

Model-based estimation of the attributable fraction for cross-sectional, case–control and cohort studies using the R package AF

Elisabeth Dahlqwist¹  · Johan Zetterqvist¹ ·
Yudi Pawitan¹ · Arvid Sjölander¹

Received: 17 December 2015 / Accepted: 3 March 2016 / Published online: 18 March 2016
© Springer Science+Business Media Dordrecht 2016

Abstract The attributable fraction (or attributable risk) is a widely used measure that quantifies the public health impact of an exposure on an outcome. Even though the theory for AF estimation is well developed, there has been a lack of up-to-date software implementations. The aim of this article is to present a new R package for AF estimation with binary exposures. The package AF allows for confounder-adjusted estimation of the AF for the three major study designs: cross-sectional, (possibly matched) case–control and cohort. The article is divided into theoretical sections and applied sections. In the theoretical sections we describe how the confounder-adjusted AF is estimated for each specific study design. These sections serve as a brief but self-consistent tutorial in AF estimation. In the applied sections we use real data examples to illustrate how the AF package is used. All datasets in these examples are publicly available and included in the AF package, so readers can easily replicate all analyses.

Keywords Attributable fraction · Attributable risk · Public health · R package · Regression model · Confounder-adjusted · Statistical software

1 Introduction

One of the main goals in public health research is to evaluate the disease burden due to a specific exposure. For this purpose, the attributable fraction (AF) is commonly used. The AF was initially defined in the 1950s and has been used extensively since then in epidemiological studies. A thorough

historical review of the AF is given by Poole [1]. Originally, the AF was defined for binary outcomes as the proportion of unfavourable outcomes that would have been prevented if the exposure of interest were eliminated from the population [2]. As such, the AF takes both the exposure–outcome association and the exposure prevalence into account, and is specific to the study population. To formalize, let X and Y be the exposure and outcome of interest. In standard counterfactual notation, the AF is defined as

$$AF = 1 - \frac{\Pr(Y_0 = 1)}{\Pr(Y = 1)}, \quad (1)$$

where $\Pr(Y = 1)$ is the factual outcome prevalence, and $\Pr(Y_0 = 1)$ is the counterfactual outcome prevalence had the exposure been eliminated (set to 0) for everyone. For instance, if the factual outcome prevalence is 10 % and the counterfactual outcome prevalence is 5 %, then $1 - 0.05/0.1 = 50 %$ of all outcomes would have been prevented, had the exposure been eliminated. We note that the exposure doesn't have to be binary per se, but the definition in Eq. (1) assumes that there is a 'zero-level' for the exposure, corresponding to the exposure being completely absent.

Recently, the AF has been extended to time-to-event outcomes [3–5]. Let T be the time-to-event of interest, e.g. time to death. The AF function is then defined as

$$AF(t) = 1 - \frac{\Pr(T_0 \leq t)}{\Pr(T \leq t)}, \quad (2)$$

where $\Pr(T \leq t)$ is the factual probability of an event at or before time t , and $\Pr(T_0 \leq t)$ is the counterfactual probability of an event at or before time t had the exposure been eliminated for everyone at baseline.

To estimate the AF from observational data, it is important to adjust for confounders for the exposure–outcome asso-

✉ Elisabeth Dahlqwist
elisabeth.dahlqwist@ki.se

¹ Karolinska Institute, Nobels väg 12A, 171 77 Stockholm, Sweden

ciation. If a covariate set \mathbf{Z} is sufficient for confounding control, then the AF can be consistently estimated by adjusting for \mathbf{Z} . For binary outcomes in cross-sectional studies, the confounder-adjusted AF can be estimated with logistic regression [6–9]. For binary outcomes in (possibly matched) case–control studies, the confounder-adjusted AF can also be estimated with logistic regression, under a ‘rare-disease’ assumption [10, 11]. For time-to-event outcomes in cohort studies, the confounder-adjusted AF function can be estimated with Cox proportional hazard regression [4].

Even though the theory for AF estimation is well developed, there is still a lack of up-to-date software implementations. In this article we focus on the open-source statistical software R [12]. To our knowledge there are three earlier packages for AF estimation available at CRAN: `epiR` [13], `attribrisk` [14] and `paf` [15]. These packages all have important limitations. The `epiR` package uses the function `epi.2by2` to estimate the AF for various sampling designs, but does not allow for model-based confounder adjustment. The `attribrisk` package allows for confounder adjustment through logistic regression, but it relies on the ‘rare-disease’ assumption and is thus essentially restricted to case–control studies. Furthermore, the `attribrisk` package only provides bootstrap and jackknife standard errors, which makes it relatively time consuming. The `paf` package estimates the AF function using Cox proportional hazard regression for confounder adjustment. However, it does not handle big data (in our simulations it breaks down for data with around 20,000 observations or more). A common limitation of all three packages is that none of them provides accurate standard errors when data are clustered, e.g. when there are repeated measures on each subject.

The aim of this article is to present a new R package for AF estimation. This new package `AF` allows for confounder-adjusted estimation of the AF for the three major study designs: cross-sectional, (possibly matched) case–control and cohort. It provides analytical standard errors for all estimates, which obviates the need for bootstrapping. When data are clustered, these standard errors are adjusted for the within-cluster correlations. The package is designed to scale up, so that it is able to handle very large datasets (up to several millions of observations).

