Published in final edited form as:

Lifetime Data Anal. 2014 April ; 20(2): 252–275. doi:10.1007/s10985-013-9258-4.

Population-based absolute risk estimation with survey data

Stephanie A. Kovalchik and

Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd.,
EPS 8047, Rockville, MD 20852, Tele: 626-319-9890, Fax: 301-402-0081

Ruth M. Pfeiffer

Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd.,
EPS 8030, Rockville, MD 20852, Tel: 301 5947832, Fax: 301-402-0081

Stephanie A. Kovalchik: kovalchiksa@mail.nih.gov; Ruth M. Pfeiffer: pfeiffer@mail.nih.gov

Abstract

Absolute risk is the probability that a cause-specific event occurs in a given time interval in the presence of competing events. We present methods to estimate population-based absolute risk from a complex survey cohort that can accommodate multiple exposure-specific competing risks. The hazard function for each event type consists of an individualized relative risk multiplied by a baseline hazard function, which is modeled nonparametrically or parametrically with a piecewise exponential model. An influence method is used to derive a Taylor-linearized variance estimate for the absolute risk estimates. We introduce novel measures of the cause-specific influences that can guide modeling choices for the competing event components of the model. To illustrate our methodology, we build and validate cause-specific absolute risk models for cardiovascular and cancer deaths using data from the National Health and Nutrition Examination Survey. Our applications demonstrate the usefulness of survey-based risk prediction models for predicting health outcomes and quantifying the potential impact of disease prevention programs at the population level.

Keywords

Absolute risk; Censored data; Crude risk; Cumulative incidence; NHANES; Survey cohort

1 Introduction

Individual risk assessment is a fundamental concern of health research. By summarizing the combined impact of risk factors and competing risks on the development of disease over time, absolute risk modeling provides a comprehensive framework for individual risk prediction, and its clinical and public health applications are numerous (Benichou and Gail, 1990; Fine and Gray, 1999). Despite the wealth of health outcome data generated by surveys, methods for absolute risk prediction from survey data are currently lacking.

Mortality-linked data from national health surveys allow researchers to assess how demographic characteristics and health behaviors impact the risk of death due to various causes. Methods have been developed to estimate hazard functions from survey data (Binder, 1992; Lin, 2000; Gray, 2009) and to estimate the average marginal predicted survival for the Cox proportional hazards model (Korn and Graubard, 1999, Chapter 8.2). However, the estimation of absolute risk from survey data has not been discussed in the literature. Building on previous work in cause-specific hazard modeling, we extend the results of Benichou and Gail (1990) for simple cohorts to estimate absolute risk for a finite

population using survey data with complex sampling designs. We use a weighted approach to reduce dependence on model assumptions and improve robustness (Patterson et al, 2002).

In Section 2.1 we define absolute risk for mutually exclusive failure types that are each described by a cause specific hazard model. In Section 2.2, we describe the structure of the survey data that is used for model development. Estimation with survey data can be conducted with a finite-population (design-based) or superpopulation (model-based) perspective. We present estimation methods in Section 3, using expansion estimators (Kish and Frankel, 1974) that yield absolute risk estimators that are consistent for finite- and superpopulation inferences. In Section 3.2, influence methods are used to derive Taylor-linearized variance estimates for the absolute risk estimates that are valid for finite-population inference. Section 3.3 introduces measures to aid the specification of the cause-specific hazard models for competing events, which are based on the absolute risk model's influence components. We illustrate our methods in Section 4 by building separate population absolute risk models of cardiovascular and cancer causes of death using data from the First and Second National Health and Nutrition Examination Surveys (NHANES I and NHANES II). We use these models to project the number of deaths by cause in the US population from an independent survey cohort (NHANES III), before concluding the paper in Section 5 with a discussion.

2 Model, Survey Setting and Data

2.1 Model

The absolute risk of experiencing the m th type of event ($m = 1, \dots, M$) within the time interval $[t_0, t_1]$ in the presence of $M - 1$ competing events, given that one is alive and failure free at t_0 , is

$$\pi^{(m)}(t_0, t_1; \mathbf{x}) = \int_{t_0}^{t_1} \lambda^{(m)}(u; \mathbf{x}^{(m)}) \exp\left\{-\int_{t_0}^u \sum_{i=1}^M \lambda^{(i)}(s; \mathbf{x}^{(i)}) ds\right\} du \quad (1)$$

where $\mathbf{x} = (\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)})$ is a set of cause-specific covariate vectors and $\lambda^{(m)}(u; \mathbf{x}^{(m)})$ is the cause-specific hazard function given covariates $\mathbf{x}^{(m)}$ (Benichou and Gail, 1990, 1995). In what follows, for ease of exposition, the superscript in $\pi^{(m)}$ will be omitted and we will regard the primary event as $m = 1$. We assume that covariates in (1) remain fixed at their values at the beginning of the projection interval, t_0 . The risk time scale could be age or time since diagnosis of a particular disease. However, as our application focuses on modeling cause-specific absolute risk of mortality, and age is a strong risk factor, we use age as the time scale.

In this work, we let each cause-specific hazard function follow a proportional hazards model

$$\lambda^{(m)}(t; \mathbf{x}^{(m)}) = \lambda_0^{(m)}(t) \exp(\boldsymbol{\beta}^{(m)'} \mathbf{x}^{(m)}) \quad (2)$$

where $\lambda_0^{(m)}(t)$ denotes the baseline hazard function at time t . In Section 3.4, we describe how our methodology can be extended to other hazard models. We present estimates when the baseline hazard function is modeled nonparametrically or parametrically with a piecewise exponential model.

