Attributable Risk Estimation in Cohort Studies

Cynthia S. Crowson
Terry M. Therneau
W. Michael O'Fallon

Technical Report #82
October 8, 2009

Department of Health Sciences Research
Mayo Clinic
Rochester, Minnesota

Contents

Abstract

Population attributable risk (PAR) or etiologic fraction is the proportion of a disease that could be prevented by elimination of a causal risk factor from the population. PAR is likely to be a function of time because both the prevalence of a risk factor and its effect on exposed individuals may change over time, as may the underlying risk of disease. In cohort studies with a wide range of follow-up, it may be important to account for this time dependency. Estimating the full PAR curve on some time scale of interest (e.g. age or time on study) can be more informative than considering a single value. Time-specific PAR can be estimated based on cumulative incidence adjusted for the competing risk of death. Cox models with time-dependent covariates can be used to obtain PAR estimates adjusted for confounders. The unique value of this approach is illustrated with examples that arose from our studies of heart failure in rheumatoid arthritis.

1 Introduction

Population attributable risk (PAR), a public health measure, was first described by Levin [Levin, 1953]. While many different names and formulae for PAR have been used, PAR is generally defined as the proportion of disease in a population that could be prevented by elimination of a causal exposure or risk factor from the population and is generally expressed as a percentage. A root definition for PAR is given by

$$PAR = \frac{P(D) - P(D|\overline{F})}{P(D)} \tag{1}$$

where P(D) is the probability of disease in the population as a whole and $P(D|\overline{F})$ is the conditional probability of disease among those without the risk factor. If the factor is associated with a higher risk of disease, i.e. $P(D|F) > P(D|\overline{F})$, then $0 < PAR < 1$.

PAR is most often estimated from case-control or cross-sectional studies using formulae that combine the prevalence of exposure to a risk factor and a measure of the impact of that risk factor on the development of disease. Using Bayes' rule and the assumption of a fixed risk factor prevalence and a fixed P(D) (and hence a fixed time t), the PAR formula is often re-written in terms of the relative risk [Levin, 1953] (RR) as

$$PAR = \frac{P(F)(RR-1)}{1 + P(F)(RR-1)}. \tag{2}$$

This is the standard formula used in case-control studies, with RR replaced by the estimated odds ratio. This formula has also been used in cohort studies, with RR replaced by the hazard ratio from a Cox model [Benichou, 2001]. While this may be adequate for short-term cohort studies, it is unclear how accurate these estimates are for long-term cohort studies. The formula ignores the complexity of PAR estimation in the longitudinal setting and oversimplifies the problem by failing to account for the possibility that PAR may be a function of when risk is assessed. When subjects are followed long enough longitudinally, some events (such as death) are not preventable, but may be delayed. In addition, the prevalence of a risk factor in a given population may be changing over time due to new exposures and to deaths. Therefore, PAR is a function of observation time, since it depends on the prevalence of the risk factor, the effect of the risk factor on exposed individuals and on the underlying probability of disease, all of which may change over time.

Motivation for this work comes from studies of rheumatoid arthritis (RA) [Crowson *et al.* , 2005, Nicola *et al.* , 2005]. A longitudinal cohort of 603 population-based RA subjects with a mean follow-up of 15 years (range: 16 days to 46.7 years) and a comparison cohort of non-RA subjects with similar age, sex, and follow-up time were assembled. RA subjects have a nearly two-fold increased risk of heart failure (HF) compared to non-RA subjects. Cardiovascular (CV) risk factors and ischemic heart disease (IHD) are known to contribute to the development of HF in the general population. The excess risk of HF in RA subjects could be due to either an increased prevalence of CV risk factors or an increased effect of these factors on the development of HF in RA subjects. If neither of these explanations is

plausible, it is possible that subjects with RA could be developing HF through a different mechanism. Attributable risk estimates can give insight into which of these possibilities lead to the excess HF in RA by quantifying the amount of the risk of HF in RA which is attributable to CV risk factors and IHD.

2 Methods

2.1 Estimating the probability of disease

The first component needed to estimate PAR as a function of time, PAR(t), is the probability of developing the disease of interest at or before time t, $P_D(t)$. If follow-up is complete through time t, then an empirical estimate of $P_D(t)$ is simply the proportion of subjects that have entered the disease state by that time. Typically, however, some subjects will be censored at times $c_i < t$, which occurs when a subject's status beyond a particular time point c_i is unknown. This can be due to subjects who are still alive and under observation at the time of analysis (administrative censoring), or loss of contact with a subject. Censoring needs to be distinguished from competing risks, such as death from another cause, after which it is known that a subject will *not* get the disease. Any number of competing events can be important in a particular study, but for the purposes of this paper, one competing event is assumed. This could be just one actual competing event (such as death) or multiple competing events (such as death, cancer, stroke, etc.) lumped together into a single outcome.

When censoring is present, but there are no competing risks, the Kaplan-Meier (KM) estimator can be used to estimate $1 - P_D(t)$ [Kaplan & Meier, 1958]. The KM estimator assumes that censored subjects' further experience, if it were available, would be unbiasedly represented by those who do have longer follow-up available. This assumption of non-informative censoring is usually feasible when the majority of censoring is administrative. When competing risks are present, the KM estimate is often modified by treating occurrence of the competing event as a censoring time c_i. Under an assumption (often highly questionable) of independence of the event types, the resulting KM curve is an estimate of what $P_D(t)$ *would be* in the population, if the competing events were to be eliminated. For estimation of PAR(t) the modified KM is not appropriate, as it does not address the right question. Since PAR is a public health measure, we want to use the prevalence of disease as experienced in the population, not what it *would be* under artificial conditions (i.e. the absence of competing risks).

