{i \leq N} e^{-\hat{H}_0(t)\hat{h}(C_i,X_i)}} \tag{13}$$

respectively for cross sectional, case control and cohort designs, with $\hat{f}_q(x)$ the estimated value for $f_q(x)$ and $\hat{P}(Y=1|x,c)$, the estimated probability of disease, when the risk factor is x and the covariates are c.

graphPAF uses these equations to estimate PAF_q across differing risk factors. Here we consider the convenient case where a group of continuous risk factors: `waist_hip_ratio`, `diet` and `lipids` all have the same set of underlying confounders, and subsequently estimated effects of each risk factor can be obtained from a single statistical model. The following code demonstrates how such a model might be specified for a case-control dataset.

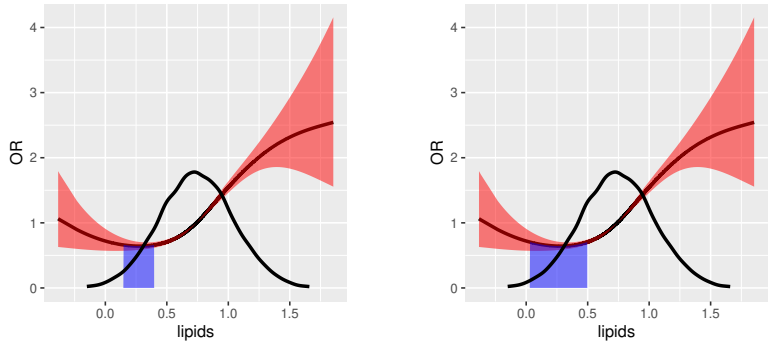
```
R> model_continuous_clogit <- clogit(formula = case ~ region*ns(age, df = 5) +
+      sex*ns(age, df = 5) + education +exercise + ns(diet, df = 3)  + alcohol
+      + stress + ns(lipids,df = 3) + ns(waist_hip_ratio, df = 3) + high_blood_pressure
+      + strata(strata), data = stroke_reduced)
```

Note that the continuous exposures `waist_hip_ratio`, `diet` and `lipids` appear in the model as natural spline terms in the model. One can evaluate the estimated shape of the exposure/outcome relationship, and visualize the interventions corresponding to a particular value of PAF_q using the function `plot_continuous`. As an example:

```
R> plot_continuous(model_continuous_clogit,riskfactor="lipids",
+    data=stroke_reduced,min_risk_q=.1,n_x = nrow(stroke_reduced))
```

```
R> plot_continuous(model_continuous_clogit,riskfactor="lipids",
+      data=stroke_reduced,min_risk_q=.2,n_x = nrow(stroke_reduced))
R> plot_continuous(model_continuous_clogit,riskfactor="lipids",
+    data=stroke_reduced,min_risk_q=.3,n_x = nrow(stroke_reduced))
```

produces estimated relationships between lipids and OR of stroke (with the median value for lipids as a reference by default), but highlights the regions representing the post-intervention ranges of the risk factor for $PAF_{0.1}$,$PAF_{0.2}$ and $PAF_{0.3}$.



(a) $PAF_{0.1}$ (b) $PAF_{0.2}$

(c) $PAF_{0.3}$

Figure 3: Estimated effects of blood lipid levels on the OR of stroke. The density of lipids and pointwise 95% confidence bands for the odds ratios are also plotted. Also shaded blue are the target regions for the intervention associated with PAF_q for various q. For instance $PAF_{0.1}$ corresponds to the smallest 10% of risk

Having fit the model, the function `PAF_calc_continuous` generates PAF_q at any desired set of quantiles. The resulting object is an S3 object of class `PAF_q`, essentially a dataframe with rows for each risk factor,PAF_q combination and columns corresponding to quantiles. This object can be printed and plotted as follows:

```
R> out <- PAF_calc_continuous(model_continuous_clogit,riskfactor_vec=
+      c("diet","lipids","waist_hip_ratio"),q_vec=c(0.01, 0.1,0.3,0.5,0.7,0.9),
+      ci=TRUE,calculation_method="B",data=stroke_reduced,boot_rep=50)
R> print(out)
```

```
        riskfactor q_val     est      bias debiased_est norm_lower norm_upper
1             diet  0.01 0.16200  2.85e-02      0.13300   0.042300     0.2240
2             diet  0.10 0.14600  3.91e-03      0.14200   0.096300     0.1880
3             diet  0.30 0.11100  4.16e-04      0.11100   0.075800     0.1460
4             diet  0.50 0.07990 -2.17e-04      0.08010   0.063000     0.0973
5             diet  0.70 0.04860 -8.75e-04      0.04950   0.034800     0.0642
6             diet  0.90 0.01600 -5.33e-04      0.01660   0.007890     0.0252
7          lipids  0.01 0.37900  8.24e-03      0.37100   0.333000     0.4090
8          lipids  0.10 0.36500  6.27e-03      0.35800   0.326000     0.3910
9          lipids  0.30 0.28200  5.31e-03      0.27700   0.251000     0.3030
10         lipids  0.50 0.17300  3.56e-03      0.17000   0.151000     0.1890
11         lipids  0.70 0.07330  4.88e-04      0.07280   0.056900     0.0887
12         lipids  0.90 0.01290 -5.19e-04      0.01350   0.005190     0.0217
13 waist_hip_ratio  0.01 0.17200  1.75e-02      0.15500   0.062600     0.2460
14 waist_hip_ratio  0.10 0.16100  1.74e-03      0.15900   0.112000     0.2050
15 waist_hip_ratio  0.30 0.11400  5.81e-04      0.11400   0.083100     0.1450
16 waist_hip_ratio  0.50 0.06810  1.09e-03      0.06700   0.048600     0.0855
17 waist_hip_ratio  0.70 0.03080  2.96e-04      0.03050   0.016400     0.0445
18 waist_hip_ratio  0.90 0.00691  2.98e-05      0.00688  -0.000203     0.0140
```

Using the argument `"calculation_method="B"`, as above uses (12) to estimate PAF_q (appropriate in case-control designs), in contrast, `"calculation_method="D"` uses (11), whereas (13) is appropriate for cohort studies with a survival response. Note that in the case of a survival response, $PAF_q(t)$ can only be evaluated at a single time,`t`, specified as `t_vector` being a single element. Plotting PAF_q against q for several risk factors allows one to assess the relative benefits of comparable and achievable interventions on differing risk factors.

For example, the results here indicate that comparable interventions targeting waist hip ratio and diet may have similar effects on disease burden, with interventions on lipids having larger effects. This might motivate an intervention on lipid levels (for example, increased statin use when appropriate) over interventions on diet or BMI, although admittedly many other factors may dictate what if any intervention may be chosen in practice.