2.2 Survey Setting

We suppose that the data for the development of the absolute risk models come from a complex multistage survey. We illustrate the setup with the NHANES Epidemiologic Followup studies that we used in our application presented in detail in Section 4.

2.2.1 NHANES Epidemiologic Followup—We used the NHANES I, 1971–1975, and NHANES II, 1976–1980, Epidemiologic Followup studies with mortality follow-up through 1993. Briefly, the NHANES samples are designed to be nationally representative of the civilian, non-institutionalized U.S. population. Participants for the surveys are selected using multistage probability sampling designs. In the first stage primary sampling units (PSUs) are selected by probability proportional to size (PPS) sampling from strata defined by geography and proportions of minority populations. For all NHANES surveys the PSUs are mostly single counties or, in a few cases, groups of contiguous counties. In the second stage the PSUs are divided up into segments (generally city blocks or their equivalent) and sample segments are selected with PPS sampling. In stage 3, a random sample of households within each segment is drawn and in stage 4 individuals from a list of all persons residing in selected households are chosen at random within designated age-sex-race/ethnicity screening subdomains. On average, 1.6 persons are selected per household.

The NHANES I sample design is approximated by 35 strata with 3 PSUs sampled from each stratum, and NHANES II has 32 strata with two PSUs sampled from each stratum. For both NHANES surveys sample weights are provided on the public use data files appropriate for obtaining national estimates.

We combined the two NHANES surveys by treating the design as coming from 67 consecutive strata, and revised the weights by dividing them by two, following Korn and Graubard (1999, Chapter 8). The combined survey is approximately representative of the US population from 1971 to 1980. Mortality follow-up data for NHANES I are based upon deaths identified during the NHANES I Epidemiologic Follow-up Study (NHEFS) follow-up periods and on the results from a probabilistic match with death certificate records from the National Death Index (NDI). Mortality follow-up was solely based on a probabilistic match with death certificate records from the NDI for NHANES II participants who were 30–75 years of age and completed a medical examination. Detailed information on the NHANES I survey, the NHANES II survey, and their follow-up studies can be found in Engel et al (1978), Cox et al (1992) and McDowell et al (1981), respectively.

Thus the ages and covariates of individuals in the sample at baseline are known, and ages at death and cause of death of individuals who died in or before 1993 are known.

2.2.2 Data—In what follows a superscript refers to cause-specific quantities and three subindices accommodate the sampling scheme. To summarize, the sampling frame of the survey is divided into I non-overlapping strata and J_i PSUs are selected from the i th stratum with inclusion probabilities $p_{i1}, \dots, p_{iJ_i}$. There can be several additional stages of complex sampling after the first stage that we will not describe further. At the end of the sampling, $k = 1, \dots, n_{ij}$ individuals are sampled from the j th sampled PSU of the i th stratum, and the final inclusion probabilities p_{ijk} of the ijk th individual are the products of the PSU level inclusion probabilities from the first stage times the conditional inclusion probabilities from the later stages. The sampling weights are $w_{ijk} = 1/p_{ijk}$, $k = 1, \dots, n_{ij}$, with possible adjustment for nonresponse and non-coverage bias.

At baseline, the age of individual ijk in the sample, denoted by T_{0ijk} , and a vector of covariates $\mathbf{X}_{ijk}$ are observed. Over the course of the study we also observe the minimum of the age at death or age at censoring, T_{ijk} , with censoring assumed to be non-informative. Using $I(\cdot)$ to denote the indicator function, the at-risk indicator at time t is $Y_{ijk}(t) = I(T_{ijk} > t)$, and the vector of cause-specific event indicators is $\delta_{ijk}(t) = (\delta_{ijk}^{(1)}(t), \dots, \delta_{ijk}^{(M)}(t))$, where $\delta_{ijk}^{(m)}(t) = 1$ if an event of type m occurs at time t or zero otherwise. Thus, the data for the ijk th individual at time t are $(T_{0ijk}, T_{ijk}, \mathbf{X}_{ijk}, w_{ijk}, \delta_{ijk}(t), Y_{ijk}(t))$.

3 Inference

3.1 Absolute Risk Estimation

Recall that the absolute risk is a function of cause-specific hazards (2). We present methods where the baseline hazard function $\lambda_0^{(m)}$ is estimated from a piecewise exponential parametric model or is estimated nonparametrically.

3.1.1 Hazard ratio parameters—Using a pseudo-likelihood formulation as suggested by Binder (1992), the estimating equations for $\beta^{(m)}(1, \dots, M)$ are

$$U(\beta^{(m)}) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} \delta_{ijk}^{(m)}(t_{ijk}) \{ \mathbf{x}_{ijk}^{(m)} - \bar{\mathbf{H}}(\beta^{(m)}, t_{ijk}) \}$$

where $\bar{\mathbf{H}}(\beta^{(m)}, t)$ is a weighted-average of the risk profiles $\mathbf{x}^{(m)}$ among the individuals still at-risk at time t ,

$$\bar{\mathbf{H}}(\beta^{(m)}, t) = \frac{\sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}^{(m)}(t) \exp(\beta^{(m)'} \mathbf{x}_{ijk}^{(m)})}{\sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}^{(m)}(t) \exp(\beta^{(m)'} \mathbf{x}_{ijk}^{(m)})}. \quad (3)$$

Standard optimization algorithms can be used to obtain the solution $\hat{\beta}_m$ such that $U(\beta^{(m)}) = 0$.