The article is organized as follows. In Sects. 2, 3 and 4 we describe how the `AF` package is used to estimate the AF in cross-sectional, case–control and cohort studies, respectively. Each section is divided into a theoretical part and an applied part. In the theoretical sections we describe how the confounder-adjusted AF is estimated for each specific study design. These sections serve as a brief but self-consistent tutorial in AF estimation. In the applied sections we use real data examples to illustrate how the `AF` package is used. All datasets in these examples are publicly available and included in the `AF` package, so readers can easily replicate all analyses.

2 Cross-sectional study and cohort study with binary outcome

2.1 Theory

In cross-sectional studies with binary outcomes, the AF is defined as in Eq. (1). The factual outcome prevalence, $\Pr(Y = 1)$, can be estimated as the observed (sample) outcome prevalence. To estimate the counterfactual outcome prevalence, $\Pr(Y_0 = 1)$, it is usually assumed that a set of observed covariates \mathbf{Z} is sufficient for confounding control. Under this assumption, $\Pr(Y_0 = 1)$ can be obtained by averaging the outcome prevalence among the unexposed, at a given value of \mathbf{Z} , over the population distribution of \mathbf{Z} :

$$\Pr(Y_0 = 1) = E\{\Pr(Y = 1 \mid X = 0, \mathbf{Z})\}.$$

In practice, $\Pr(Y = 1 \mid X = 0, \mathbf{Z})$ is usually estimated with a logistic regression model

$$\text{logit}\{\Pr(Y = 1 \mid X, \mathbf{Z})\} = g(X, \mathbf{Z}; \boldsymbol{\beta}), \quad (3)$$

where $g()$ is a specified function indexed by the parameter vector $\boldsymbol{\beta}$. For example, $g()$ could be specified as $\beta_0 + \beta_1 X + \beta_2 \mathbf{Z}$. However, $g()$ could also involve interactions and higher order terms. The model in Eq. (3) is fitted to obtain an estimate of $\boldsymbol{\beta}$. Then, for each subject i with covariate vector \mathbf{Z}_i we use $\text{expit}\{g(X = 0, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})\}$ as a prediction of $\Pr(Y = 1 \mid X = 0, \mathbf{Z}_i)$. These predictions are averaged to obtain an estimate of $\Pr(Y_0 = 1)$:

$$\widehat{Pr}(Y_0 = 1) = \frac{1}{n} \sum_{i=1}^n \text{expit}\{g(X = 0, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})\}. \quad (4)$$

The estimates of $\Pr(Y = 1)$ and $\Pr(Y_0 = 1)$ are plugged into Eq. (1), to produce an estimate of the AF. The standard error for the resulting estimate can be obtained by combining the sandwich formula with the delta method [8, 9].

We end this section by noting that neither the definition in Eq. (1), nor the estimation procedure described in this section, requires a cross-sectional study design per se. For instance, they are also applicable in cohort studies with time-to-event outcomes, if the outcome is dichotomized as having the event before a fixed time point, e.g. 5 years from baseline. However, when censoring is present, as is often the case in cohort studies, it more natural to use the time-to-event analysis described in Sect. 4.

2.2 Applied example

To illustrate the theory we use a dataset on 487 births among 188 women described in Juul and Frydenberg [16]. For each birth, the following variables are measured: parity (`birth`),

a binary indicator of whether the mother smoked during pregnancy (`smoker`), race of the mother (`race`: white, black or other), age of the mother (`age`), a unique identification number for each mother (`id`), weight of the mother at last menstrual period in pounds (`lwt`), birthweight of the newborn child (`bwt`), and a binary indicator of whether the newborn has low birthweight (defined as birthweight smaller or equal to 2500 grams) (`lbw`). These variables are stored in the data frame `clslowbwt`, which is included in the `AF` package.

We are interested in the effect of smoking during pregnancy on the child's birthweight. We will adjust for age and race, since both these variables are potential confounders. Initially, we assume the following standard logistic regression model:

$$\begin{aligned} \text{logit}\{\text{Pr}(\text{lbw} = 1 \mid \text{smoker}, \text{age}, \text{race})\} \\ = \beta_0 + \beta_1 \text{smoker} + \beta_2 I(\text{race} = \text{"black"}) \\ + \beta_3 I(\text{race} = \text{"other"}) + \beta_4 \text{age}. \end{aligned} \tag{5}$$

To estimate the AF under this model we can use the `AF.cs` function in the `AF` package. This function fits the model 'internally', and then outputs the estimated AF. However, for illustrational purpose we first fit the model separately, and discuss the output.

In R, we fit the model in Eq. (5) with the `glm` function by typing

```
fit <- glm(formula = lbw ~ smoker + race + age, family = binomial, data = clslowbwt)
```

Summarizing the output gives

```
> summary(fit)

Call:
glm(formula = lbw ~ smoker + race + age, family = binomial, data = clslowbwt)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.2326  -0.8936  -0.6491   1.2249   1.9808

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.35946    0.53281  -2.551  0.01073 *
smoker       0.60080    0.21693   2.770  0.00561 **
race2. Black -0.85852    0.32331  -2.655  0.00792 **
race3. Other -0.70449    0.24624  -2.861  0.00422 **
age          0.02399    0.01785   1.344  0.17900
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 603.05  on 486  degrees of freedom
Residual deviance: 570.53  on 482  degrees of freedom
AIC: 580.53

Number of Fisher Scoring iterations: 4
```