4. Pathway-specific PAF calculations

While PAF provides an overall measure of the importance of a particular disease risk factor in causing disease on a population level, the mechanisms by which the risk factor effects disease may also be of interest. For instance, perhaps physical inactivity increases blood pressure which subsequently increases the risk of stroke. Alternatively, physical inactivity might indirectly increase the risk of stroke through weight gain or increased cholesterol levels. In this context, the variables blood pressure, weight gain and cholesterol are regarded as 'mediators', that is they are intermediate variables on differing causal pathways each partially explaining the causal relationship between inactivity and stroke. How important might each pathway be in disease pathogenesis? In O'Connell and Ferguson (2022), this question is addressed by defining an attributable fraction for a particular mediating pathway. Roughly this 'pathway-specific' attributable fraction (PS-PAF for short) can be interpreted as the relative decrease in disease prevalence if a particular mediating pathway didn't exist. For

```
R> plot(out)
```

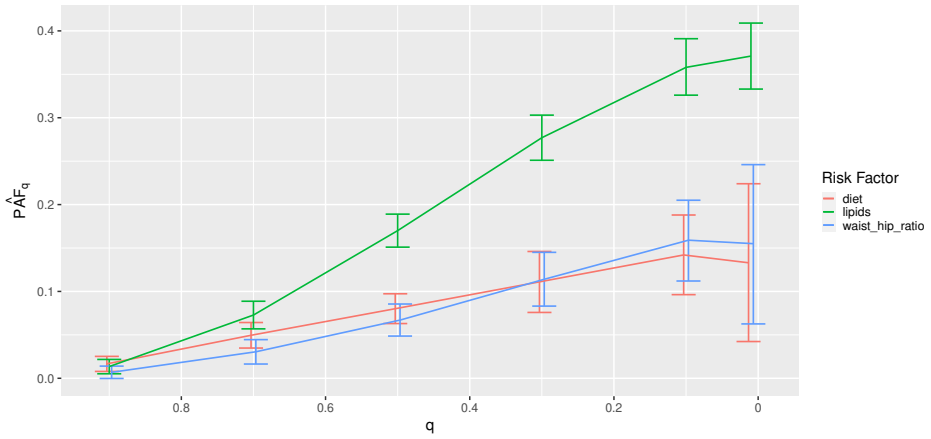


Figure 4: plotting PAF_q over multiple risk factors. The figure indicates that comparable interventions on diet and waist hip ratio (for instance shifting 50% of exposure values as is the case in $PAF_{0.5}$ may have similar effects for diet and waist hip ratio, but much larger effects for lipids). As can be seen in the plot, the confidence intervals for PAF_q are much narrower for smaller q, reflecting the fact that PAF_q is easier to estimate than PAF in addition to representing more realistic interventions

instance imagine there was no effect of physical inactivity on blood pressure; what percentage of stroke might be avoided in such a population? Letting $M^1, ..., M^K$ represent K known mediators of the risk factor outcome relationship, $A \in \{0, 1\}$ a risk factor and $Y \in \{0, 1\}$ a disease outcome, the PS-PAF for mediator $k \leq K$ is denoted as:

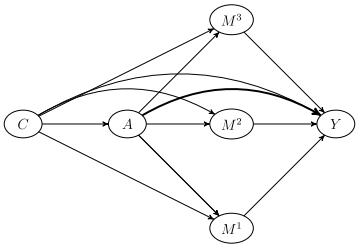


Figure 5: Mediators on separate causal pathways. M^1, M^2 and M^3 mediate the causal relationship between A and Y. These mediators represent independent mechanisms by which A affects Y in that any pathway of direct arrows originating from A and ending at Y can only involve one of the three mediators.

$$PAF_{A->M^k->Y} = \frac{P(Y = 1) - P(Y_{A,M_0^k} = 1)}{P(Y = 1)} \tag{14}$$

$P(Y_{A,M_0^k} = 1)$ can be interpreted as disease prevalence in a hypothetical population which mirrors the actual population in the values of the risk factor A, but where the values for mediator, M^k, behave as if the risk factor didn't exist (note that on an individual level M_0^k is the potential outcome for the k^{th} mediator assuming no exposure to the risk factor, that is

$A = 0$). As described in O'Connell and Ferguson (2022), interpretations for pathway-specific attributable fractions subtly differ based on the causal identifiability assumptions assumed. We are describing the mechanistic interpretation here, although two other interpretations exist. We won't go into these details here and instead refer the interested reader to O'Connell and Ferguson (2022).

In addition to pathway specific PAF for indirect pathways, one can also define an attributable fraction for all 'unobserved' or unknown pathways:

$$PAF_{A->Y} = \frac{P(Y=1) - P(Y_{0,M^1,...,M^K}=1)}{P(Y=1)} \tag{15}$$

(15) denotes the 'direct' pathway specific population attributable fraction, and represents the contribution of mechanisms by which the risk factor affects disease, other than those represented by pathways through M^1,....,M^K (Note that $P(Y_{0,M^1,...,M^K}=1)$ can be interpreted as the disease prevalence in a population where the risk factor was eliminated but with the joint distribution of mediators $M^1, ..., M^K$ being unaffected).

Under the assumptions listed in O'Connell and Ferguson (2022) (with the additional assumption that mediators are on separate causal pathways between the risk ractor and disease), estimating (14) requires fitting a model for the mediator M^k conditional on both the risk factor, A, and the confounder-vector for the exposure outcome relationship C (Note that these models estimate $P(M^k = m|A, C)$ for a discrete mediator and $E(M^k|A, C)$ for a continuous mediator), in addition to fitting a model for the disease outcome, Y, conditional on the exposure A, mediators $M^1, ..., M^K$ and the same set of confounders C. (This second model estimates $P(Y = 1|A, C, M^1, ..., M^K)$). When M^k is continuous, the following estimator for (14) is used:

$$\widehat{PAF}_{A->M^k->Y} = \frac{\sum w_i Y_i - \sum_i w_i \widehat{P(Y=1|A_i, C_i, \hat{M}_i^k, \mathbf{M_i^{\neq k}})}}{\sum w_i Y_i} \tag{16}$$

with $\hat{M}_i^k = M_i^k - E(M^k|\widehat{A=0, C_i})$, with C_i and M_i^k representing the observed values of the confounder vector and k^{th} mediator for person i, and $\mathbf{M_i^{\neq k}}$, the observed values for other mediators for the same individual. Weights w_i are used to account for possible case-control structure. For representative cross sectional samples, these weights should be set to 1 (the default). In contrast, for case control data, these weights can be set based on estimated disease prevalence. In the case that the mediator is discrete, a slightly different estimator is used:

$$\widehat{PAF}_{A->M^k->Y} = \frac{\sum w_i Y_i - \sum_i w_i \sum_{m \in \mathcal{M}^k} \widehat{P(M^k = m|A_i = 0, C_i)} \widehat{P(Y=1|A_i, C_i, M^k = m, \mathbf{M_i^{\neq k}})}}{\sum w_i Y_i} \tag{17}$$