3.1.2 Piecewise exponential baseline hazard model—We define a set of ordered time intervals $\tau_1, \tau_2, \dots$, with $\tau_q = [\tau_{0q}, \tau_{1q})$ and $\tau_{1q} = \tau_{0(q+1)}$. Denote the constant cause-specific hazard during the q th time interval by $\lambda_0^{(m)}(\tau_q)$. The baseline hazard function under the piecewise exponential model is

$$\lambda_0^{(m)}(t) = \sum_q \lambda_0^{(m)}(\tau_q) I(t \in \tau_q).$$

Following the formulation proposed by Benichou and Gail (1990), the estimate for the cause-specific probability of the primary event type $m = 1$ occurring between $[\tau_{0l}, \tau_{1n})$ is

$$\hat{\pi}(\tau_{0l}, \tau_{1n}; \mathbf{x}) = \sum_{q=l}^n \hat{S}(\tau_q) \hat{P}^{(1)}(\tau_q) \left[1 - \exp\left\{ - \sum_i \hat{\lambda}_0^{(i)}(\tau_q) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)}) (\tau_{1q} - \tau_{0q}) \right\} \right], \quad (4)$$

where $\hat{P}^{(1)}(\tau_q) = \hat{\lambda}_0^{(1)}(\tau_q) \exp(\hat{\beta}^{(1)'} \mathbf{x}^{(1)}) / \sum_i \hat{\lambda}_0^{(i)}(\tau_q) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)})$ and

$$\hat{S}(\tau_q) = \begin{cases} 1 & q=l \\ \prod_{j=l}^{q-1} \exp\left\{ - \sum_i \hat{\lambda}_0^{(i)}(\tau_j) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)}) (\tau_{1j} - \tau_{0j}) \right\} & q>l \end{cases}. \quad (5)$$

Let t_{0ijk} be the time that the k th individual from the j th PSU and i th stratum is first at-risk. Define the at-risk time during the q th interval as

$$\mathcal{A}_{ijk}(\tau_q) = I(t_{0ijk} < \tau_{1q} \cap t_{ijk} > \tau_{0q}) (\min\{t_{ijk}, \tau_{1q}\} - \max\{t_{0ijk}, \tau_{0q}\}). \quad (6)$$

Setting $D^{(m)}(\tau_q)$ equal to the total weighted person-time during the q th interval

$$D^{(m)}(\tau_q) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} \mathcal{A}_{ijk}(\tau_q) w_{ijk} \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}),$$

and given $d^{(m)}(\tau_q) = \sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} \delta_{ijk}^{(m)}(t_{ijk}) I(\tau_{0q} \leq t_{ijk} < \tau_{1q})$ sample-weighted observed events of the m th type in $[\tau_{0q}, \tau_{1q})$, the estimate for $\lambda_0^{(m)}(\tau_q)$ is

$$\hat{\lambda}_0^{(m)}(\tau_q) = d^{(m)}(\tau_q) / D^{(m)}(\tau_q).$$

Under the piecewise exponential model, the baseline cumulative hazard up to time t given survival to τ_{0q} is

$$\hat{\Lambda}_0^{(m)}(t) = \sum_{t > \tau_{0q}} \hat{\lambda}_0^{(m)}(\tau_q) (\min\{t, \tau_{1q}\} - \tau_{0q}).$$

3.1.3 Nonparametric baseline hazard model—When no distributional assumption is made for $\lambda_0^{(m)}$, the estimator for the cause-specific risk of the primary event within the interval $[t_0, t_1)$, given $\mathbf{x}$, is

$$\hat{\pi}(t_0, t_1; \mathbf{x}) = \exp(\hat{\beta}^{(1)'} \mathbf{x}^{(1)}) \sum_{t_0 \leq u_j < t_1} \hat{\lambda}_0^{(1)}(u_j) \exp\left\{-\sum_{i=1}^M \sum_{t_0 \leq s_k < u_j} \hat{\Lambda}_0^{(i)}(s_k) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)})\right\},$$

where $\hat{\Lambda}_0^{(i)}(s)$ is the cause-specific baseline cumulative hazard function for event i , $\hat{\Lambda}_0^{(1)}(s)$ the primary-event baseline hazard function at time s , and u_j are the observed event times occurring in $[t_0, t_1)$. A weighted Nelson-Aalen estimator (Aalen, 1978) for the non-parametric cause-specific baseline hazard function is

$$\hat{\lambda}_0^{(m)}(t) = \frac{\sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}(t) \delta_{ijk}^{(m)}(t)}{\sum_{i=1}^I \sum_{j=1}^{J_i} \sum_{k=1}^{n_{ij}} w_{ijk} y_{ijk}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)})},$$

which uses Breslow's method for handling ties (Breslow, 1974). The cause-specific baseline cumulative hazard at time t is estimated as

$$\hat{\Lambda}_0^{(m)}(t) = \sum_{u_i^{(m)} \leq t} \hat{\lambda}_0^{(m)}(u_i^{(m)}), \quad (7)$$

with $u^{(m)}$ denoting the observed event times for the m th event type.