The output indicates that the odds of getting a low birthweight is about $e^{0.6} \approx 1.8$ times higher among children born to smokers as compared to children born to non-smokers. The effect is highly significant, with a p-value equal to 0.00561. To estimate the proportion of low birthweights that would

have been prevented if no mother had smoked during pregnancy we use the `AF.cs` function.

```
> AFest <- AF.cs(formula = lbw ~ smoker + race + age,
data = clslowbwt, exposure = "smoker")
```

Like the `glm` function, the `AF.cs` function has a `formula` argument and a `data` argument. Since the outcome is by definition binary in Eq. (1), the `AF.cs` function always uses logistic regression, and thus has no `family` argument. The name of the exposure variable is specified by the `exposure` argument. Summarizing the output gives

```
> summary(AFest)

Call:
AF.cs(formula = lbw ~ smoker + race + age, data = clslowbwt,
exposure = "smoker")

Estimated attributable fraction (AF)and untransformed 95% Wald CI:

      AF Std.Error z value Pr(>|z|) Lower limit Upper limit
0.1697446 0.06357109 2.670154 0.007581644 0.04514756 0.2943417

Exposure : smoker
Outcome   : lbw

Observations Cases
      487      151

Method for confounder adjustment: Logistic regression

Formula: lbw ~ smoker + race + age
```

The output indicates that approximately 17% of all low birthweights would have been prevented if no mother had smoked during pregnancy. The AF is highly significant, with a p-value equal to 0.0076. However, the 95% CI is quite wide, ranging from 5 to 29%. The default CI is untransformed, but the `summary` function also allows for log- and logit-transformed CIs, which sometimes have more accurate coverage probabilities [17].

There are two problems with the analysis above. First, children born to the same mothers are correlated, which is not accounted for in the standard errors and p-values. Second, the model in Eq. (5) is perhaps unrealistically simple, since it assumes that the effect of smoking is the same regardless of age and race. Thus, we next consider the following model, which allows for interactions between all predictors:

$$\begin{aligned} \text{logit}\{\text{Pr}(\text{lbw} = 1 \mid \text{smoker}, \text{age}, \text{race})\} = \beta_0 + \beta_1 \text{smoker} \\ + \beta_2 I(\text{race} = \text{"black"}) + \beta_3 I(\text{race} = \text{"other"}) + \beta_4 \text{age} \\ + \beta_5 \text{smoker} \times I(\text{race} = \text{"black"}) + \beta_6 \text{smoker} \times I(\text{race} = \text{"other"}) \\ + \beta_7 \text{smoker} \times \text{age} \\ + \beta_8 I(\text{race} = \text{"black"}) \times \text{age} + \beta_9 I(\text{race} = \text{"other"}) \times \text{age}. \end{aligned} \tag{6}$$

The `glm` function has no facilities for handling clustered data. Thus, we instead fit the model with the `gee` function from the `drgee` package:

```
> fit <- gee(formula = lbw ~ (smoker + race + age)^2, link = "logit",
data = clslowbwt, clusterid = "id")
```

By typing $(\text{smoker} + \text{race} + \text{age})^2$, the formula automatically constructs all possible main effects and interactions between `smoker`, `race` and `age`. By specifying the `clusterid` argument, cluster-robust standard errors are calculated. Summarizing the output gives

```
> summary(fit)

Call: gee(formula = lbw ~ (smoker + race + age)^2, link = "logit",
data = clslowbwt, clusterid = "id")

Model: lbw ~ smoker + race2. Black + race3. Other + age + smoker:race2. Black +
smoker:race3. Other + smoker:age + race2. Black:age + race3. Other:age

Link function: logit

              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.91927    1.06155    1.808 0.070607 .
smoker         -3.81160    1.28115   -2.975 0.002928 **
race2. Black   -4.79348    1.99482   -2.403 0.016263 *
race3. Other   -2.99103    1.47110   -2.033 0.042033 *
age            -0.08611    0.03586   -2.401 0.016335 *

smoker:race2. Black  1.43758    0.80736    1.781 0.074977 .
smoker:race3. Other  1.00778    0.77086    1.307 0.191093
smoker:age          0.14763    0.04482    3.294 0.000988 ***
race2. Black:age    0.10954    0.06123    1.789 0.073625 .
race3. Other:age    0.06936    0.05244    1.323 0.185915

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

487 complete observations used

Cluster-robust Std. errors
188 clusters
```

We observe that the interaction term between `smoker` and `age` is highly significant, with a p-value < 0.001 . Thus, it seems like the model in Eq. (5), which only includes main effects, may indeed be too simplistic. On the other hand, the more realistic model in Eq. (6) is harder to interpret and communicate, since the smoking effect in this model is captured by four parameters (one main effect plus three interaction terms). This is not a problem for the AF though, since the AF is always a single number, regardless of the complexity of the underlying model. Estimating the AF, together with cluster-robust standard errors, gives:

```
> AFest <- AF.cs(lbw ~ (smoker + race + age)^2, data = clslowbwt,
exposure = "smoker", clusterid = "id")
> summary(AFest)

Call:
AF.cs(formula = lbw ~ (smoker + race + age)^2, data = clslowbwt,
exposure = "smoker", clusterid = "id")

Estimated attributable fraction (AF)and untransformed 95% Wald CI:

      AF Robust SE  z value Pr(>|z|) Lower limit Upper limit
0.1006186 0.0943206 1.066772 0.2860749 -0.08424642  0.2854836

Exposure : smoker
Outcome   : lbw

Observations Cases Clusters
      487      151      188

Method for confounder adjustment: Logistic regression

Formula: lbw ~ (smoker + race + age)^2
```