Direct PS-PAF is slightly easier to estimate, as one only needs to fit the outcome model that conditions on the risk factor, A, covariates C and mediators, M^1,...,M^K:

$$\widehat{PAF}_{A->Y} = \frac{\sum w_i Y_i - \sum_i w_i \widehat{P(Y=1|A_i = 0, C_i, M_i^1, ..., M_i^K)}}{\sum w_i Y_i} \tag{18}$$

Examples with graphPAF

To illustrate these calculations with **graphPAF**, suppose we wish to estimate pathway-specific attributable fractions for the 4 pathways from physical inactivity to stroke through waist hip ratio, through blood lipid counts, through high-blood pressure, and through any mediating pathways other than waist hip ratio, blood pressure and lipids from the simulated dataset `stroke_reduced` included in the **graphPAF** library. Since `stroke_reduced` is a case control dataset, weighted models for each mediator and the response need to be fit, to replicate the fits one would expect from a representative sample of the population. In `stroke_reduced`, these weights are already in the dataset and are based on an average incidence of 0.0035 new strokes per person per year (as explained earlier). To calculate the `weights` vector for a different prevalence, `stroke_reduced` could be sent to the `data_clean` function. For instance, if we instead thought that 0.01 was the correct incidence, we could use `stroke_reduced_2 <- data_clean(stroke_reduced,vars=colnames(stroke_reduced),prev=0.01)`. A column of weights would in this case be included in the dataframe `stroke_reduced_2`. Having calculated these weights, models for the response and a list of models for the mediators can be specified:

```
R> response_model <- glm(case ~ region * ns(age, df = 5) + sex * ns(age, df = 5) +
+                          education + exercise + ns(diet, df = 3) +
+                          smoking + alcohol + stress + ns(lipids, df = 3) +
+                          ns(waist_hip_ratio, df = 3) + high_blood_pressure,
+                        data=stroke_reduced,family='binomial', weights=weights)
R> mediator_models <- list(
+    glm(high_blood_pressure ~ region * ns(age, df = 5) + sex * ns(age, df = 5) +
+          education   +exercise + ns(diet, df = 3) + smoking + alcohol + stress,
+        data=stroke_reduced,family='binomial',weights=weights),
+    lm(lipids ~ region * ns(age, df = 5) + sex * ns(age, df = 5) +education +
+         exercise + ns(diet, df = 3) + smoking + alcohol + stress, weights=weights,
+       data=stroke_reduced),
+    lm(waist_hip_ratio ~ region * ns(age, df = 5) + sex * ns(age, df = 5) + education
+       + exercise + ns(diet, df = 3) + smoking + alcohol + stress, weights=weights,
+       data=stroke_reduced))
```

The response model and list of mediator models is then sent to `ps_paf`, which implements the estimators: (16),(17) or (18) with the fitted models. Again, for case control datasets, the argument `prev` needs to be specified for correct calculation of the weights.

```
R> ps_paf(response_model=response_model, mediator_models=mediator_models,
+         riskfactor="exercise",refval=0,data=stroke_reduced,prev=0.0035,
+         ci=TRUE,boot_rep=100,ci_type="norm")

                       est      bias debiased_est norm_lower norm_upper
Direct              0.3350  0.005050       0.3300   0.260000     0.3990
high_blood_pressure 0.0198 -0.001030       0.0208  -0.007260     0.0489
lipids              0.0207  0.000713       0.0200  -0.000896     0.0409
waist_hip_ratio     0.0314  0.000390       0.0310   0.019100     0.0429
```

The results indicate that only a small proportion of the disease burden due to physical inactivity is attributable to pathways involving lipids, blood pressure and waist hip ratio. For instance, if the pathway from physical inactivity to stroke through waist hip ratio were disabled (in that physical inactivity had no deleterious affect on waist hip ratio), relative stroke prevalence would only decrease by 3.1%, with similar interpretations and small PS-PAFs for the pathways through lipids and high blood pressure.

5. Joint PAF

Joint attributable fractions refer to the collective disease burden that can be appropriated to a collection of risk factors. For instance the INTERSTROKE study O'Donnell *et al.* (2016) estimates that roughly 90 % of incident strokes might be avoided if 10 major modifiable stroke risk factors were removed from the population. More formally, the joint population attributable fraction for a set of risk factors, $\mathbf{S}$ can be defined as:

$$PAF_{\mathbf{S}} = \frac{P(Y=1) - P(Y(\mathbf{0_S}=1)}{P(Y=1)}, \tag{19}$$

with the shorthand: $Y(\mathbf{0_S})$ representing the potential outcome where the subset of risk factors $\mathbf{S}$ have been set to their reference levels. Traditionally, such calculations were performed via multivariable regression models that include the set of variables that are to be eliminated. For instance to estimate a joint PAF for stroke associated with stress and a diagnosis of diabetes, disease risk in the data-collected might be compared to predicted disease risk if diabetes status and stress were set to their reference levels, with the predicted disaese risk being computed via a single fitted logistic model. While this approach may be fine if diabetes status and stress share the same set of confounding variables (proviso that the model for stroke risk includes these confounders and is correctly specified) bias may result when effects of one of the risk factors (for example, increased blood pressure is an effect of stress according to Figure 6) confounds the relationship between the response and other risk factors of interest. This is the case here as blood pressure confounds the relationship between diabetes and stroke according to Figure 6 (Note that confounders are variables that cause the risk factor under investigation, and in addition cause the outcome). For these kinds of causal structures, while predicted risks derived via a single regression may correctly reflect the probability that an individual in the dataset with the reference values of stress and diabetes has disease, they will not reflect the probability of disease in a population with all individuals having the reference levels. In other words, the associated estimated joint PAF will not have a causal interpretation.

Ferguson *et al.* (2020b) describes how the intervention corresponding to a joint population attributable fraction (the intervention being the elimination of a subset of risk factors) can be conceptualized via recursive application of Pearl's do-operator Pearl (2009) on the true causal graph (assumed to be a directed acyclic graph or DAG), linking risk factors, outcome and associated risk factor/outcome confounders. This observation facilitates asymptotically unbiased estimation of joint attributable fractions under general causal structures. To achieve this in practice, we need to first know the causal DAG, second have collected data on individuals $i = 1, ..., N$ for all variables represented in the DAG, and finally correctly specify and fit statistical models linking each node in the causal DAG to all of its direct causes (the direct causes being those variables with arrows pointing to the node of interest). Having done this, one can use these fitted models to simulate from the joint-distribution of all variables in the graph (confounders, risk factors and outcome) corresponding to each application of the do-operator. For each application of the do-operator (corresponding to a population level elimination of a single risk factor), this simulation is itself recursive. For instance, if smoking is eliminated, smoking is first set to its reference level (no smoking) for all individuals in the current simulated dataset. Values for the direct effects of smoking (that is the nodes for which smoking is a parent in the causal graph) are then simulated from the conditional distribution of these variables assuming no smoking. Suppose blood pressure is one of the effects of smok-

ing. Next the direct effects of variables such as blood pressure are simulated, conditional on prior simulated values for blood pressure and the other direct effects of smoking. This process (simulations of a particular node being made conditional on the simulated values for parent nodes) is continued until the response node is simulated. More details are given in Ferguson *et al.* (2020b).