For both the piecewise and semiparametric approaches, given a baseline cumulative hazard estimate, the individualized cumulative hazard to time t for an individual with risk profile $\mathbf{x}^{(m)}$ is

$$\hat{\Lambda}^{(m)}(t; \mathbf{x}^{(m)}) = \sum_{u_i^{(m)} \leq t} \hat{\lambda}_0^{(m)}(u_i^{(m)}) \exp(\hat{\beta}^{(m)' } \mathbf{x}^{(m)}).$$

3.2 Variance estimation

Taylor linearization is a non-iterative method to obtain the variance of complex estimators that can be expressed as sums of differentiable functions of parameters and observed data, and it has been widely applied to survey statistics for design-based inference (Woodruff, 1971; Demnati and Rao, 2010). The Taylor-linearized variance can be derived with the empirical influence function $\Delta\{\cdot\}$, which yields the analytic jackknife residual, or Taylor deviate, for each observation. Influence functions have been used to estimate the variance of M-estimators (Hampel, 1974), complex survey statistics (Graubard and Fears, 2005), and the hazard ratios of the proportional hazards model for cohort data (Reid and Crepeau, 1985). Key properties of the influence operator have been previously discussed by Shah (2002) and Deville (1999).

We provide the outline for the variance estimation of $\tau(\hat{t}_0, t_1; \mathbf{x})$ given the Taylor deviates of $\hat{\beta}^{(m)}$ and $\hat{\lambda}_0^{(m)}$. Derivations for the Taylor deviates are provided in the Appendix.

We write the absolute risk estimate for (1) as a sum over the observed event times for the primary cause, $u_1, \dots, u_L$, occurring within $[t_0, t_1)$,

$$\hat{\pi}(t_0, t_1; \mathbf{x}) = \sum_{l=1}^L \hat{\lambda}_0^{(1)}(u_l) \exp(\hat{\beta}^{(1)' } \mathbf{x}^{(1)}) \prod_{m=1}^M \exp\{-\hat{\Lambda}_0^{(m)}(u_l) \exp(\hat{\beta}^{(m)' } \mathbf{x}^{(m)})\} = \sum_{l=1}^L \prod_{m=1}^M \theta^{(m)}(u_l),$$

where

$$\theta^{(1)}(u_l) = \hat{\lambda}_0^{(1)}(u_l) \exp(\hat{\beta}^{(1)' } \mathbf{x}^{(1)}) \exp\{-\hat{\Lambda}_0^{(1)}(u_l) \exp(\hat{\beta}^{(1)' } \mathbf{x}^{(1)})\},$$

and for all competing events, $m > 1$,

$$\theta^{(m)}(u_l) = \exp\{-\hat{\Lambda}_0^{(m)}(u_l) \exp(\hat{\beta}^{(m)' } \mathbf{x}^{(m)})\}.$$

The Taylor deviates for the absolute risk estimate are weighted sums of the deviates of θ , weighted by the partial derivatives of the absolute risk with respect to θ . The form of the ijk th deviate is

$$\Delta_{ijk}\{\hat{\pi}(t_0, t_1; \mathbf{x})\} = \sum_{l=1}^L \sum_{m=1}^M \frac{\partial \hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \theta^{(m)}(u_l)} \Delta_{ijk}\{\theta^{(m)}(u_l)\} = \sum_{m=1}^M \Delta_{ijk}^{(m)}\{\hat{\pi}(t_0, t_1; \mathbf{x})\}, \quad (8)$$

where $\Delta_{ijk}^{(m)}\{\hat{\pi}(t_0, t_1; \mathbf{x})\}$ is the total influence due to a specific cause.

Given the Taylor deviates $\Delta_{ijk}\{\tau(\hat{t}_0, t_1; \mathbf{x})\}$ in (8), the variance estimate for the absolute risk is

$$\widehat{\text{Var}}(\hat{\pi}(t_0, t_1; \mathbf{x})) = \sum_{i=1}^I \frac{J_i}{J_i - 1} \sum_{j=1}^{J_i} (\Delta_{ij} \{\hat{\pi}(t_0, t_1; \mathbf{x})\} - \bar{\Delta}_{i..} \{\hat{\pi}(t_0, t_1; \mathbf{x})\})^2 \quad (9)$$

where $\Delta_{ij} \{\hat{\pi}(t_0, t_1; \mathbf{x})\} = \sum_{k=1}^{n_{ij}} \Delta_{ijk} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}$ is the sum over all units of the ij th PSU and $\bar{\Delta}_{i..} \{\hat{\pi}(t_0, t_1; \mathbf{x})\} = J_i^{-1} \sum_{j=1}^{J_i} \Delta_{ij} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}$ is the average of these totals.

3.3 Novel measures for competing risk model selection

Standard methods for model diagnosis and covariate selection of the cause-specific hazard model have limited applicability in the absolute risk setting because they do not consider the other components of the model. Refinement of the competing risk models is particularly challenging, owing to their second-order effects on the absolute risk estimate. In the following sections, we introduce two measures based on the absolute risk influence components (8) to assist with the development of the prediction model.

3.3.1 Relative influence and competing risk selection—When there are multiple competing risks, there may be uncertainty about how they should be classified, that is, how narrowly events should be defined. The relative information a competing cause contributes to the absolute risk model relative to all competing causes is one basis for deciding whether a cause is sufficiently informative to warrant a separate hazard model. For this purpose, we propose the measure $\rho^{(m)}$,

$$\rho^{(m)}(t_0, t_1) = \int \sum_{i,j,k} \Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}^2 / \sum_{m=2}^M \sum_{i,j,k} \Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}^2 \mathbf{p}(t_0; \mathbf{x}) d\mathbf{x}, \quad (10)$$

where $\mathbf{p}(t_0; \mathbf{x})$ denotes the distribution of the covariates in the population at t_0 . Thus, $\rho^{(m)}(t_0, t_1)$ is the expected proportion of variation in a risk prediction over the interval $[t_0, t_1]$ that is due to the m th cause out of all competing causes.