The estimated AF is now reduced to 10%, and not statistically significant. At first glance, it may appear surprising that the AF is not statistically significant, given that both the main effect of `smoker` and the interaction between `smoker`

and `age` are statistically significant in the `gee` output. There are two reasons for this discrepancy. First, the main effect of `smoker` is negative, but the `smoker`-`age` interaction is positive. This means that for young women, smoking appears to decrease the risk of low birthweight in the newborn, whereas for old women smoking appears to increase the risk; for white women the ‘switch’ occurs at $3.8116/0.14763 = 25.8$ years. Thus, when averaging over all ages, as in the AF, these effects may cancel out. Second, the statistical uncertainty in the AF does not only depend on the uncertainty in the model parameters, but also on the sampling variability in the distribution of the confounders. To see this, note that even if we would replace the estimated parameter $\hat{\beta}$ in Eq. (4) with the true value β , there would be residual uncertainty in $\widehat{Pr}(Y_0 = 1)$ due to the sampling variability in \mathbf{Z}_i when estimating the distribution of \mathbf{Z}_i with the empirical distribution.

3 Case-control study

3.1 Theory

Case-control sampling distorts the outcome distribution, so that $\Pr(Y = 1)$ and $\Pr(Y_0 = 1)$ in Eq. (1) cannot be separately estimated from the data. However, by assuming that the covariate vector \mathbf{Z} is sufficient for confounding control, and applying Bayes’ theorem [10, 11], the AF can be reformulated as

$$\text{AF} = 1 - E\{\text{RR}(\mathbf{Z})^{-X} | Y = 1\} \quad (7)$$

where

$$\text{RR}(\mathbf{Z}) = \frac{\Pr(Y = 1 | X = 1, \mathbf{Z})}{\Pr(Y = 1 | X = 0, \mathbf{Z})} \quad (8)$$

is the conditional risk ratio, given \mathbf{Z} . If the outcome is rare, then the risk ratio $\text{RR}(\mathbf{Z})$ can be approximated by the conditional odds ratio

$$\text{OR}(\mathbf{Z}) = \frac{\Pr(Y = 1 | X = 1, \mathbf{Z})\Pr(Y = 0 | X = 0, \mathbf{Z})}{\Pr(Y = 0 | X = 1, \mathbf{Z})\Pr(Y = 1 | X = 0, \mathbf{Z})}. \quad (9)$$

The AF is thus approximately equal to $1 - E\{\text{OR}(\mathbf{Z})^{-X} | Y = 1\}$. Estimation of the AF proceeds as follows. First, a logistic regression model is fitted to the data. Then, for each subject i with covariate vector \mathbf{Z}_i the model is used to estimate $\text{OR}^{-X_i}(\mathbf{Z}_i)$. For exposed subjects (those with $X_i = 1$), $\text{OR}^{-X_i}(\mathbf{Z}_i) = \text{OR}^{-1}(\mathbf{Z}_i)$. For unexposed subjects (those with $X_i = 0$), $\text{OR}^{-X_i}(\mathbf{Z}_i) = 1$. The predictions of $\text{OR}^{-X_i}(\mathbf{Z}_i)$ are then averaged among the cases (those with $Y_i = 1$), to produce an estimate of AF:

$$\widehat{AF} = 1 - \frac{\sum_{i=1}^n Y_i OR^{-X_i}(\mathbf{Z}_i)}{\sum_{i=1}^n Y_i}$$

For instance, if we assume a logistic model without interactions between X and \mathbf{Z}

$$\text{logit}\{\text{Pr}(Y = 1|X, \mathbf{Z})\} = \beta_0 + \beta_1 X + \beta_2 \mathbf{Z}, \tag{10}$$

then $OR(\mathbf{Z}) = e^{\beta_1}$. It follows that \widehat{AF} simplifies to

$$\widehat{AF} = 1 - \frac{\sum_{i=1}^n Y_i e^{-\hat{\beta}_1 X_i}}{\sum_{i=1}^n Y_i} = \hat{P}_r(X = 1|Y = 1)(1 - e^{-\hat{\beta}_1}), \tag{11}$$

where $\hat{P}_r(X = 1|Y = 1)$ is the sample proportion of exposed among the cases. If we assume a more complicated model that allows for interactions between X and \mathbf{Z}

$$\text{logit}\{\text{Pr}(Y = 1|X, \mathbf{Z})\} = \beta_0 + \beta_1 X + \beta_2 \mathbf{Z} + \beta_3 X\mathbf{Z}, \tag{12}$$

then $OR(\mathbf{Z}) = e^{\beta_1 + \beta_3 \mathbf{Z}}$. Under this model, \widehat{AF} does not simplify as in Eq. (11).

The estimation procedure outlined above applies to matched case–control studies as well, where the conditional logistic regression is commonly used instead of the ordinary logistic regression. The standard error for the resulting estimate can be obtained by combining the sandwich formula with the delta method [8,9].