The appropriate estimator for our case is the cumulative incidence (CI), which accounts for competing risks [Kalbfleisch & Prentice, 2002]. In this approach, $P_D(t)$ is estimated using $CI_1(t) = \int_0^t \lambda_1(u)S(u^-)du$, where $\lambda_1(t)$ is the cause-specific risk for the event of interest at time t, and $S(t^-)$ is the probability of being at risk for the event of interest just before time t. $\hat{S}(t)$ can be estimated using the KM for "any event", restricting censoring to observations actually censored. Provided that there are no tied event times (actually, no ties that involve more than a single event type), $\hat{S}(t) = \hat{S}_1(t) * \hat{S}_2(t)$, where $S_1(t)$ is the modified KM for the event of interest (treating all other events as censored) and $S_2(t)$ is the modified KM for the competing events. The product form is particularly useful when adjusting for covariates. Note that the CI estimate depends on the overall probability, $\hat{S}(t)$, of still being at risk. The factorization, $S(t) = S_1(t) * S_2(t)$, provides a convenient way to compute $\hat{S}(t)$, particularly in the presence of covariates. Although independence is absolutely critical for $S_1(t)$ and $S_2(t)$ to estimate the cause-specific survival curves that "would occur" if the other cause were removed, independence is not necessary for computation of the overall survival, S(t). Using a Nelson estimate for the cause-specific hazard, λ, the CI can then easily be calculated. Hence, the CI estimator can be used to provide an estimate of the probability of disease for use in PAR(t) estimation.

2.2 Estimating the conditional probability of disease without the risk factor

The second component needed to estimate PAR(t) is the probability of disease at time t if the risk factor of interest were eliminated from the population, $P_D(t|\overline{F})$. This quantity could be estimated using a standard CI estimator, restricted to the subjects without the risk factor. This can be problematic because the subset of such patients may be small or, even more importantly, may be biased. For

instance, those without hypertension may be younger than those with hypertension. Time-dependent risk factors, such as are found in the RA study, are a particularly thorny problem. Consider F= hypertension as a particular case. A computation that includes a patient in the $\overline{F}$ estimate, censored at the time of development of hypertension, is making the very strong assumption of non-informative censoring. That is, that the subject who developed hypertension was essentially a "random selection" from among all those at risk, with respect to their likelihood of future events. This assumption is likely to be invalid because hypertension is not a random occurrence. The same genetic and environmental factors which led to hypertension may likely play a role in the occurrence of future events such as heart disease.

A better approach involves explicitly incorporating covariates into the CI estimator. This will allow estimation of $P_D(t|\overline{F})$ and will also accommodate adjustment for potential confounders. In this approach, the probability of disease at time t for patient i, $P_{Di}(t)$, is estimated using $CI_{i1}(t|x_i) = \int_0^t \lambda_{i1}(u|x_i)S(u^-|x_i)du$ where x_i represents the vector of risk factors of interest for patient i. Various modeling methods can be used to estimate the two necessary components $\lambda_{i1}(t|x_i)$ and $S(t^-|x_i)$. A common choice is the Cox proportional hazards model, which assumes that $\lambda_i(t|x_i) = \lambda_0(t)e^{x_i(t)\beta}$ for each individual cause of failure. Hence, the components can be estimated using two Cox models, one for the disease of interest and one for competing events. The overall survival, $S(t|x)$, can be estimated as $exp[-\Lambda_1(t|x) + \Lambda_2(t|x)] = S_1(t) * S_2(t)$, which combines the cumulative hazards from the two Cox models.

Since the Cox models are hazard models, patients are censored at the time when they experience the event of interest. If the model includes all the relevant covariates for the endpoint of interest, then informative censoring is not an issue as it was with the simple population CI. Instead of assuming that censored subjects are similar to *all* subjects still under observation, the model assumes that censored subjects are similar to those *with the same covariates* who are still under observation.

In cases such as the RA study, where risk factors develop over time, a Cox model with time-dependent covariates can be used to capture this dynamic information (for both the risk factor of interest and the other determinants of risk). The data setup for such models has become fairly standard; one way is to divide each subject into disjoint intervals each of which becomes a separate observation in the data set [Therneau & Grambsch, 2000]. The variable values for each observation indicate the person's covariate level during that interval.

The fitted models can be used to obtain estimates of the hazard function, $\lambda_i(t|x_i)$, for any pre-specified set of covariates or for a specified covariate path over time in the case of time-dependent covariates. Using these hazard function estimates, $CI(t|x_i)$ can be estimated as described above. Then estimating the probability of disease in the absence of the risk factor is simply a matter of specifying the appropriate set of covariates. Therefore, estimates of $P_D(t|\overline{F})$ can be obtained using the CI estimator.

2.3 Obtaining estimates of $P_D(t)$ that represent the cohort

Since attributable risk is a public health measure, estimates that represent $P_D(t)$ and $P_D(t|\overline{F})$ in a population are needed in order to use the corresponding estimates of PAR to make inferences about the population. Using the models described above, estimates of $P_D(t)$ and $P_D(t|\overline{F})$ can be obtained for any specific risk profile. However, no single risk profile will adequately represent the cohort, especially in the presence of time-dependent covariates. Almost every multivariate covariate path $x_i(t)$ in the study will be unique. The obvious estimate of a population risk, $\frac{1}{n}\sum_{i=1}^{n} CI_i(t|x_i(t))$, would require that we extrapolate all of the observed covariate paths out to the maximum observation time. To avoid extrapolation predicted $CI(t|x_i)$ estimates can be obtained for each subject's covariate path and averaged at each time point. In the presence of censoring, the average CI(t) estimates will be discontinuous because averages at different time points include different subsets of subjects. This problem can be avoided by transforming the $CI(t|x_i)$ estimates to hazards and averaging the underlying hazard estimates instead of averaging the $CI(t|x_i)$ estimates directly. This method corresponds to the conditional method for expected survival estimation [Therneau & Grambsch, 2000, Ederer & Heise, 1977]. One advantage of the conditional method is that it remains consistent when differential censoring is present. Hence the

conditional method is used to obtain $\overline{CI}(t|x_i)$, an estimate of $P_D(t)$ which represents the cohort for use in PAR estimation.