An empirical estimate of Equation (10) is,

$$\hat{\rho}^{(m)}(t_0, t_1) = \sum_l \left\{ \sum_{i,j,k} \Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x}_l)\}^2 / \sum_{m=2}^M \sum_{i,j,k} \Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x}_l)\}^2 \right\} \hat{\mathbf{p}}(\mathbf{x}_l(t_0)) \quad (11)$$

where $\mathbf{x}_l(t_0)$ is the l th of $1, \dots, L$ unique covariate combinations in the sample at time t_0 and $\hat{\mathbf{p}}(\mathbf{x}_l(t_0)) = w_l / \sum_j w_j$ is the population based frequency of $\mathbf{x}_l(t_0)$.

[Clarify expectation here.] Letting $\nu^{(m)}(t_0, t_1; \mathbf{x}) = E[\Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}^2]$, it is easy to see that Equation (12) is a consistent estimate of

$$\rho^{(m)}(t_0, t_1) = \int \left\{ \nu^{(m)}(t_0, t_1; \mathbf{x}) / \sum_{k=2}^M \nu^{(k)}(t_0, t_1; \mathbf{x}) \right\} \mathbf{p}(t_0; \mathbf{x}) d\mathbf{x},$$

Further, if the second moments for each squared influence $\Delta_{ijk}^{(m)} \{\hat{\pi}(t_0, t_1; \mathbf{x})\}^2$, are finite, then $\hat{\rho}^{(m)}(t_0, t_1)$ converges to a Gaussian distribution.

3.3.2 Competing risk covariate selection—If a risk factor is going to be added to the competing risk model, it is desirable that its addition increases the precision of the competing cause's parameters. Suppose the m th competing event's hazard model is defined by q_1 factors (model H_1) and consider the nested model (H_0) with $q_0 < q_1$ factors. The expected average variance reduction (reduction per variable) due to the larger model is

$$\phi(t_0, t_1)_{H_1|H_0}^{(m)} = \int \frac{\sum_{i,j,k} \left(\Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_0 \}^2 / q_0 - \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_1 \}^2 / q_1 \right)}{\sum_{i,j,k} \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_0 \}^2 / q_0} \mathbf{p}(t_0; \mathbf{x}) d\mathbf{x},$$

where $\mathbf{x}$ are the covariates corresponding to model H_1 and $\mathbf{p}(t_0; \mathbf{x})$ can be estimated as described in Section (3.3.3).

Because of the relationship between the influence curve of a functional and its variation, ϕ is conceptually similar to the partial F test in linear modeling, with the key distinctions that ϕ is a function of the projection interval and not guaranteed to be strictly positive for all time periods. Also, since no large-sample statistical test is available for $\phi(t_0, t_1)_{H_1|H_0}^{(m)}$, we instead use the percentage of the variance reduction as a selection criterion.

3.3.3 Relative influence and competing risk selection—When there are multiple competing risks, there may be uncertainty about how they should be classified, that is, how narrowly events should be defined. The relative information a competing cause contributes to the absolute risk model relative to all competing causes is one basis for deciding whether a cause is sufficiently informative to warrant a separate hazard model. For this purpose, we propose the measure

$$\hat{\rho}^{(m)}(t_0, t_1) = \sum_l \left\{ \frac{\sum_{i,j,k} \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}_l) \}^2}{\sum_{m=2}^M \sum_{i,j,k} \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}_l) \}^2} \right\} \hat{\mathbf{p}}(\mathbf{x}_l(t_0)) \quad (12)$$

where $\mathbf{x}_l(t_0)$ is the l th of $1, \dots, L$ unique covariate combinations in the sample at time t_0 and $\hat{\mathbf{p}}(\mathbf{x}_l(t_0)) = w_l / \sum_j w_j$ is the population based frequency of $\mathbf{x}_l(t_0)$. Thus, $\hat{\rho}^{(m)}(t_0, t_1)$ estimates the population expected proportion of variation in a risk prediction over the interval $[t_0, t_1]$ that is due to the m th cause out of all competing causes.

We let $\nu^{(m)}(t_0, t_1; \mathbf{x}) = E[\Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) \}^2]$, where the expectation is taken with respect to all possible samples from the source population consistent with the particular design (Särndal et al, 1992, p. 40). It is easy to see that Equation (12) is a consistent estimate of

$$\rho^{(m)}(t_0, t_1) = \int \left\{ \nu^{(m)}(t_0, t_1; \mathbf{x}) / \sum_{k=2}^M \nu^{(k)}(t_0, t_1; \mathbf{x}) \right\} \mathbf{p}(t_0; \mathbf{x}) d\mathbf{x},$$

where $\mathbf{p}(t_0; \mathbf{x})$ denotes the distribution of the covariates in the population at t_0 . Further, if the second moments for each squared influence $\Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) \}^2$, are finite, then $\hat{\rho}^{(m)}(t_0, t_1)$ converges to a Gaussian distribution.

3.3.4 Competing risk covariate selection—If a risk factor is going to be added to the competing risk model, it is desirable that its addition increases the precision of the competing cause's parameters. Suppose the m th competing event's hazard model is defined

by q_1 factors (model H_1) and consider the nested model (H_0) with $q_0 < q_1$ factors. The expected average variance reduction (reduction per variable) due to the larger model is

$$\phi(t_0, t_1)_{H_1|H_0}^{(m)} = \sum_l \frac{\sum_{i,j,k} \left(\Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_0 \}^2 / q_0 - \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_1 \}^2 / q_1 \right)}{\sum_{i,j,k} \Delta_{ijk}^{(m)} \{ \hat{\pi}(t_0, t_1; \mathbf{x}) | H_0 \}^2 / q_0} \hat{\mathbf{p}}(t_0; \mathbf{x}),$$

where $\mathbf{x}$ are the covariates corresponding to model H_1 and $\hat{\mathbf{p}}(t_0; \mathbf{x})$ is computed as described in Section (3.3.3).