3.2 Applied example

In a study on causes of oesophageal cancer, cases and controls were collected from hospitals in Singapore during 1970–1972 [18]. Each case was individually matched to 4 controls on sex and age within 5 years intervals. In this article we re-analyse a publicly available subset of these data, consisting of 80 male cases and their 320 matched male controls. De Jong et al.[18] considered various potential risk factors for oesophageal cancer, such as intake of bread, potato, bananas and beverages at burning hot temperatures, smoking, and alcohol intake. In this article we focus on intake of beverages at burning hot temperatures, which was observed to be highly associated with oesophageal cancer by De Jong et al. [18]. The available variables are the patient’s age (Age), dialect group (Dial: 1 for Hokhien/Teochew and 0 for Cantonese/Other), a binary indicator of whether the patient consumes Samsu wine on a daily basis (Samsu), number of cigarettes smoked per day (Cigs), a binary indicator of whether the patient drinks beverages at ‘burning hot’ temperatures on a daily basis (Everhotbev), a unique identification number for each matched set (Set), and a binary case–control indicator (Oesophagealcancer). These variables are stored in the data frame `singapore`

which is included in the AF package. We will adjust for age, dialect, if the patient drinks Samsu wine and number of smoked cigarettes per day since these variables are potential confounders. We assume the following conditional logistic regression model:

$$\begin{aligned} \text{logit}\{\text{Pr}(\text{Oesophagealcancer} \\ = 1 | \text{Everhotbev}, \text{Age}, \text{Dial}, \text{Samsu}, \text{Cigs}, \text{Set})\} \\ = \beta_{\text{Set}} + \beta_1 \text{Everhotbev} + \beta_2 \text{Age} \\ + \beta_3 \text{Dial} + \beta_4 \text{Samsu} + \beta_5 \text{Cigs}, \end{aligned} \tag{13}$$

where β_{Set} is a set-specific intercept. The model in Eq. (13) can be fitted in R with the `clogit` function from the `Survival` package, as follows:

```
> fit <- clogit(Oesophagealcancer ~ Everhotbev + Age + Dial + Samsu +
              Cigs + strata(Set), data = singapore)
```

Summarizing the output gives:

```
> fit
Call:
clogit(Oesophagealcancer ~ Everhotbev + Age + Dial + Samsu +
      Cigs + strata(Set), data = singapore)
```

	coef	exp(coef)	se(coef)	z	p
Everhotbev	1.16912	3.21915	0.29239	4.00	6.4e-05
Age	0.00786	1.00790	0.05500	0.14	0.88629
Dial	1.22857	3.41635	0.32017	3.84	0.00012
Samsu	0.48253	1.62017	0.28595	1.69	0.09152
Cigs	0.01114	1.01121	0.00962	1.16	0.24649

Likelihood ratio test=51.7 on 5 df, p=6.16e-10
n= 400, number of events= 80

The output indicates that the odds of getting oesophageal cancer is about $e^{1.16912} \approx 3.22$ times higher among those who drink at least one beverage at burning hot temperature every day compared to those who do not. The effect is highly significant, with a p-value of 0.000064. To estimate the proportion of cases of oesophageal cancer that would have been prevented if no patient had consumed beverage at burning hot temperatures we use the `AF.cc` function:

```
> AFest <- AF.cc(formula = Oesophagealcancer ~ Everhotbev + Age + Dial + Samsu + Cigs,
                data = singapore, exposure = "Everhotbev", clusterid = "Set", matched = TRUE)
```

By setting the argument `matched=TRUE`, conditional logistic regression is used instead of ordinary logistic regression (the default). Summarizing the output gives:

```
> summary(AFest)
Call:
AF.cc(formula = Oesophagealcancer ~ Everhotbev + Age + Dial +
      Samsu + Cigs, data = singapore, exposure = "Everhotbev",
      clusterid = "Set", matched = TRUE)
Estimated attributable fraction (AF) and untransformed 95% Wald CI:
      AF Robust SE z value Pr(>|z|) Lower limit Upper limit
0.3360627 0.05542807 6.063041 1.335717e-09 0.2274256 0.4446997
Exposure : Everhotbev
```

```
Outcome : Oesophagealcancer
Observations Cases Clusters
          400    80    80
Method for confounder adjustment: Conditional logistic regression
Formula: Oesophagealcancer ~ Everhotbev + Age + Dial + Samsu + Cigs
```

The output indicates that approximately 34 % all cases of oesophageal cancer would have been prevented if no patient had consumed beverages at burning hot temperature. The AF is highly significant, with a p-value close to zero, and a 95 % CI ranging from 23 to 44 %.

4 Cohort study with time-to-event outcome

4.1 Theory

In cohort studies with time-to-event outcomes, the AF function is defined as in Eq. (2) [3]. Equivalently, the AF function can be expressed as

$$\text{AF} = 1 - \frac{\{1 - S_0(t)\}}{\{1 - S(t)\}}, \quad (14)$$

where $S(t) = 1 - \Pr(T \leq t)$ is the factual survival function, and $S_0(t) = 1 - \Pr(T_0 \leq t)$ is the counterfactual survival function had the exposure been eliminated for everyone at baseline.