2.4 Obtaining estimates of $P_D(t|\overline{F})$ that represent the cohort

The conditional method can be used to obtain an estimate of $P_D(t|\overline{F})$ which represents the cohort using the concept of a target value. An indicator variable for the risk factor of interest, x*, can be used to obtain a predicted CI for a cohort of patients identical to the observed cohort but without a particular risk factor by defining x* = 0 (or some desired target value) for all subjects in the cohort. Creative target values (e.g. 10% decrease in systolic blood pressure) can be used to mimic the effects of specific interventions. Using the conditional method described above, an estimate of $P_D(t|\overline{F})$ can be obtained at each time point by estimating $\overline{CI}(t|x^*)$. PAR(t) can then be estimated using these estimates. Plots of PAR(t) vs. time are useful for visualizing the pattern of PAR over time.

3 Confidence intervals

A closed form expression for the variance of PAR(t) is difficult to obtain because PAR(t) involves a ratio of two complex quantities. Therefore, we recommend the use of bootstrap sampling to generate pointwise confidence intervals for PAR(t). Each replicate sample should be constrained to have the same number of endpoints of each type (censored, event of interest, competing event) as the original sample, which essentially conditions on the total amount of information as an ancillary statistic [Efron & Tibshirani, 1993]. Several methods can be used to obtain confidence intervals based on the bootstrap samples [DiCiccio & Efron, 1996]. The bias-corrected (BC) estimates of confidence intervals with 1000 replicates were used in the RA example which appears later in the manuscript.

4 Comparison of PAR(t) to simple PAR estimates

Simple PAR estimates are often obtained in cohort studies by using a hazard ratio from a Cox model in place of the relative risk in standard PAR formulae as recommended by Benichou [Benichou, 2001]. For studies with a wide range of follow-up times, ignoring the complexity of PAR estimation is not justified. However, Appendix A shows that the simple PAR estimate approximates PAR(t) when certain assumptions are met and the mean of the risk factor prevalence at the event times is used.

In order to better understand the factors that influence PAR in a longitudinal study, consider the following example. The probability of disease for a subject without a risk factor is shown in curve B in Figure 1, assuming the hazard for the outcome of interest follows a Weibull distribution with risk starting at age 55 and parameters of $\lambda = 0.017$ and p = 1.5. Curve A in Figure 1, displays the probability of disease for a subject with a risk factor that increases the early risk of disease (obtained using $\lambda = 0.034$). Assuming a prevalence of the risk factor of 50%, the probability of disease in the population is the average of these 2 curves, which yields the dashed line in the figure. Of note, this event curve is similar in shape and size to the observed cumulative incidence of HF in the RA example, which follows this section. The dashed line in Figure 2 shows PAR(t) as a function of age for the population in Figure 1. In this particular case, PAR can be computed in closed form and the resulting PAR(t) curve shows for each age what PAR would be with complete follow-up to that age. Early in the follow-up, when the event rate is low, PAR(t) agrees with the approximation obtained using the standard PAR formula. This is consistent with the rare disease assumption commonly used to argue that an odds ratio, risk ratio and hazard ratio are all equivalent when a disease is rare. When the incidence rate of the event is much lower, or equivalently when the competing risk makes the observed incidence of the event small, the PAR(t) curve is much flatter. This is shown in the solid line on Figure 2, where $\lambda = 0.017/5$ vs. $0.34/5$ yielding a $P_D(t)$ which is 1/5 of the prior values. This observation is consistent with the proof in Appendix A, which showed PAR(t) and the standard PAR agree assuming the disease is rare and in the absence of competing risks.

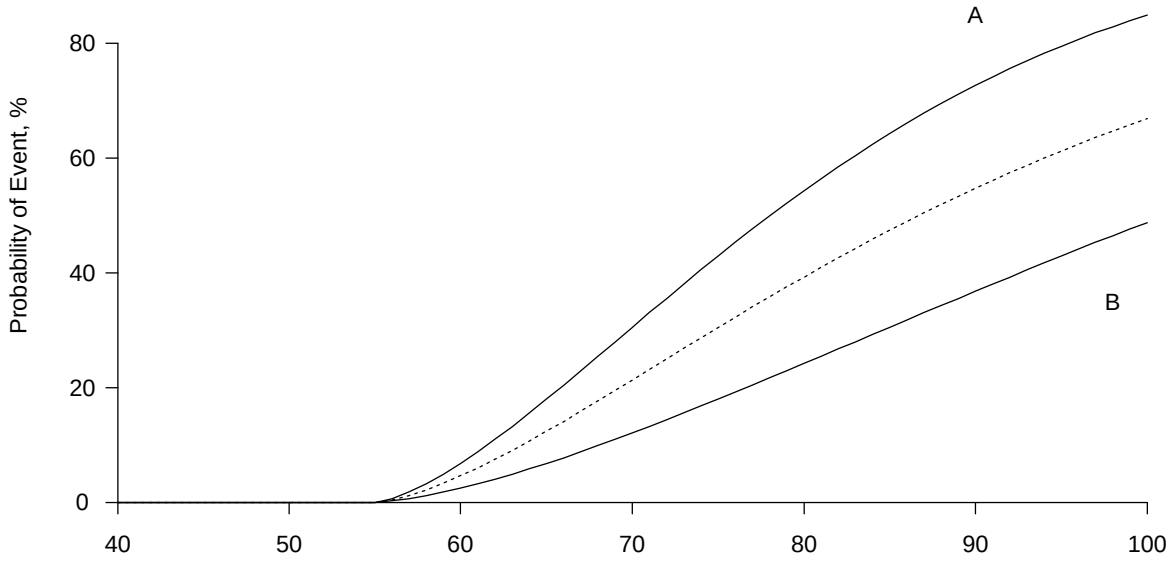

Figure 1: *Probability of disease vs. age for a hypothetical example. Curve A displays the probability of disease for a subject with a risk factor that increases the early risk of disease. Curve B displays the probability of disease for a subject without the risk factor. The dashed line displays probability of disease for a population with a risk factor prevalence of 50%.*