Suppose then that upon elimination of a subset $\mathbf{S}$ of risk factors, the population distribution of all variables in the causal graph is $\mathbf{P_S}$, and via the recursive algorithm above, we have simulated new data $\mathbf{D_S}$ for all variables in the causal graph (exlcuding the response) under $\mathbf{P_S}$. Our estimate for (19) is then:

$$PAF_{\mathbf{S}} = \frac{\sum_{i \leq N}[w_i Y_i - w_i \hat{P}(Y_i = 1|\mathbf{D_S})]}{\sum_{i \leq N} w_i Y_i} \tag{20}$$

, where $\hat{P}(Y_i = 1|\mathbf{D_S})$ represents the estimated probability of disease for individual i under the simulated data structure for risk factors and confounders represented by $\mathbf{D_S}$ (this probability depends on $\mathbf{D_S}$ through the simulated values for individual i at those risk factors and covariates that are assumed to directly affect the outcome). This approach can be applied to cross-sectional and case-control datasets, where as before the argument `prev` is utilized to change the weighting in case-control datasets. Note that the above estimator may be randomized, that is estimating joint PAF twice using the same data may give slightly different results, since only differing simulated datasets $\mathbf{D_S}$ will be used in the differing estimates. This will generally be fine for large datasets and is recommended, although for smaller datasets the estimator (20) can be averaged over several independently simulated versions of $\mathbf{D_S}$ to reduce the variability due to the randomized nature of the estimator. In some cases, D_S may not vary over over differing simulations. For instance, for reasons described O'Connell and Ferguson (2022), continuous variables in $\mathbf{D_S}$ are simulated by adding model predicted residuals to the predicted values given the current values of their parents. As a result, randomness in $\mathbf{D_S}$ can only be generated by discrete risk factors or confounders that are graph-descendants of risk factors that are eliminated.

5.1. Data examples

The `joint_paf` function in graph PAF implements the procedure described above. As an example, suppose we are interested in estimating the joint PAF for stroke due to stress and blood pressure. First we need to specify the causal graph linking stress, blood pressure and stroke. In doing this, one must ensure that the confounders of any two nodes in the graph are also specified: for instance, any joint causes of stress and blood pressure must also be included. In Figure 5, we illustrate our assumed causal structure for INTERSTROKE risk factors, which inlcudes many confounders and risk factors other than stress and blood pressure. However, in the context of this estimation problem (and assuming Figure 5 is correct), we can give **graphPAF** a reduced causal structure: we actually don't need to specify pre-clinical disease variables `PCD` or physiology variables `P`, other than blood pressure, since they are not common causes of variables in the set (stress, blood pressure and stroke). In graphPAF we specify the causal graph with a list of the parents of all relevant variables in the graph (`parent_list`), together with a vector of variable names (corresponding to the nodes of the graph) (`node_vec`). When doing this it is important that `node_vec` and `parent_list` are in the same order. In addition, `node_vec` should be ordered so that parent nodes (that is

causes) are positioned in the vector before their children (that is their effects).

```
R> node_vec=c("exercise","diet","smoking","alcohol","stress",
+      "high_blood_pressure","case")
R> parents_exercise <- c("education")
R> parents_diet <- c("education")
R> parents_smoking <- c("education")
R> parents_alcohol <- c("education")
R> parents_stress <- c("education")
R> parents_high_blood_pressure <- c("education","exercise","diet",
+        "smoking","alcohol","stress")
R> parents_case <- c("education","exercise","diet","smoking",
+       "alcohol","stress","high_blood_pressure")
R> parent_list <- list(parents_exercise,parents_diet,parents_smoking,
+  parents_alcohol,parents_stress,parents_high_blood_pressure,parents_case)
```

Next, models for each variable (each time conditioning on its parents) need to be fit. In the context of joint PAF (as well as the sequential and average attributable fractions detailed in the following section), **graphPAF** supports simulation from linear models (fit using `lm`), logistic models (fit using `glm`) and ordinal logistic models (fit using `polr` from the R library **MASS**). Given that specification of multiple models can be time-consuming, **graphPAF** has a function `automatic_fit` that automatically fits additive models for each node in `node_vec`, conditioned on the parents of that node. This function can also fit non-linear relationships for continuous riskfactors or confounders using the `spline_nodes` argument. In the code below, `diet` is assumed to have a non-linear effect. Common interactions between variables that appear in all of the models can be specified by the argument `common`. However, in reality some of these models may require individual specification of interactions, in which case the models must be fit separately with either `lm`, `glm` or `polr`, before populating `model_list`. For case-control datasets, these models need to be fit with appropriate weighting (so that the weighted dataset set could be regarded as a representative sample) as described earlier. If `automatic_fit` is used, this can again be achieved automatically my specifying the `prev` argument. As mentioned earlier, weights can be calculated by passing the original dataset through `data_clean` if models are to be individually specified.

```
R> model_list=automatic_fit(data=stroke_reduced, parent_list=parent_list,
+      node_vec=node_vec, prev=.0035,common="region*ns(age,df=5)
+      +sex*ns(age,df=5)",spline_nodes=c("diet"))
```

Once `model_list` is specified, it can be passed to `joint_paf` for estimating joint PAF. Below, we compare estimated single risk factor attributable fractions for smoking and blood pressure to the joint attributable fraction for both smoking and blood pressure together. Note that the estimated joint attributable fraction (0.373) is slightly less than the sum of individual attributable fractions (0.116+0.268=0.384). This is actually expected Rowe, Powell, and Flanders (2004): as some of the disease cases that might be prevented in a population where nobody smokes would equally be prevented in a population where nobody was hypertensive. As mentioned earlier, `joint_paf` can average (20) over multiple independently estimated datasets using the argument `nsim`. However, since no discrete graph-descendants of `smoking` (other than `high_blood_pressure`) are specified in the causal graph specified in `joint_paf`, $\mathbf{D_S}$ will not vary over differing simulation iterates in this example.

```
R> joint_paf(data=stroke_reduced, model_list=model_list, parent_list=parent_list,
+            node_vec=node_vec,vars=c("smoking"), prev=.0035,ci=TRUE,boot_rep=50)

            est    bias debiased_est norm_lower norm_upper
joint PAF 0.113 0.00718        0.105     0.0857      0.125

R> joint_paf(data=stroke_reduced, model_list=model_list, parent_list=parent_list,
+            node_vec=node_vec,vars=c("high_blood_pressure"), prev=.0035,ci=TRUE,
+            boot_rep=50)

            est    bias debiased_est norm_lower norm_upper
joint PAF 0.269 0.00331        0.266      0.248      0.284

R> joint_paf(data=stroke_reduced, model_list=model_list, parent_list=parent_list,
+            node_vec=node_vec,vars=c("smoking","high_blood_pressure"), prev=.0035,
+            ci=TRUE,boot_rep=50)

            est     bias debiased_est norm_lower norm_upper
joint PAF 0.375 0.000738        0.374      0.349      0.399
```