As before, we assume that a set of observed covariates, \mathbf{Z} , is sufficient for confounding control. Under this assumption, $S_0(t)$ can be obtained by averaging the survival function among the unexposed at a given value of \mathbf{Z} over the population distribution of \mathbf{Z} :

$$S_0(t) = E\{S(t) \mid X = 0, \mathbf{Z}\}.$$

In practice, $S(t \mid X, \mathbf{Z})$ is usually estimated with a Cox proportional hazards model

$$\lambda(t \mid X, \mathbf{Z}) = h(t)e^{g(X, \mathbf{Z}; \boldsymbol{\beta})}, \quad (15)$$

where $\lambda(t \mid X, \mathbf{Z})$ is the conditional hazard at time t , given X and \mathbf{Z} , $h(t)$ is the unspecified baseline hazard, and $g(X, \mathbf{Z}; \boldsymbol{\beta})$ is a specified function of the exposure X and confounders \mathbf{Z} indexed by the parameter vector $\boldsymbol{\beta}$. For example, $g(X, \mathbf{Z}; \boldsymbol{\beta})$ can be specified as $g(X, \mathbf{Z}; \boldsymbol{\beta}) = \beta_1 X + \beta_2 \mathbf{Z}$. However, $g()$ could also involve interactions and higher-order terms.

The model in Eq. (15) is fitted to obtain the partial likelihood estimate of $\boldsymbol{\beta}$ [19] and the Breslow estimate of the cumulative baseline hazard function $\Lambda(t) = \int_{u=0}^t h(u) du$ [20]. Then, for each fixed value of t we proceed as follows. For each subject i with exposure level X_i and covariate vector \mathbf{Z}_i , we use $e^{-e^{g(X_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})} \hat{\Lambda}(t)}$ as a prediction of $S(t \mid X_i, \mathbf{Z}_i)$.

These predictions are averaged to obtain an estimate of $S(t)$:

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^n e^{-e^{g(X_i, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})} \hat{\Lambda}(t)}. \quad (16)$$

Similarly, for each subject i with covariate vector \mathbf{Z}_i , we use $e^{-e^{g(X=0, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})} \hat{\Lambda}(t)}$ as a prediction of $S(t \mid X = 0, \mathbf{Z}_i)$. These predictions are averaged to obtain an estimate of $S_0(t)$:

$$\hat{S}_0(t) = \frac{1}{n} \sum_{i=1}^n e^{-e^{g(X=0, \mathbf{Z}_i; \hat{\boldsymbol{\beta}})} \hat{\Lambda}(t)}. \quad (17)$$

The estimates of $S(t)$ and $S_0(t)$ are plugged in to Eq. (14) to produce an estimate of the AF function. The standard error of the resulting estimate can be obtained by combining the sandwich formula with the delta method [5].

4.2 Applied example

To illustrate the theory we use records from 2982 women with primary breast cancer from the Rotterdam tumor bank in the Netherlands. The Rotterdam breast cancer dataset is thoroughly described in Sauerbrei et al. [21] and Royston and Lambert [22]. The follow-up time ranges from 1 to 231 months. The outcome variables are the time, measured in months, that the patient is under study (`rf`), and an indicator of whether the patient experienced death or relapse before censoring (`rfi`). Seven prognostic variables are recorded: age at surgery (`age`), menopausal status (`meno`: 0=pre and 1=post), tumor size in three classes (`size`: '<=20 mm', '>20–50 mm' and '>50 mm'), tumor grade (`grade`: 2 or 3), progesterone receptors, (`pr`: fmol/l), oestrogen receptors, (`er`: fmol/l) and the number of positive lymph nodes (`nodes`: ranging between 0 and 34). In our example, we consider absence of chemotherapy as the exposure, i.e. we wish to estimate the proportion of deaths that would have been prevented before a given time, if all patients had been given chemotherapy at baseline. Absence of chemotherapy is measured by the variable `no.chemo`, with levels 0 for 'yes' and 1 for 'no'. These variables are stored in the data frame `rott2`, which is included in the AF package.

In the analysis we will adjust for the seven prognostic factors as well as age since these variables are potential confounders. We assume the following Cox proportional hazards model:

$$\begin{aligned} \lambda(t \mid \text{no.chemo}, \text{age}, \text{meno}, \text{size}, \text{grade}, \text{nodes}, \text{pr}, \text{er}) \\ = h(t) e^{\beta_1 \text{no.chemo} + \beta_2 \text{age} + \beta_3 \text{meno} + \beta_4 \text{size} + \beta_5 \text{grade} + \beta_6 e^{-0.12 \text{nodes}} + \beta_7 \text{pr} + \beta_8 \text{er}}. \end{aligned} \quad (18)$$

In this model we used the transformation $e^{-0.12 \text{nodes}}$, since Sauerbrei et al. [21] noted that this transformation gave a bet-

ter model fit. The model in Eq. (18) can be fitted in R with the `coxph` function from the `survival` package, as follows:

```
> fit <- coxph(Surv(rf,rfi) ~ no.chemo + age + meno + size + grade+ as.factor(nodes) + pr + er,
  data = rott2)
```

Summarizing the output gives:

```
> fit
Call:
coxph(formula = Surv(rf, rfi) ~ no.chemo + age + meno + size +
  grade + I(exp(-0.12 * nodes)) + pr + er, data = rott2)

              coef exp(coef) se(coef)      z      p
no.chemo      2.82e-01  1.33e+00  7.23e-02  3.90 9.8e-05
age           -1.60e-02  9.84e-01  3.52e-03 -4.54 5.7e-06
menopre      -1.32e-01  8.76e-01  8.95e-02 -1.48  0.14
size>20-50mm  2.89e-01  1.33e+00  5.85e-02  4.93 8.2e-07
size>50mm     4.81e-01  1.62e+00  8.91e-02  5.40 6.8e-08
grade         3.47e-01  1.42e+00  6.50e-02  5.34 9.1e-08
I(exp(-0.12 * nodes)) -1.85e+00  1.57e-01  9.79e-02 -18.92 < 2e-16
pr            -9.55e-05  1.00e+00  1.06e-04 -0.90  0.37
er            -5.09e-05  1.00e+00  1.04e-04 -0.49  0.63
```

```
Likelihood ratio test=574 on 9 df, p=0
n= 2982, number of events= 1518
```

The output indicates that the hazard of death or relapse is about 1.33 times higher among those who did not receive chemotherapy compared to those who did.