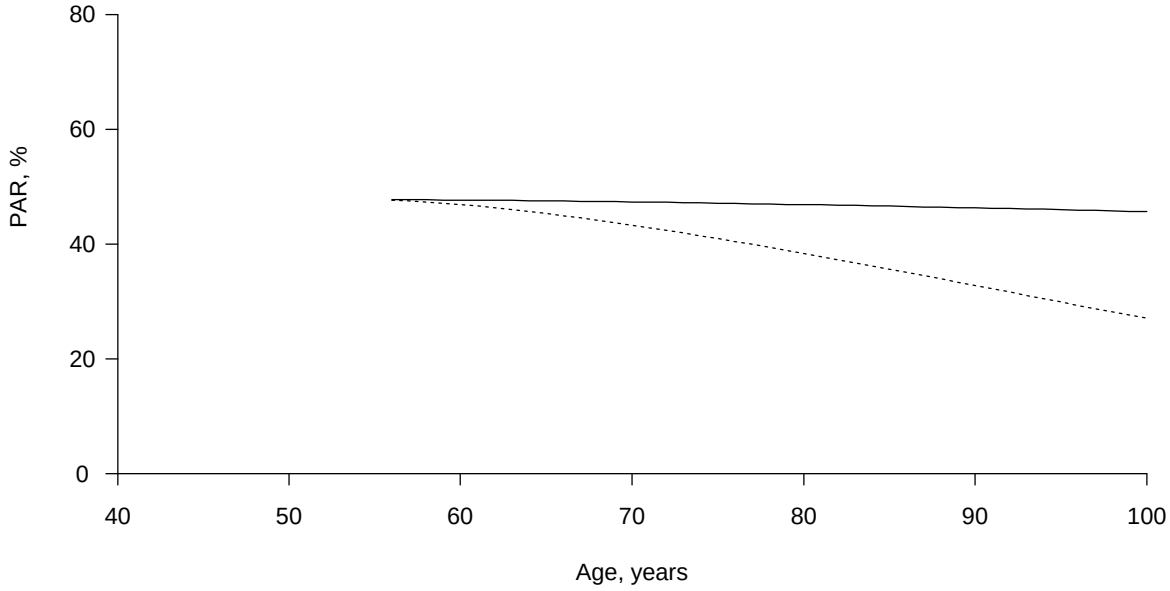

Figure 2: *Population attributable risk (PAR) vs. age. The dashed line displays PAR for the example in figure 1. The solid line displays PAR for a population with a lower probability of disease, possibly due to a competing event.*

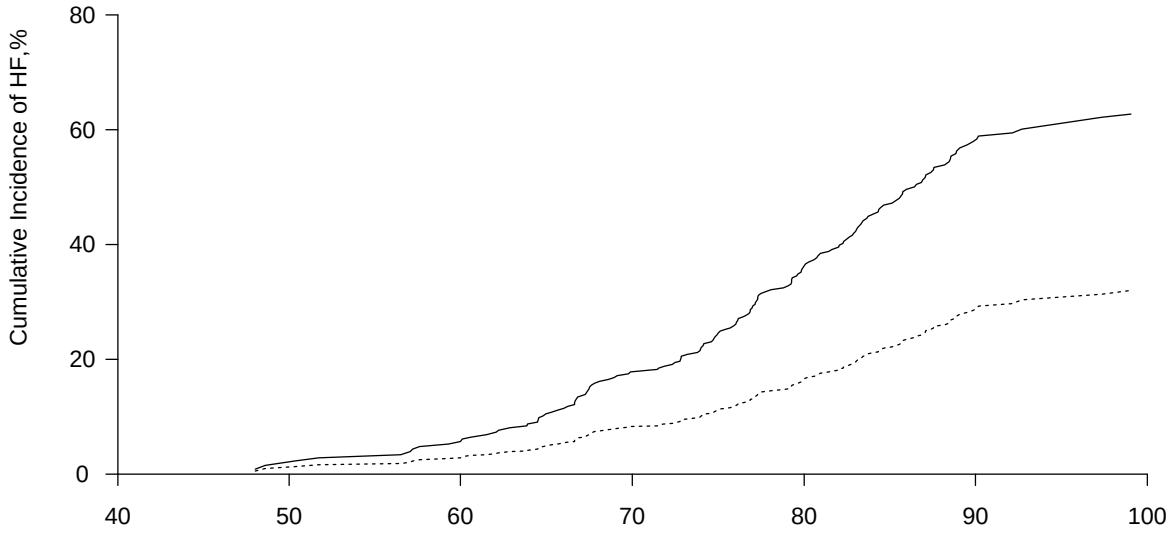

Figure 3: *Cumulative Incidence of heart failure (HF) vs. age among 575 rheumatoid arthritis (RA) subjects (solid line) along with estimated cumulative incidence of HF in the absence of all cardiovascular risk factors and ischemic heart disease (dashed line).*