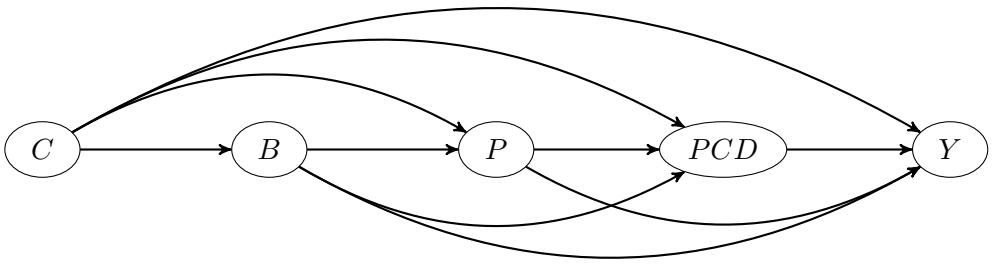


Figure 6: DAG showing causal structure linking risk factors at multiple levels. For the simulated INTERSTROKE dataset, we might assume that each node represents multiple risk factors as follows: C represents the Confounders: {`age`, `region`, `sex` and `education`}, B represents Behavioural risk factors: {`exer`,`alcohol`,`smoking`,`diet`,`exer`,`stress`}, P represents risk factors indicating physiology: {`high_blood_pressure`, `waist_hip_ratio`, `lipids`}, PCD represents pre-clinical disease: {`diabetes` and `early_stage_heart_disease`}. Y is a 0/1 indicator for stroke occurence

6. Sequential and Average PAF

Sequential attributable fractions (SAF), first described by Eide and Gefeller (1995) are closely related to joint attributable fractions as discussed in the previous section. They pertain to the incremental disease burden attributable to a risk factor (or more specificially to the removal of that risk factor from the population) in a population where a subset of risk factors have already been eliminated. Suppose that we number disease risk factors under consideration as: $\{1, ..., K\}$. We can define the sequential PAF for eliminating risk factor $j \leq K$, conditional on the subset of risk factors $\mathbf{S} \subset \{1, ..., K\} \setminus \{j\}$ already having being removed from

the population, as the difference in joint PAF pertaining to removing $\mathbf{S} \cup \{\mathbf{j}\}$ and the PAF pertaining to removing $\mathbf{S}$ alone:

$$PAF_{j|\mathbf{S}} = PAF_{\mathbf{S}\cup\{j\}} - PAF_{\mathbf{S}} \tag{21}$$

Given this link between joint and sequential PAF, the same issues (in particular risk factors of interest acting as confounders of causally downstream risk factors of interest) mentioned in the section above can also cause biases in estimating sequential PAF and average PAF. These can again be handled by recursive application of the do-operator and simulation from the corresponding distributions. Practically sequential PAF may be of interest if population health interventions are to be applied incrementally (for instance, what would be the next risk factor to target in a health intervention after a successful intervention that targets smoking?), but another use is in the definition and estimation of average population attributable fractions, again first introduced in Eide and Gefeller (1995).

As explained above, individual population attributable fractions for differing risk factors in a set are not expected to sum to the joint PAF corresponding to eliminating all risk factors in the set. Over the years, differing proposals have been made to contstruct versions of attributable fractions for individual risk factors that do form a partition for the joint PAF (that is they sum up to the corresponding joint PAF). The most convincing of convincing of these are average attributable fractions. Again suppose there are K risk factors, labeled again $\{1...K\}$. Imagine eliminating these K risk factors in some sequence. This can be done in $K!$ different ways. Each of these $K!$ permutations can be represented as $\sigma = \sigma(1), ..., \sigma(K)$, where $\sigma(j) = k$ if the risk factor k is the j^{th} risk factor eliminated according to that ordering, and as such each permutation is associated with a sequential PAF for each risk factor. For instance, in the previous example the sequential PAF for risk factor k according to σ would be $SAF_{k|\{\sigma(1)...\sigma(j-1)\}}$ if $j \geq 2$ or just the PAF for risk factor k if $j = 1$. The average PAF, $APAF_k$, for risk farctor k is the average of the sequential PAFs over all $K!$ different permutations. By definition, the sequential PAFs for differing risk factors corresponding to a particular permutation must add to the joint PAF. From this it follows easily that the average of these sequential PAFs for each risk factor across differing permutations (that is the average PAF) must also add over differing risk factors to the joint PAF.

6.1. Estimation

At first look, it seems that one must calculate $K!$ differing sequential PAFs to calculate average PAF for a risk factor. However, examining (21) we see that any sequential PAF is the difference between two differing joint PAFs. The number of joint PAF calculations is the same as the number of non-empty subsets of $\{1...K\}$ (that is $2^K - 1$, much smaller than $K!$). Provided the number of risk factors isn't too large (say 10 or fewer) this it is quite feasible to calculate all possible sequential PAFs utilizing this approach. Average PAF for risk factor $k \leq K$ can then be calculated using:

$$APAF_k = \frac{\sum_{j=1}^{K}(K-j)!(j-1)!\sum_{\mathbf{S}\subset\{1,...,K\}\setminus k:|\mathbf{S}|=j-1} PAF_{k|\mathbf{S}}}{K!}. \tag{22}$$

The 'exact' approach to estimating $APAF_k$ is to first estimate $PAF_{k|\mathbf{S}}$ for all possible subsets: $\mathbf{S} \subset \{1, ...K\} \setminus k$ of risk factors sets that exclude k, and then plug these estimates into (22). This is done most efficiently when calculating $APAF$ for all K risk factors together.

When 2^K is very large, estimating (22) exactly may be too time consuming. Recognizing instead that $APAF_k$ is a 'population' average of $K!$ sequential PAFs, each sequential PAF corresponding to a single permutation (with admittedly many of these permutations lead to the same SAF), one can approximate the $APAF$ by randomly sampling a smaller number `nperm` $< K!$ of permutations. Obviously, the larger `nperm` is, the smaller the approximation error from this step, which like any sample average decreases probabilistically at rate $\frac{1}{\sqrt{nperm}}$ as `nperm` increases. In practice, `nperm`= 1000 has been suggested to achieve acceptable accuracy Ferguson *et al.* (2018). Stratified sampling of permutations (ensuring for instance that each risk factor appears in position 1 in the elimination order an equal number of times in the `nperm` permutations) can somewhat reduce the approximation error. We will describe this in the next section.