Because of the relationship between the influence curve of a functional and its variation, ϕ is conceptually similar to the partial F test in linear modeling, with the key distinctions that ϕ is a function of the projection interval and not guaranteed to be strictly positive for all time

periods. Also, since no large-sample statistical test is available for $\phi(t_0, t_1)_{H_1|H_0}^{(m)}$, we instead use the percentage of the variance reduction as a selection criterion.

3.4 Remarks

3.4.1 Alternative hazard models—Our methodology can be applied to a broader set of hazard models by generalizing Equation (2), so that

$$\lambda^{(m)}(t; \mathbf{x}^{(m)}) = \lambda_0^{(m)}(t) r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)}) \quad (13)$$

where $r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)})$ is a strictly positive scalar function of $\boldsymbol{\beta}^{(m)}$ and covariates $\mathbf{x}^{(m)}$. If the $\mathbf{x}^{(m)}$ include a continuous exposure variable, $\mathbf{x}_0^{(m)}$ (e.g. cumulative pack-years smoked or absorbed radiation) one might consider an excess relative risk (ERR) model (Preston et al, 1993), such as the linear ERR model

$$r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)}) = (1 + \beta_0^{(m)} x_0^{(m)}) \exp(\boldsymbol{\beta}_1^{(m)'} \mathbf{x}_1^{(m)}).$$

To obtain the absolute risk estimates and variances for any model of the form of Equation (13), one could derive estimating equations $\boldsymbol{\beta}^{(m)}$ with a weighted partial likelihood, replacing all instances of $\exp(\boldsymbol{\beta}^{(m)'} \mathbf{x}^{(m)})$ and its first and second derivative with $r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)})$,

$\frac{\partial r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)})}{\partial \boldsymbol{\beta}^{(m)}}$, and $\frac{\partial^2 r(\boldsymbol{\beta}^{(m)}; \mathbf{x}^{(m)})}{\partial \boldsymbol{\beta}^{(m)} \partial \boldsymbol{\beta}^{(m)'}}$, respectively. Given the estimator $\hat{\boldsymbol{\beta}}^{(m)}$, one would substitute $r(\hat{\boldsymbol{\beta}}^{(m)}; \mathbf{x}^{(m)})$ and its first derivative with their corresponding quantities in the formulae of this paper to derive point estimates and variances for absolute risk.

3.4.2 Absolute risk estimation for a subpopulation—Absolute risk estimates for a subpopulation A can be derived with the modified sampling weights $w_{ijk}^A = I_{ijk}^A w_{ijk}$, where I_{ijk}^A is an indicator for the k th individual's membership in subpopulation A , and the n_{ij} are replaced with $n_{ij}^A = \sum_{k=1}^{n_{ij}} I_{ijk}^A$. The computation in Eq. (9) is otherwise unchanged (Korn and Graubard, 1999, Chapter 5.4).

3.4.3 Clustering—When individuals of the survey cohort are drawn from clusters within PSUs, the point estimates we present still apply but the variance computation must be modified to account for the correlation between individuals from the same cluster. To accommodate clustering, use k to index n_{ij} clusters within PSU and l (1, ..., L_{ijk}) to index

individuals within clusters. The totals of Eq. (9) are taken over all individuals and clusters of a given PSU and stratum, so that the PSU-level variance becomes

$$\widehat{\text{Var}}(\hat{\pi}(t_0, t_1; x)) = \sum_{i=1}^I \frac{J_i}{J_i - 1} \sum_{j=1}^{J_i} (\Delta_{ij..} \{\hat{\pi}(t_0, t_1; x)\} - \bar{\Delta}_{i...} \{\hat{\pi}(t_0, t_1; x)\})^2,$$

$$\text{with } \Delta_{ij..} \{\hat{\pi}(t_0, t_1; x)\} = \sum_{k=1}^{n_{ij}} \sum_{l=1}^{L_{ijk}} \Delta_{ijkl} \{\hat{\pi}(t_0, t_1; x)\}.$$

4 Application

4.1 Data and Methods

To illustrate our methodology, we built population-based models of cause-specific deaths from the NHANES I and II, pooling data from the two surveys to increase sample size and improve the accuracy of our estimates. The design of each cohort are detailed in Section 2.2.1.

We classified underlying causes of death as cancer (140–239), cardiovascular disease (390–459), or ‘other causes’, according to International Classification of Diseases, Ninth Revision. In the pooled NHANES I/II sample of 23, 659 individuals, who were followed up to December 31, 1993, cardiovascular disease accounted for 49% of observed deaths in men and 51% in women, while cancer was the underlying cause for 25% of observed deaths in men and 24% in women. As there were notable differences in the leading causes of cancer-related and other deaths by gender (Table 1), we built separate cause-specific absolute risk models for men and women. We followed the approach described in Section ?? to obtain subpopulation estimates for each gender.

The covariates examined were baseline race, income, marital status, body mass index (BMI; kg/height in meters squared), smoking, diabetes diagnosis, hypertension, alcohol consumption, and serum cholesterol level. BMI, hypertension, and serum cholesterol levels were based on physical measurements conducted as part of NHANES; all remaining variables were self-reported.