The AF function can be estimated by the function `AF.ch` as follows:

```
> AFest <- AF.ch(Surv(rf,rfi) ~ no.chemo + age + meno + size + grade + I(exp(-0.12 * nodes)) + pr + er,
  data = rott2, exposure = "no.chemo",
  times = c(12, 24, 36, 48, 60))
```

The formula argument is on the same format as the formula argument in the `coxph` function. The `times` argument specifies the time points at which the AF is supposed to be estimated. If not specified, `times` defaults to all observed event times. In the call to `AF.ch` above we have asked for the AF at years 1 to 5 (12 to 60 months). Summarizing the output gives:

```
> summary(AFest)
Call:
AF.ch(formula = Surv(rf, rfi) ~ no.chemo + age + meno + size +
  grade + I(exp(-0.12 * nodes)) + pr + er, data = rott2,
  exposure = "no.chemo", times = c(12, 24, 36, 48, 60))

Estimated attributable fraction (AF) and untransformed 95% Wald CI:

Time   AF   Std.Error   z value   Pr(>|z|)   Lower limit   Upper limit
12    0.1818045  0.04398121  4.133686  3.569912e-05  0.09560292  0.2680061
24    0.1668268  0.04080641  4.088250  4.346397e-05  0.08684770  0.2468059
36    0.1557651  0.03845976  4.050080  5.120099e-05  0.08038538  0.2311449
48    0.1479756  0.03679384  4.021749  5.776757e-05  0.07586099  0.2200902
60    0.1416357  0.03544215  3.996251  6.435344e-05  0.07217040  0.2111011

Exposure : no.chemo
Event      : rfi

Observations Events
          2982    1518

Method for confounder adjustment: Cox Proportional Hazards model
Formula: Surv(rf, rfi) ~ no.chemo + age + meno + size + grade + I(exp(-0.12 * nodes)) + pr + er
```

The output indicates that the AF function decreases over the 5 year period, from 18% at 1 year (12 months) after baseline to 14% at 5 years (60 months) after baseline. The AF is statistically significant at all time points.

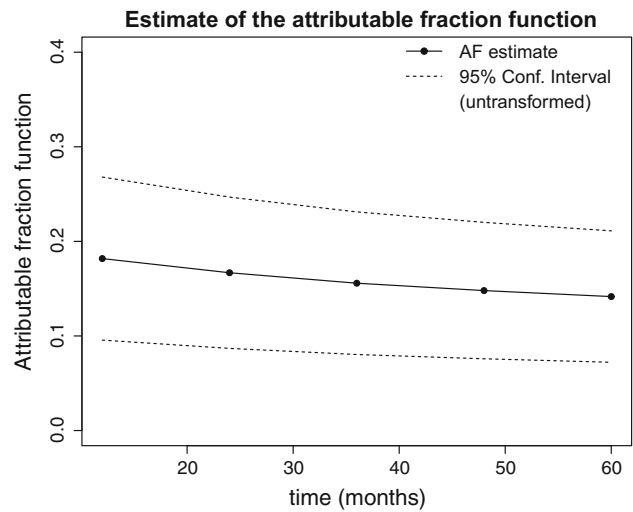


Fig. 1 Estimated AF function over 5 years for relapse-free survival time in the Rotterdam dataset (solid line) together with point-wise 95% confidence intervals (dashed lines)

A convenient way to visualize the AF function is to use the `plot` function in the `AF` package:

```
> plot(AFest, CI = TRUE, ylim = c(0, 0.4), xlab = "time (months)")
```

where `AFest` is an object of class `AF`, estimated by the `AF.ch` function. This function call produces the plot in Fig. 1.

5 Discussion

In this article we have presented the new R package `AF`, developed for epidemiologists and biostatisticians. The package `AF` estimates the confounder-adjusted AF for cross-sectional studies, case-control studies (matched and unmatched) and cohort studies with time-to-event outcomes, using the functions `AF.cs`, `AF.cc` and `AF.ch`, respectively. We have used three datasets (`clslowbwt`, `singapore` and `rott2`) for illustration. These datasets are all publicly available and included in the `AF` package. Thus, readers of the paper can easily replicate all analyses that we have presented.

In order for the estimated AF to have a causal interpretation, it is necessary that those covariates that are adjusted for are sufficient for confounding control. In practice, important confounders might be unknown and/or unmeasured, which implies that the estimated AF should always be interpreted cautiously. For instance, in our applied example in Section 4.2 comorbidity is a potential confounder for the association between chemotherapy and relapse-free survival, since those patients who have severe comorbidities are less likely to be prescribed chemotherapy due to the health risks of this invasive treatment, and also more likely to die early during follow-up. Thus, if comorbidities are not adjusted for,

then the protective effect of chemotherapy may be overestimated. As a consequence, the AF may be overestimated as well.