6.2. Examples

Let's extend the example from earlier where we looked at the joint PAF for `smoking` and `high_blood_pressure`, to include a 3rd risk factor `diabetes`. Note that `lipids` and `waist_hip_ratio` are joint causes of diabetes and stroke (see Figure 5), and we now need to extend our causal graph and associated list of statistical models to include these variables in addition to `diabetes`.

```
R> node_vec=c("exercise","diet","smoking","alcohol","stress",
+      "high_blood_pressure","waist_hip_ratio","lipids","diabetes","case")
R> parents_exercise <- c("education")
R> parents_diet <- c("education")
R> parents_smoking <- c("education")
R> parents_alcohol <- c("education")
R> parents_stress <- c("education")
R> parents_high_blood_pressure <- c("education","exercise","diet",
+        "smoking","alcohol","stress")
R> parents_waist_hip_ratio <- c("education","exercise","diet",
+        "smoking","alcohol","stress")
R> parents_lipids <- c("education","exercise","diet",
+        "smoking","alcohol","stress")
R> parents_diabetes <- c("education","exercise","diet",
+        "smoking","alcohol","stress","high_blood_pressure","waist_hip_ratio","lipids")
R> parents_case <- c("education","exercise","diet","smoking","alcohol","stress",
+        "high_blood_pressure","waist_hip_ratio","lipids","diabetes")
R> parent_list <- list(parents_exercise,parents_diet,parents_smoking,
+    parents_alcohol,parents_stress,parents_high_blood_pressure,
+    parents_waist_hip_ratio,parents_lipids,parents_diabetes,parents_case)
```

Again we can automatically specify models using the `automatic_fit` function which now will fit models for the extra variables specified in `node_vec`. Note that `lipids` and `waist_hip_ratio` are also continuous risk factors and we can allow non-linear effects by adding these variable names to the `spline_nodes` argument.

```
R> model_list=automatic_fit(data=stroke_reduced, parent_list=parent_list,
+      node_vec=node_vec, prev=.0035,common="region*ns(age,df=5)+sex*ns(age,df=5)",
+      spline_nodes = c("waist_hip_ratio","lipids","diet"))
```

Single sequential PAFs can be estimated with the function `seq_paf`, which has the same structure as `joint_paf`. The most important argument is `vars`, a vector of risk factors. Sequential PAF is estimated for the risk factor specified in the last position of `vars` conditional on the risk factors in earlier positions. For instance, the code below estimates sequential PAF for eliminating `diabetes`, in a population where `smoking` and `high_blood_pressure` are already eliminated. As can be seen below, this estimator is randomized: the estimate varies slightly based on the simulated data. The reason for this is that now the discrete variable `diabetes` is included in the dataset: $\mathbf{D}_{\{\text{smoking, high\_blood\_pressure}\}}$, and the simulated value for `diabetes` under interventions for `smoking` and `high_blood_pressure` will vary slightly from simulation to simulation. Nevertheless as demonstrated below, the variation over simulation repetitions is fairly minimal and this variability will be accounted for in the Bootstrap confidence interval. Overall, this analysis suggests that in a population where `smoking` and `high_blood pressure` were already eliminated, an extra 2.4% of strokes (taken as a percentage of the number of strokes in the current population) might be prevented if there was no diabetes.

```
R> seq_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+          vars=c("smoking","high_blood_pressure","diabetes"),ci=FALSE,nsim=1)

[1] 0.02382662

R> seq_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+          vars=c("smoking","high_blood_pressure","diabetes"),ci=FALSE,nsim=1)

[1] 0.02267426

R> seq_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+          vars=c("smoking","high_blood_pressure","diabetes"),ci=TRUE,nsim=1,
+          boot_rep=50)

                 est     bias debiased_est norm_lower norm_upper
sequential PAF 0.024 2.94e-05        0.024     0.0156     0.0323
```

The function `average_paf` generates results for average PAF for the three risk factors: `smoking`, `high_blood_pressure` and `diabetes`. The results are stored in a S3 object of class `SAF_summary`, which can be printed or plotted using print and plot. The default estimation method is to first estimate joint PAF for all possible risk factor subsets, $\mathbf{S}$, next to estimate all sequential PAFs, $PAF_{j|\mathbf{S}}$, from the vector of joint PAFs and finally substitute these estimated sequential PAF into (22). Recall that in estimating joint PAF for the risk factor set $\mathbf{S}$, a data set $D_{\mathbf{S}}$ corresponding to this joint intervention is simulated recursively. The recursive nature of this simulation can be exploited to perform the estimation of all 2^K joint PAFs efficiently. For instance, when simulating data: $D_{\mathbf{S}\cup\{j\}}$ corresponding to eliminating risk factors: $\mathbf{S}\cup\{j\}$, with j being the final risk factor eliminated, data corresponding to eliminating the risk factors in $\mathbf{S}$, $D_{\mathbf{S}}$ has already been simulated. `average_paf` calculates joint PAF for the 2^K risk factor subsets in an order that allows extensive use of this fact. As illustrated in the results below, estimated average PAF is highest for `high_blood_pressure` at 0.259, with `smoking` at 0.106 and `diabetes` at 0.0395. These three quantities sum to estimated joint PAF, 0.405, as expected. In addition, average sequential PAF by elimination position for each risk factor is provided. Note that the sequential PAF for diabetes is most

effected by elimination position. This makes sense based on its position in the causal graph (causally upstream of `smoking` and `high_blood_pressure`)

```
R> out <- average_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+    vars=c("smoking","high_blood_pressure","diabetes"),ci=FALSE,exact=TRUE)
R> print(out)

                 position         risk factor   estimate
1  elimination position 1             smoking 0.10530974
2  elimination position 2             smoking 0.10030958
3  elimination position 3             smoking 0.10349950
4  elimination position 1 high_blood_pressure 0.27667668
5  elimination position 2 high_blood_pressure 0.26135515
6  elimination position 3 high_blood_pressure 0.25422371
7  elimination position 1            diabetes 0.05407963
8  elimination position 2            diabetes 0.03472414
9  elimination position 3            diabetes 0.02355873
10               Average             smoking 0.10303961
11               Average high_blood_pressure 0.26408518
12               Average            diabetes 0.03745417
13                 Joint                     0.40457895
```