The shape of PAR(t) is affected by several factors. First, the distribution of event times will affect the pattern of decline in PAR(t). If the event times are not uniformly distributed, steeper declines will appear at times where more events occur. Second, competing risks affect the shape of PAR(t). The presence of a competing risk will lower the prevalence of the event of interest, which will flatten the PAR(t) curve. Finally, the prevalence of the risk factor over time will change the shape of PAR(t). A risk factor with low prevalence at baseline, but increasing prevalence over time, will cause a rise in PAR(t). Time-dependent covariates can easily be used to represent risk factors which develop over time in the PAR(t) estimation. Given all these factors, predicting the shape of the PAR(t) curves is non-trivial. When the prevalence of the risk factor and the outcome are both reasonably large, no single estimate of PAR is adequate. It is quite possible that a risk factor with increasing prevalence will increase PAR(t) while an important competing risk will attenuate this increase. In fact, this phenomenon will be demonstrated in the following example.

5 Example

The rheumatoid arthritis data serve as an example to illustrate the PAR estimation methods described above. Excluding those with HF prior to their RA incidence date (or index date for those without RA), 575 RA subjects with a mean age of 57.1 years (range: 18 to 92 years) and 583 non-RA subjects with a mean age of 57.5 years (range: 18 to 94 years) were examined. Of these, 165 RA and 115 non-RA subjects developed HF during follow-up. Study details, descriptive statistics, and multivariable Cox model results have been previously published [Crowson *et al.* , 2005, Nicola *et al.* , 2005]. Age was used as the time scale for these analyses because CV events like HF are more highly related to age than to RA disease duration.

Figure 3 displays the observed cumulative incidence of HF as a function of age among the RA subjects along with the predicted cumulative incidence of HF in the absence of CV risk factors (in-

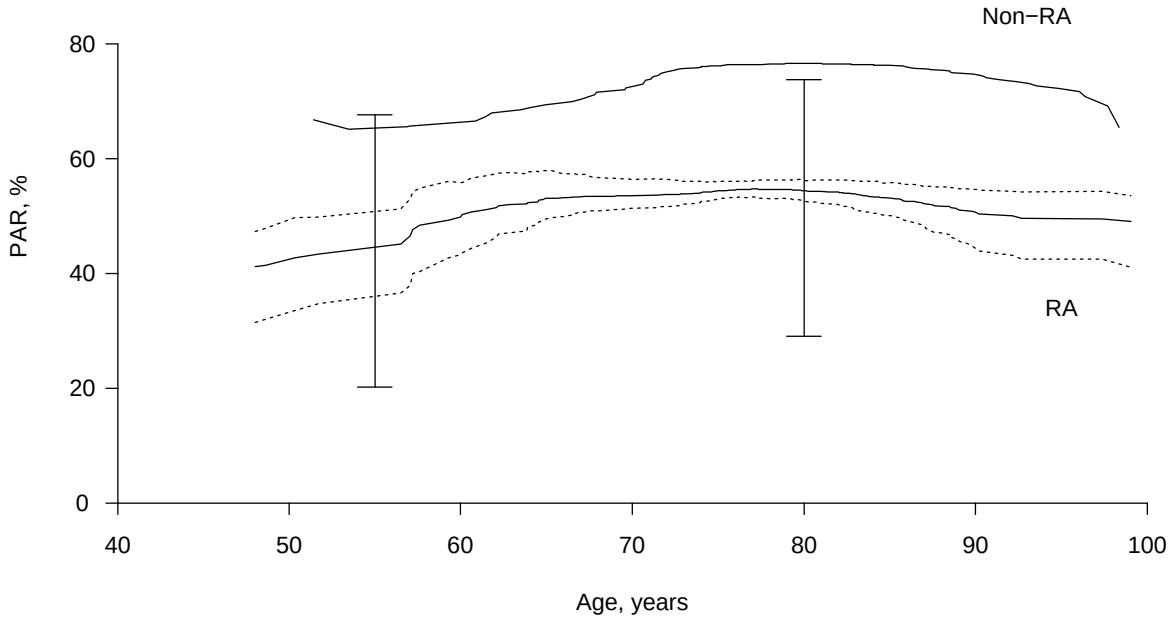

Figure 4: *Population attributable risk (PAR) of HF among RA subjects with 95% pointwise confidence intervals at ages 55 and 80 and the PAR of HF among non-RA subjects, for comparison. Centered 95% confidence intervals for assessment of shape are shown with dashed lines.*

cluding family history of IHD, cigarette smoking, hypertension, dyslipidemia, abnormal body mass index, diabetes mellitus, and alcohol abuse) and IHD. The cumulative incidence of HF rose steadily with age (Figure 3). The observed cumulative incidence of HF at age 80 was 36% for RA subjects and the predicted cumulative incidence of HF in the absence the risk factors was 17% at age 80 for RA subjects. The difference between the observed cumulative incidence of HF and the predicted cumulative incidence of HF in the absence of the risk factors (i.e. the risk difference) was 19% at age 80; this is a measure of the absolute impact of these factors on the risk of HF in RA subjects.