The main advantage of the AF package, as compared to other R packages for AF estimation (`epi2by2`, `attribrisk` and `paf`), is that it offers a uniform way to estimate the confounder-adjusted AF for all three major study designs. The package has a standard input/output interface, which makes it easy to use for practitioners who has some familiarity with R. Another important advantage is that the AF package provides analytic standard errors, based on the delta method and the sandwich formula, which alleviates the need for time-consuming bootstrap or jackknife methods. Finally, the AF package produces correct standard errors when data are clustered, e.g. when there are repeated measures on each individual or when the dataset contains related individuals.

The AF package covers the most fundamental estimation strategies developed for the most common study designs. Possible extensions include so-called ‘partial attributable fractions’ [11,23,24], which allow for multiple exposures, and so-called ‘generalized impact fractions’ [25–27], which allow for continuous exposures. Another possible extension is to allow for more advanced models for time-to-event outcomes, such as flexible parametric models [22]. We plan to make these extensions in the future.

References

- Poole C. A history of the population attributable fraction and related measures. *Ann Epidemiol.* 2015;25(3):147–54.
- Levin ML. The occurrence of lung cancer in man. *Acta Union Int Contr.* 1953;9(3):531–41.
- Chen YQ, Hu C, Wang Y. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics.* 2006;7(4):515–29.
- Chen L, Lin DY, Zeng D. Attributable fraction functions for censored event times. *Biometrika.* 2010;97(3):713–26.
- Sjölander A, Vansteelandt S. Doubly robust estimation of attributable fractions in survival analysis. *Stat Methods Med Res.* 2014. doi:10.1177/0962280214564003.
- Sturmans F, Mulder PG, Valkenburg HA. Estimation of the possible effect of interventive measures in the area of ischemic heart diseases by the attributable risk percentage. *Am J Epidemiol.* 1977;105(3):281–9.
- Deubner DC, Wilkinson WE, Helms MJ, Tyroler HA, Hames CG. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County. *Ga Am J Epidemiol.* 1980;112(1):135–43.
- Greenland S, Drescher K. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics.* 1993;49(3):865–72.
- Sjölander A, Vansteelandt S. Doubly robust estimation of attributable fractions. *Biostatistics.* 2011;12(1):112–21.
- Miettinen OS. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am J Epidemiol.* 1974;99(5):325–32.
- Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol.* 1985;122(5):904–14.
- Core Team R. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. <https://www.r-project.org/>. Accessed 25 Nov 2015.
- Nunes T, Heuer C, Marshall J, Sanchez J, Thornton R, Reiczigel J, Robison-Cox J, Sebastiani P, Solyms P, Yoshida K, Jones G, Firestone SPaS. `epiR`: Tools for the analysis of epidemiological data. CRAN. 2015. <https://cran.r-project.org/web/packages/epiR/index.html>. Accessed 25 Nov 2015.
- Schenck L, Atkinson E, Crowson C, Therneau T. `attribrisk`: population attributable risk. CRAN. 2014. <https://cran.r-project.org/web/packages/attribrisk/index.html>. Accessed 25 Nov 2015.
- Chen L. `paf`: attributable fraction function for censored survival data. CRAN. 2014. <https://cran.r-project.org/web/packages/paf/index.html>. Accessed 25 Nov 2015.
- Juul S, Frydenberg M. An introduction to Stata for health researchers. 3rd ed. College Station, Texas: Stata Press; 2010.
- Lehnert-Batar A, Pfahlberg A, Gefeller O. Comparison of confidence intervals for adjusted attributable risk estimates under multinomial sampling. *Biom J.* 2006;48(5):805–19.
- De Jong UW, Breslow N, Hong JG, Sridharan M, Shanmugaratnam K. Aetiological factors in oesophageal cancer in Singapore Chinese. *Int J Cancer.* 1974;13(3):291–303.
- Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Stat Methodol.* 1972;34(2):187–220.
- Breslow N. Discussion of the paper by D. R. Cox. *J R Stat Soc B.* 1972;34:216–7.
- Sauerbrei W, Royston P, Look M. A new proposal for multivariable modelling of time-varying effects in survival data based on fractional polynomial time-transformation. *Biom J.* 2007;49(3):453–73.
- Royston P, Lambert PC. Flexible Parametric survival analysis using Stata: beyond the Cox model. 1st ed. College Station, Texas: Stata Press; 2011.
- Eide GE, Gefeller O. Sequential and average attributable fractions as aids in the selection of preventive strategies. *J Clin Epidemiol.* 1995;48(5):645–55.
- Rämsch C, Pfahlberg AB, Gefeller O. Point and interval estimation of partial attributable risks from case-control data using the R-package ‘`pARccs`’. *Comput Meth Prog Biol.* 2009;94(1):88–95.
- Morgenstern H, Bursic ES. A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population. *J Commun Health.* 1982;7(4):292–309.
- Drescher K, Becher H. Estimating the generalized impact fraction from case-control data. *Biometrics.* 1997;53(3):1170–6.
- Taguri M, Matsuyama Y, Ohashi Y, Harada A, Ueshima H. Doubly robust estimation of the generalized impact fraction. *Biostatistics.* 2012;13(3):455–67.