For analyses with the piecewise exponential model, the baseline hazard function was divided into the seven age intervals [30, 55), [55, 60), [60, 65), [65, 70), [70, 75), [75, 80), [80, 100).

4.2 Model Selection

4.2.1 Cause-Specific Risk Factors—Using the weighted pseudo-likelihood described in Section 3.1.1, we selected risk factors with backward stepwise regression and the design-based significance test suggested by Rao and Scott (1987) with a 10% significance level. Confidence intervals for the hazard ratios were constructed by exponentiating the lower and upper endpoints for the corresponding large-sample confidence interval of the log-hazard ratios. For men and women, smoking and alcohol consumption had significant associations with all three causes of death (Table 2). Although the set of significant risk factors for cardiovascular and other death were similar for men and women—each including weight, diabetic status, and sociodemographic variables—, fewer risk factors were associated with cancer deaths among women.

4.2.2 Competing Risk Models—We illustrate the use of the relative influence measure ρ with the semiparametric cancer mortality model for women. The overall expected relative influences for projections between ages 50 and 80 years show that competing deaths due to cardiovascular disease constitute the majority of information about competing deaths

(94.7%, Table 3). This suggests that there is no need to further separate events among other causes of death. However, when we examined how the relative influence varied by smoking and BMI status, we found that CVD and other causes of death were equally informative for predictions of underweight women (Table 3).

Table 3 also shows the use of ϕ for evaluating the reduction in variance with the inclusion of the smoking variable for each competing risk model (i.e. CVD or other mortality). Overall, the inclusion of the smoking variable resulted in an average 34% and 17% reduction in the variance of the CVD and other competing risk models, respectively. Stratifying ϕ on BMI showed a lower variance reduction for the CVD model in the underweight subpopulation. The reductions within each model were otherwise similar across BMI categories.

4.3 Finite-Sample Risk of Cause-Specific Death

Based on an internal cross-validation of the NHANES I/II the cause-specific mortality models were well-calibrated at the population level (data not shown). We thus report selected absolute risk estimates for CVD and cancer death between ages 65 and 75 for a married White male with high income, a BMI of 22, who consumes more than one alcoholic drink per week and less than one per day on average, has no diabetes, no hypertension, and a baseline serum cholesterol level of 130 mg/dl (Table 4). To illustrate the impact of smoking for this particular individual, Table 4 shows risk estimates for each smoking exposure type. In these examples, the piecewise exponential and the semiparametric hazard functions yielded similar absolute risk estimates and standard errors.

Owing to the asymptotic equivalence of the influence-based and jackknife variance estimators, we compared our influence-based variance estimates to those obtained from a jackknife resampling procedure as a verification of our derivation. For computational efficiency we used a subsample of 20 randomly selected males from each of the 169 stratum-PSUs in the NHANES I/II dataset. Letting i index the full sample and j the subsample, population estimates were obtained from the subsample by modifying the weight of the k th individual in the subsample to $w_k^* = w_k (\sum_i w_i / \sum_j w_j)$. There was good agreement between the influence and jackknife variance estimators (Table 4).

A useful application of the finite-population mortality models is the estimation of the preventable number of cause-specific deaths in the male US population represented by NHANES I/II over a ten-year period under two smoking intervention scenarios. For the ‘cessation’ scenario, we treated current smokers as former smokers at the beginning of the projection period (Table 5). For the ‘no exposure’ scenario, all former and current smokers were regarded as never smokers. We do not consider the feasibility of these scenarios but, instead, use them to suggest the possible public health utility of survey-based prediction models for quantifying the impact of disease or death outcomes resulting from changes in the distribution of risk factors in the population.

4.4 Population Projections of Cause-Specific Death

Although the cause-specific mortality models were internally well-calibrated, to assess their value for making projections into the future the models need to be validated in independent samples of later birth cohorts. We, therefore, performed an external validation with data from the NHANES III. This survey was conducted between 1988 and 1994 and collected interview data on 33,994 individuals, of whom 31,111 also underwent a medical examination. Sampling weights for obtaining national estimates are provided on the public-use data files. For variance estimation, we approximate the design by the sampling of two pseudo-PSUs from each of 49 pseudo-strata (Ezzati et al, 1992).

We used data on 12,517 individuals from NHANES III who were ages 30 to 75 at the time of the interview, and for whom death information was available through December 31, 2006. We first compared the number of cardiovascular deaths predicted by a piecewise exponential absolute risk model to the observed counts. Within the NHANES III sample, the model was adequately calibrated (Table 6). We then projected the cardiovascular death estimates to the US population for a ten-year period and report the counts and standard errors for the population predictions in Table 6. These results demonstrate the importance of population-based measures of validation. While the sample comparisons suggest good calibration for other race in men and women (Sample Obs./Exp.), because of over sampling in these minority population, these ratios do not necessarily reflect the accuracy of the model at the population level (Population Obs./Exp.).

5 Discussion

We have shown how to estimate population-based absolute risk from survey data with prospective follow-up, using a cause-specific formulation. Our methodology builds on work for estimating survival functions from complex survey cohort under the Cox model (Binder, 1992; Lin and Wei, 1989; Lin, 2000), which are currently implemented by a number of statistical packages including the R package *survey* (Lumley, 2011, 2004), PROC SURVEYPHREG (SAS v.9.22) and SUDAAN's SURVIVAL procedure (Version 10.0). We extend these ideas to estimate an individualized probability of disease within a given time interval in the presence of competing events. The framework we have presented can accommodate multiple competing risks and each event type has some flexibility in the form of its hazard model. We derived Taylorlinearized variance estimators for the absolute risk models that, when working with large survey cohorts, could have computational advantages over iterative methods. Further, the influence methodology for variance estimation allows the calculation of two novel measures for guiding selection of the competing risk models.