The PAR rose from 45% at age 55 to 54% at age 80 (Figure 4). The BC 95% confidence interval at age 80 ranged from 29% to 74%. The pointwise confidence intervals for PAR(t) were surprisingly large. For the younger ages, this was largely a consequence of taking the ratio of two small numbers; both $\overline{CI}(t|x)$ and $\overline{CI}(t|x^*)$ were near zero. At the later ages, it was due to the uncertainty in the Cox model estimates, $\hat{\beta}$, as $\overline{CI}(t|x^*)$ is highly dependent on $\hat{\beta}$. The same issue of large confidence limits resulting from x* far from $\overline{x}$ exists when computing the usual survival curves from a Cox model [Tsiatis, 1981, Andersen *et al.* , 1993]. The wide confidence intervals also seem to indicate that the apparent trend in PAR across time is not statistically significant, but although the pointwise intervals are the correct measure for predicting absolute PAR at a particular x, they are less satisfactory for judging shape. Examination of the bootstrap samples revealed the shape of the PAR curve to be fairly consistent across samples; most of the variability in the pointwise estimates is in the intercept of the PAR estimates (i.e. the curve shifts up or down as a whole while maintaining its shape). The same issue arises in assessing non-linearity in generalized additive models (GAM). Chambers and Hastie [Chambers & Hastie, 1992] show that centered intervals are more appropriate for assessment of shape; these are implemented in the standard GAM plots from Splus and R. Figure 4 also shows the confidence band for the RA curve based on centering. The centered BC 95% confidence intervals show an apparent difference between the PAR at ages 55 and age 80; however, this increase in PAR was not thought to be clinically meaningful.

Due to the almost two-fold increased risk of HF in RA subjects, the goal of this analysis was to

compare the PAR of HF among RA subjects to the PAR in a non-RA comparison cohort of similar age and sex. Among non-RA subjects, the risk of HF attributable to the risk factors rose from 66% at age 55 to 77% at age 80 (Figure 4). In other words, about one-fourth (23%) of HF at age 80 was not attributable to the risk factors among the non-RA, while nearly one-half (46%) of HF at age 80 was not attributable to these factors in the RA cohort. This observed difference of 23% in the PAR of HF at age 80 in the RA cohort compared to the non-RA cohort is statistically significant (BC 95% confidence interval: 7% to 36%). This difference in PAR means the excess risk of HF in RA was not entirely explained by these risk factors. In addition, the risk difference was similar in both cohorts (19% in RA and 16% in non-RA at age 80), implying these factors had similar absolute impact in both cohorts. This suggests that the increased risk of HF in RA was not completely explained by an increase in the prevalence or effect of these risk factors and is due at least partially to other causes.

6 Discussion

Estimation of PAR in longitudinal studies is complex, but quite doable, as all components are available in standard Cox model software. When the prevalence of the risk factor and the outcome are both reasonably large, a single estimate of PAR is inadequate. As shown in the closed form example, when the prevalence of the risk factor is moderately large and the outcome is not rare, PAR can change substantially over time. In the RA example, the PAR curves were reasonably flat, but this does not obviate the need for methods which account for the issues relevant to PAR estimation in the longitudinal setting. The prevalence of a risk factor in a given study population may be increasing over time due to new exposures, but decreasing over time due to deaths. This necessitates the use of time-dependent covariates and adjustment for the competing risk of death. When subjects are followed long enough longitudinally, some events (such as death) are not preventable and the entire population would reach the endpoint regardless of risk factor exposure, resulting in PAR = 0. Adjustment for competing risks yields more realistic estimates. Examining PAR as a function of time can be more informative than considering a single value. In some instances, the elimination of a risk factor may significantly delay the onset of the disease, but may not change the lifetime risk or ultimately prevent the disease. This phenomenon would cause the time-specific PAR to first rise and then fall. We have shown that PAR as a function of time can be estimated using cumulative incidence adjusted for the competing risk of death. In addition, Cox models can be used to obtain estimates which incorporate time-dependent covariates for the risk factor of interest and adjust for potential confounders.

It is also important to remember that PAR is a public health measure and as such it needs to represent a population of interest. In certain studies, it may be necessary to weight the sampled subjects to obtain estimates which represent the population of interest. This could be done by incorporating case weights into the Cox models used to estimate CI(t) and also using weighted averages to obtain the estimates of $P_D(t)$ and $P_D(t|\overline{F})$ needed to estimate PAR(t). In the RA study, all incidence cases of RA in Rochester, MN residents, over a 40-year time span were assembled, so estimates based on this cohort are logically representative of the cohort of RA patients in this community.

The use of competing risks in the estimation of PAR in longitudinal studies was first discussed by Robins and Greenland [Robins & Greenland, 1989]. They focused on estimating a lower bound for PAR, rather than examining PAR as a function of time. However, their methodology is rarely used in PAR estimation, likely due to its complexity and the lack of software. Chen et al [Chen *et al.* , 2006] utilize proportional hazards models to estimate PAR over time and they utilize a normal approximation to obtain confidence intervals. Samuelsen and Eide [Samuelsen & Eide, 2008] focus on a measure they call the attributable hazard and they recommend bootstrap confidence intervals similar to ours.Cox et al [Cox *et al.* , 2009] introduce attributable survival as an alternative to attributable risk, which may be more meaningful for some studies. None of these articles use methodology that adjusts for competing risks. While this adjustment was important in our example, there are many studies that are not impacted by competing risks, such as those with death as an outcome.

There are several benefits to estimating PAR as a function of time using the methodology we have described. First, these methods allow for realistic estimation of PAR in longitudinal studies, which may make this analysis more common. Second, these methods capitalize on the inherent flexibility of Cox

models to allow for time-dependent risk factors, adjusting for potential confounders, incorporating case weights, and realistically representing the changes in risk factors associated with specific interventions (e.g. eliminating preventing new teenage smokers vs. smoking cessation among senior citizens). Finally, the trends in PAR(t) can be used to target interventions to the age groups where the greatest impact can be achieved. For instance, if the PAR declines substantially by age 80, an intervention designed to eliminate a risk factor may be futile for people in this age group.