In the above analysis, the estimator is again randomized. While all $2^K - 1$ joint PAFs need to be estimated to enable this calculation for all risk factors; each estimated joint PAF corresponds to a single simulated data set $D_{\mathbf{S}}$, which can generate substantial Monte Carlo variability for small datasets and small K. As an alternative, one can sample `nperm` differing permutations of $\{1, ..., K\}$: corresponding to differing risk factor elimination orders, calculate sequential PAFs associated with each permutation and average the associated sequential PAF for a particular risk factor. For small K and `nperm` $> 2^K$ this approach is likely to have reduced Monte Carlo error (compared to the 'exact' approach), despite sampling permutations. If the argument `exact=FALSE`, this approach is used in place of the estimator based on (22). Stratified sampling of permutations (so that the joint empirical distribution of permutation positions $\sigma(1), ..., \sigma(S)$ for some $S < K$ is uniform (as it would be if we calculated sequential PAFs for all $K!$ permutations), can help further reduce Monte Carlo error. For K risk factors, an integer multiple of $K(K-1)...(K-S+1)$ permutations are needed to implement such a strategy. Such stratified sampling of permutations is implemented through the argument `correct_order` (`correct_order`=S in the preceding example). Here there only $K = 3$ risk factors, and averaging all 3! permutations sampled with `correct_order`=2 generates an exact calculation of average PAF (similar to (22)). For larger K, averaging sequential PAF over a number of sampled permutations `nperm`$\approx 2^K$ or `nperm`$< 2^K$ may be less accurate than (22) due to the Monte Carlo error associated with sampling permutations. When confidence intervals are not requested an upper bound on the margin of error of the point estimate (in terms of how close to the calculation with `nperm` $= \infty$) is given (with 95%) confidence as calculated in Ferguson *et al.* (2018), provided permutations are sampled (`exact=FALSE`). Note that this margin of error assumes non-stratified sampling rather than the more accurate stratified sampling implemented here. The results below indicate that the three average PAFs are calculated to within an accuracy of 0.004 (with 95% confidence) compared to the exact estimate when `nperm` $\to \infty$.

```
R> out <- average_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+    vars=c("smoking","high_blood_pressure","diabetes"),ci=FALSE,exact=FALSE, correct_order=2, nperm=60)
R> print(out)
```

```
                position         risk factor   estimate Margin error lower bound Upper bound
1  elimination position 1             smoking 0.10441103 2.786493e-03  0.10162454  0.10719752
2  elimination position 2             smoking 0.10406168 1.177682e-03  0.10288399  0.10523936
3  elimination position 3             smoking 0.10349950 0.000000e+00  0.10349950  0.10349950
4  elimination position 1 high_blood_pressure 0.27582328 6.291276e-04  0.27519415  0.27645241
5  elimination position 2 high_blood_pressure 0.26066718 6.768796e-03  0.25389838  0.26743598
6  elimination position 3 high_blood_pressure 0.24764498 2.129574e-03  0.24551541  0.24977455
7  elimination position 1            diabetes 0.05407963 0.000000e+00  0.05407963  0.05407963
8  elimination position 2            diabetes 0.03942847 6.971034e-03  0.03245744  0.04639951
9  elimination position 3            diabetes 0.02412110 3.563301e-04  0.02376477  0.02447743
10                Average             smoking 0.10399074 9.525977e-04  0.10303814  0.10494333
11                Average high_blood_pressure 0.26137848 3.738582e-03  0.25763990  0.26511706
12                Average            diabetes 0.03920974 3.864373e-03  0.03534536  0.04307411
13                  Joint                     0.40457895 5.903706e-18  0.40457895  0.40457895
```

Of courses, sampling error also needs to be accounted for when making a statement about estimation accuracy. Confidence intervals (with the Bootstrap) suggest comparable accuracy overall with the method (that estimates all average PAF just from 8 joint PAF) and also with the slower method which uses 60 permutations to calculate average PAF.

```
R> full_results_a <- average_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+    vars=c("smoking","high_blood_pressure","diabetes"),exact=TRUE, ci=TRUE,boot_rep=50)
R> print(full_results_a)
```

```
                position         risk factor    est      bias debiased_est norm_lower norm_upper
1  elimination position 1             smoking 0.1100 -0.002300       0.1130     0.0918     0.1330
2  elimination position 2             smoking 0.1060  0.000482       0.1050     0.0892     0.1210
3  elimination position 3             smoking 0.1030  0.000395       0.1030     0.0890     0.1170
4  elimination position 1 high_blood_pressure 0.2750 -0.001090       0.2760     0.2570     0.2960
5  elimination position 2 high_blood_pressure 0.2580 -0.000116       0.2590     0.2360     0.2810
6  elimination position 3 high_blood_pressure 0.2440 -0.002010       0.2460     0.2180     0.2740
7  elimination position 1            diabetes 0.0541  0.000731       0.0533     0.0354     0.0713
8  elimination position 2            diabetes 0.0379  0.001700       0.0362     0.0203     0.0522
9  elimination position 3            diabetes 0.0244 -0.000195       0.0246     0.0153     0.0339
10                Average             smoking 0.1060 -0.000475       0.1070     0.0916     0.1220
11                Average high_blood_pressure 0.2590 -0.001070       0.2600     0.2380     0.2830
12                Average            diabetes 0.0388  0.000746       0.0381     0.0243     0.0518
13                  Joint                 PAF 0.4050 -0.000801       0.4050     0.3790     0.4320
```

```
R> full_results_b <- average_paf(stroke_reduced,model_list,parent_list,node_vec,prev=0.0035,
+    vars=c("smoking","high_blood_pressure","diabetes"),ci=TRUE,exact=FALSE,boot_rep=50,
+    correct_order=2, nperm=60)
R> print(full_results_b)
```

```
                position         risk factor    est      bias debiased_est norm_lower norm_upper
1  elimination position 1             smoking 0.1040  9.29e-04       0.1030     0.0815     0.1250
2  elimination position 2             smoking 0.1040  5.65e-04       0.1040     0.0862     0.1210
3  elimination position 3             smoking 0.1030  2.37e-05       0.1030     0.0881     0.1190
4  elimination position 1 high_blood_pressure 0.2760  4.34e-05       0.2760     0.2550     0.2970
5  elimination position 2 high_blood_pressure 0.2620 -3.45e-04       0.2620     0.2400     0.2840
6  elimination position 3 high_blood_pressure 0.2460 -2.25e-04       0.2460     0.2230     0.2690
7  elimination position 1            diabetes 0.0541  6.78e-04       0.0534     0.0371     0.0697
8  elimination position 2            diabetes 0.0396  3.95e-04       0.0392     0.0261     0.0522
9  elimination position 3            diabetes 0.0242  7.41e-05       0.0241     0.0155     0.0327
10                Average             smoking 0.1040  5.06e-04       0.1040     0.0855     0.1220
11                Average high_blood_pressure 0.2610 -1.76e-04       0.2610     0.2400     0.2830
12                Average            diabetes 0.0393  3.82e-04       0.0389     0.0263     0.0515
13                  Joint                 PAF 0.4050  7.13e-04       0.4040     0.3770     0.4310
```

Results (average PAF and sequential PAF by elimination position, along with associated variability bands) can be plotted over differing risk factors as follows:

```
R> plot(full_results_b,1, 0.35)
```