There are also some limitations to estimating PAR as a function of time. An important limitation is the wide confidence intervals associated with small samples sizes. Larger numbers of events than those found in the RA study would ease comparisons. Other factors which could reduce the variability include choosing less extreme target values (e.g., elimination of hypertension in 10% of patients, instead of all patients) or modifying the Cox models to minimize the residual variability. In addition, the complexity of the dataset manipulation required to use time-dependent covariates in Cox models and to account for competing risks in the estimation of cumulative incidence can be non-trivial for those who are less familiar with these methods. Finally, estimating confidence intervals for PAR(t) using bootstrap sampling is computationally intensive. Generalized software could be developed incorporating techniques to optimize performance.

Better estimates of the effect of the risk factor on disease development may also improve accuracy of the PAR estimates. Further research to examine other failure time models which allow for changing the relative risk associated with a risk factor over time is needed. Although no evidence of non-proportional hazards was found in the RA study, other models may be more appropriate for chronic diseases, such as RA. Alternative models could easily be used in place of Cox models within the framework described here for estimating PAR(t).

In conclusion, realistic PAR estimation is possible in longitudinal studies. This capability may help make PAR estimation more prevalent in epidemiologic studies. PAR estimates can be used to identify risk factors with the greatest potential for reducing disease burden, which can help target preventive interventions. The additional information obtained by estimating PAR across time can further inform preventive efforts.

7 Funding

This work was funded by a grant from the National Institutes of Health, NIAMS (R01 AR46849) and made possible by a grant from the National Institutes of Health, NIAMS (AR-30582).

8 Acknowledgements

The authors wish to thank Dr. Sherine E. Gabriel for encouragement and for providing the data used for the example, Dr. Paulo J. Nicola for posing the question and Dr. Kenneth J. Koehler and Mr. Eric J. Bergstralh for editorial suggestions.

9 Appendix

9.1 Theorem

Levin's PAR formula, namely

$$PAR = \frac{P(F)(RR-1)}{1+P(F)(RR-1)},$$

approximates an average over time of the PAR formula presented in this paper, namely

$$PAR(t) = \frac{\overline{CI}(t|x) - \overline{CI}(t|x^*)}{\overline{CI}(t|x)},$$

which was derived from the root PAR formula of

$$PAR = \frac{P(D) - P(D|\overline{F})}{P(D)},$$

when competing risks are ignored, an "average" P(F) is used, and the disease of interest is "rare".

9.2 Proof

9.2.1 Simplifying the first piece, $\overline{CI}(t|x)$

Ignoring competing risks and using the conditional method,

$$\overline{CI}(t|x) = 1 - exp[-\int_0^t \frac{\sum_{i=1}^n Y_i(u)\lambda_i(u|x_i)}{\sum_{i=1}^n Y_i(u)}du].$$

Since $\lambda_i(t|x_i) = \lambda_0(t)e^{x_i(t)\beta}$ in the Cox model,

$$\begin{aligned}\overline{CI}(t|x) &= 1 - exp[-\int_0^t \frac{\sum_{i=1}^n Y_i(u)\lambda_0(u)e^{x_i(u)\beta}}{\sum_{i=1}^n Y_i(u)}du] \\ &= 1 - exp[-\int_0^t \frac{\sum_{i=1}^{p(u)} Y_i(u)\lambda_0(u)e^{1\beta} + \sum_{i=p(u)+1}^n Y_i(u)\lambda_0(u)e^{0\beta}}{\sum_{i=1}^n Y_i(u)}du]\end{aligned}$$

where $x(t) = 1$, if the factor is present, and otherwise, $x(t) = 0$, and where p(t) is the number of subjects with the factor at time t. By pulling constants out of the sums,

$$\overline{CI}(t|x) = 1 - exp[-\int_0^t \frac{\lambda_0(u)e^{\beta}\sum_{i=1}^{p(u)} Y_i(u)}{\sum_{i=1}^n Y_i(u)} + \frac{\lambda_0(u)\sum_{i=p(u)+1}^n Y_i(u)}{\sum_{i=1}^n Y_i(u)}du].$$

Define $P_F(u) = \frac{\sum_{i=1}^{p(u)} Y_i(u)}{\sum_{i=1}^n Y_i(u)}$. Then

$$\begin{aligned}\overline{CI}(t|x) &= 1 - exp[-\int_0^t \lambda_0(u)e^{\beta}P_F(u) + \lambda_0(u)[1-P_F(u)]du] \\ &= 1 - exp[-\int_0^t \lambda_0(u)[1 + P_F(u)(e^{\beta}-1)]du].\end{aligned}$$

9.2.2 Simplifying the second piece, $\overline{CI}(t|x^*)$

Ignoring competing risks and using the conditional method and since $\lambda_i(t|x_i^*) = \lambda_0(t)e^{x_i^*(t)\beta}$ in the Cox model and $x^*(t) = 0$ for all x and t,

$$\begin{aligned}\overline{CI}(t|x^*) &= 1 - exp[-\int_0^t \frac{\sum_{i=1}^n Y_i(u)\lambda_0(u)e^{x_i^*(t)\beta}}{\sum_{i=1}^n Y_i(u)}du] \\ &= 1 - exp[-\int_0^t \lambda_0(u)du].\end{aligned}$$

9.2.3 Put the pieces in the PAR(t) formula

$$\begin{aligned} PAR(t) &= \frac{\overline{CI}(t|x) - \overline{CI}(t|x^*)}{\overline{CI}(t|x)} = 1 - \frac{\overline{CI}(t|x^*)}{\overline{CI}(t|x)} \\ &= 1 - \frac{1 - exp - \int_0^t \lambda_0(u)du}{1 - exp[-\int_0^t \lambda_0(u)1 + P_F(u)(e^\beta - 1)du]}. \end{aligned}$$

Considering the integral in the denominator, if we make a change of variable, multiply and divide by $\Lambda_0(t)$, and think of $\frac{\lambda_0(u)}{\Lambda_0(u)}$ as a PDF,

$$\begin{aligned} \int_0^t \lambda_0(u)[1 + P_F(u)(e^\beta - 1)]du &= \int_0^t [1 + P_F(u)(e^\beta - 1)]d\lambda_0(u) \\ &= \Lambda_0(t) \int_0^t [1 + P_F(u)(e^\beta - 1)] \frac{d\lambda_0(u)}{\Lambda_0(u)} \\ &= \Lambda_0(t)[1 + \overline{P}_F(e^\beta - 1)], \end{aligned}$$

where $\overline{P}_F$ is the mean prevalence with respect to this distribution.