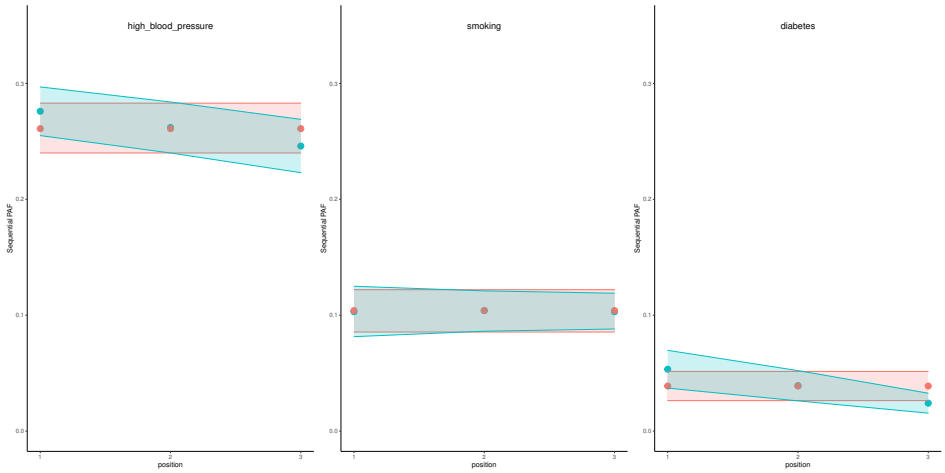


Figure 7: Estimated average PAF and sequential PAF for the group of risk factors: smoking, blood pressure and diabetes. Risk factors are plotted in decreasing order of estimated $APAF$. Average PAF is shaded in pink, and the average sequential PAF for particular risk factors by elimination position in blue. One expects sequential PAF to decrease over elimination position which is what we observe here.

Note that if `exact='FALSE'` and `ci='FALSE'`, the plotted variability bands will not be interpretable as confidence intervals, but rather as bands for the degree of possibility approximation error in the point estimate.

6.3. Computational considerations

As described here, **graphPAF**, facilitates incorporation of causal structure into estimation of joint, sequential and average PAF, essentially by incorporating recursive simulation methods based on an assumed causal structure. Ignoring such causal structure, as other approaches have in the past (for example, Rückinger, von Kries, and Toschke (2009), Ferguson *et al.* (2018)) may lead to bias. A drawback of theis simulation based strategy is computational cost. Techniques such as Bootstrap-parallelization (through the `boot` library), intelligent ordering of calculations when calculating joint PAF for differing risk factor subsets, stratified sampling of permutations when the number of risk factors is large and the use of the more efficient formula for average PAF (22) can somewhat reduce these computational requirements. Computational cost depends jointly the size of the Bayesian networks and the size of the underlying dataset. The dataset `stroke_reduced` used in this manuscript has 13,712 rows and the algorithms described here can be run in reasonable time on most modern laptops when using this data. For larger datasets, splitting into independent subsets and rerunning these methods independently on each subset before averaging might be recommended to avoid memory management problems.

7. Conclusions

In addition to implementing standard PAF estimation, **graphPAF** collates many recently developed tools for estimation of disease burden in non-standard settings into one package. We hope it will be useful to statisticians and epidemiologists who are interested in comparisons of disease burden over multiple risk factors, both discrete and continuous.

Acknowledgments

Dr. Ferguson and Dr. O'Connell are supported by the grant EIA-017-2017 from the Health Research Board in Ireland

References

Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C (1985). "Estimating the population attributable risk for multiple risk factors using case-control data." *American journal of epidemiology*, **122**(5), 904–914.

Bulterys M, Morgenstern H, Weed D (1997). "Quantifying the expected vs potential impact of a risk-factor intervention program." *American journal of public health*, **87**(5), 867–868.

Camacho-García-Formentí D, Zepeda-Tello R (2019). "Vignette: Introduction to the pipaf package: Introduction to the pifpaf package."

Chen L, Lin D, Zeng D (2010). "Attributable fraction functions for censored event times." *Biometrika*, **97**(3), 713–726.

Dahlqwist E, Zetterqvist J, Pawitan Y, Sjölander A (2016). "Model-based estimation of the attributable fraction for cross-sectional, case–control and cohort studies using the R package AF." *European journal of epidemiology*, **31**(6), 575–582.

Eide GE, Gefeller O (1995). "Sequential and average attributable fractions as aids in the selection of preventive strategies." *Journal of clinical epidemiology*, **48**, 645–655.

Ferguson J, Alvarez-Iglesias A, Newell J, Hinde J, O'Donnell M (2018). "Estimating average attributable fractions with confidence intervals for cohort and case–control studies." *Stat Methods Med Res*, **27**, 1141–1152.

Ferguson J, Maturo F, Yusuf S, O'Donnell M (2020a). "Population attributable fractions for continuously distributed exposures." *Epidemiologic Methods*, **9**(1).

Ferguson J, O'Connell M, O'Donnell M (2020b). "Revisiting sequential attributable fractions." *Arch Public Health*, **78**, 1–9.

Ferguson J, O'Leary N, Maturo F, Yusuf S, O'Donnell M (2019). "Graphical comparisons of relative disease burden across multiple risk factors." *BMC Med Res Methodol*, **19**, 186.

Laaksonen MA, Härkänen T, Knekt P, Virtala E, Oja H (2010). "Estimation of population attributable fraction (PAF) for disease occurrence in a cohort study design." *Statistics in medicine*, **29**(7-8), 860–874.

Louis Schenck Elizabeth Atkinson CCTT (2014). "Attribrisk package."

O'Connell M, Ferguson J (2021). "causalPAF." R package, version 1.2.5.

O'Connell M, Ferguson J (2022). "Pathway-specific population attributable fractions." *International Journal of epidemiology*, **10:dyac079. doi: 10.1093/ije/dyac079. Epub ahead of print.**

O'Donnell MJ, Chin SL, Sumathy R, et al (2016). "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study." *Lancet*, **388**, 761–775.

Pearl J (2009). *Causality*. Cambridge: Cambridge university press.

Rowe AK, Powell KE, Flanders WD (2004). "Why population attributable fractions can sum to more than one." *Am J Prev Med*, **26**, 243–249.

Rubin DB (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, **66**(5), 688.

Rückinger S, von Kries R, Toschke AM (2009). "An illustration of and programs estimating attributable fractions in large scale surveys considering multiple risk factors." *BMC medical research methodology*, **9**(1), 1–6.

Sinha A, Ning H, Carnethon MR, Allen NB, Wilkins JT, Lloyd-Jones DM, Khan SS (2021). "Race-and sex-specific population attributable fractions of incident heart failure: a population-based cohort study from the lifetime risk pooling project." *Circulation: Heart Failure*, **14**(4), e008113.

Sjölander A (2018). "Estimation of causal effect measures with the R-package stdReg." *European journal of epidemiology*, **33**(9), 847–858.

Therneau T, Crowson C, Atkinson E (2021). "Multi-state models and competing risks." *Technical report*, Mayo Clinic.

Affiliation:

John Ferguson
Biostatistics Unit
HRB Clinical Research Facility
University of Galway
Ireland
E-mail: john.ferguson@universityofgalway.ie