Substituting into the PAR formula,

$$PAR(t) = 1 - \frac{1 - exp[-\Lambda_0(t)]}{1 - exp[-\Lambda_0(t)[1 + \overline{P}_F(e^\beta - 1)]]}.$$

Taking the first two terms of the Taylor expansion gives

$$\begin{aligned} PAR(t) &\approx 1 - \frac{1}{1 + \overline{P}_F(e^\beta - 1)} - \Lambda_0(t) \frac{\overline{P}_F(e^\beta - 1)}{2[1 + \overline{P}_F(e^\beta - 1)]} \\ &\approx \frac{\overline{P}_F(e^\beta - 1)}{1 + \overline{P}_F(e^\beta - 1)} (1 - \frac{\Lambda_0(t)}{2}). \end{aligned}$$

9.3 Conclusion

Levin's formula corresponds to the 0^{th} order Taylor expansion,

$$\begin{aligned} PAR(t) &\approx \frac{\overline{P}_F(e^\beta - 1)}{1 + \overline{P}_F(e^\beta - 1)} \\ &\approx \frac{P(F)(RR - 1)}{1 + P(F)(RR - 1)} \end{aligned}$$

using $P(F) = \overline{P}_F$ and $RR \approx e^\beta$. The 0^{th} order Taylor expansion is reasonable when $\Lambda_0(t)$ is small. Note that small $\Lambda_0(t)$ corresponds to the usual rare disease assumption commonly used to argue that $OR \approx RR$. Also $\overline{P}_F$ is the mean of the risk factor prevalence at the event times.

References

[Andersen *et al.* , 1993] Andersen, P.K., Borgan, O., Gill, R.D., & Keiding, N. 1993. *Statistical Models Based on Counting Processes.* New York: Springer-Verlag.

[Benichou, 2001] Benichou, J. 2001. A review of adjusted estimators of attributable risk. *Statistical Methods in Medical Research*, **10**, 195–216.

[Chambers & Hastie, 1992] Chambers, J.M., & Hastie, T.J. 1992. *Statistical Models in S.* Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books and Software.

[Chen *et al.* , 2006] Chen, Y. Q., Hu, C., & Wang, Y. 2006. Attributable risk function in the proportional hazards model for censored time-to-event. *Biostatistics*, **7(4)**, 515–529.

[Cox *et al.* , 2009] Cox, C., Chu, H., & Munoz, A. 2009. Survival attributable to an exposure. *Statistics in Medicine*, **online**.

[Crowson *et al.* , 2005] Crowson, C.S., Nicola, P.J., Maradit-Kremers, H., O'Fallon, W.M., Therneau, T.M., Jacobsen, S. J., Roger, V.L., Ballman, K.V., & Gabriel, S.E. 2005. How much of the increased incidence of heart failure in rheumatoid arthritis is attributable to traditional cardiovascular risk factors and ischemic heart disease? *Arthritis and Rheumatism*, **52(10)**, 3039–3044.

[DiCiccio & Efron, 1996] DiCiccio, T.J., & Efron, B. 1996. Bootstrap Confidence Intervals. *Statistical Science*, **11(3)**, 189–228.

[Ederer & Heise, 1977] Ederer, F., & Heise, H. 1977. *Instructions to IBM 650 programmers in processing survival computations.* Methodological Note No. 10. End Results Evaluation Section, National Cancer Institute.

[Efron & Tibshirani, 1993] Efron, B., & Tibshirani, R.J. 1993. *An introduction to the bootstrap.* New York: Chapman and Hall.

[Kalbfleisch & Prentice, 2002] Kalbfleisch, J.D., & Prentice, R.L. 2002. *The Statistical Analysis of Failure Time Data: Second Edition.* New York: Wiley.

[Kaplan & Meier, 1958] Kaplan, E.L., & Meier, P. 1958. Non-parametric estimation from incomplete observations. *Journal of the America Statistical Association*, **53**, 457–481.

[Levin, 1953] Levin, M.L. 1953. The occurrence of lung cancer in man. *Acta Un Intern cancer*, **19**, 531–541.

[Nicola *et al.* , 2005] Nicola, P.J., Maradit-Kremers, H., Roger, V.L., Jacobsen, S. J., Crowson, C.S., Ballman, K.V., & Gabriel, S.E. 2005. the risk of congestive heart failure in rheumatoid arthritis: a population-based study over 46 years. *Arthritis and Rheumatism*, **52(2)**, 412–420.

[Robins & Greenland, 1989] Robins, J.M., & Greenland, S. 1989. Estimability and estimation of excess and etiologic fractions. *Statistics in Medicine*, **8(7)**, 845–859.

[Samuelsen & Eide, 2008] Samuelsen, S. O., & Eide, G. E. 2008. Attributable fractions with survival data. *Statistics in Medicine*, **27**, 1147–1467.

[Therneau & Grambsch, 2000] Therneau, T.M., & Grambsch, P.M. 2000. *Modeling Survival Data: Extending the Cox Model.* New York: Springer-Verlag.

[Tsiatis, 1981] Tsiatis, A.A. 1981. A large sample study of Cox's regression model. *Annals of Statistics*, **9**, 93–108.