Our work builds on the cohort methods presented by Benichou and Gail (1990) for simple cohorts. Like these authors, we have presented parametric and semiparametric approaches for estimating the cause-specific hazard function. For cohorts drawn with simple random sampling, Benichou and Gail showed through simulation studies that correctly specified parametric models for the hazard could considerably improve the efficiency of absolute risk estimates in comparison to semiparametric methods. However, the efficiency of estimates obtained with the piecewise exponential model in comparison to estimates obtained with Kaplan-Meier's method approached equality as the number of time intervals (pieces) increased. The piecewise exponential should, therefore, be preferred to more general models when it provides a good approximation to the shape of the baseline hazard without requiring a large number of time intervals. The reasonableness of a parsimonious piecewise exponential can be assessed by inspecting the Kaplan-Meier plot of the baseline hazard function. The piecewise exponential model for the baseline hazard is also useful when the exact time of event for each individual in a survey is not available, for example, when age at death is only given in categories (e.g. ages 30 – 34, 35 – 39, ...). It would be useful to extend even more flexible parametric forms for the baseline hazard function, such as the simple mixed-model proposed by Cai et al (2002), to survey data. One approach to estimation with general hazards would be the pseudo-likelihood proposed by Kalbfleisch and Lawless (1988), that extends the approach by Binder (1992).

In related work, methods for estimating absolute risk from individually matched nested case-control (Langholz and Borgan, 1997) and from case-cohort studies (Langholz and Jiao, 2007) also use weights to accommodate the sampling of cases, with various control sampling designs. However, the hazard for the competing causes discussed by Langholz and

Borgan (1997) and Langholz and Jiao (2007) does not depend on covariates, and thus the variance computations are appreciably simpler.

Although the family of cause-specific hazard models we consider allow some flexibility in the relative risk component, all models in the family have the relative risk acting multiplicatively on the baseline hazard. An additive hazard model might also be of interest. Using standard cohort data, Shen and Cheng (1999) presented methods for absolute risk estimation under Aalen's additive model. Future work will also consider extending this model to the survey setting.

Accommodating the survey design in the model building is important, as ignoring differential sampling rates in a survey sample can lead to biased population estimates (Kish and Frankel, 1974; Korn and Graubard, 1995). In addition, designs that include clustering induce correlation between observations from the same cluster. Inference methods that ignore this structure will underestimate the standard errors of population estimates (Williams, 1995). Thus weights informative about the sampling design should be incorporated into the inferential procedure to obtain unbiased estimates with accurate measures of uncertainty.

Our methods yield consistent estimates of absolute risk for the finite and superpopulation but the variance estimators are only valid for the finite population. The superpopulation framework regards the finite population defining the sampling frame of the survey as a sample from an infinite population (Graubard and Korn, 2002). When predictions are made for a future, or an otherwise distinct, cohort, the superpopulation framework may be more relevant. In future work, we plan to extend our methods to superpopulation inference and give more attention to the practical distinctions between design-based and model-based inference for absolute risk prediction. The large-sample results of Lin (2000) for superpopulation variance of semiparametric hazard models could be the starting point for deriving inferential methods for model-based absolute risk. Lin stated that the extra variance arising from the superpopulation component is often numerically negligible when inclusion probabilities are small, under the assumption that the superpopulation model was correctly specified. With respect to survey-based risk prediction, it is unclear when adopting a model-based framework is recommendable.

The finite-population methods we have developed could be important for public health research. The presented techniques will enable survey datasets linked to health outcomes to be a resource for building and validating population-based absolute risk prediction models. Also, by projecting the expected number of events for the current population to some future point in time, a validated risk model could aid the design of surveys of health outcomes and the identification of changing trends in the state of a population's health. These uses will require further development of population-based validation methodology and performance criteria.

Acknowledgments

This research was supported by the intramural research program of the National Cancer Institute.

We thank the reviewers for their helpful comments. We are grateful to Dr. Barry Graubard for suggestions he provided to us during the writing of this paper.

Appendix

Derivatives and Taylor deviates for the piecewise exponential hazard function

To simplify the notation, we express the absolute risk estimate of (4) in a more compact form,

$$\hat{\pi}(\tau_{0n_0}, \tau_{1n_1}; \mathbf{x}) = \sum_{q=n_0}^{n_1} \hat{S}(\tau_q) A_q (1 - B_q)$$

where

$$A_q = \frac{\hat{\lambda}_0^{(1)}(\tau_q) \exp(\hat{\beta}^{(1)'} \mathbf{x}^{(1)})}{\sum_i \hat{\lambda}_0^{(i)}(\tau_q) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)})}, \quad B_q = \exp\left\{-\sum_i \hat{\lambda}_0^{(i)}(\tau_q) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)}) (\tau_{1q} - \tau_{0q})\right\},$$

and $\hat{S}(\tau_q)$ is as defined in Equation (5). Then the deviates for $\tau(\tau_{0n_0}, \tau_{1n_1}; \mathbf{x})$ are

$$\Delta_{ijk}\{\hat{\pi}(\tau_{0n_0}, \tau_{1n_1}; \mathbf{x})\} = \sum_{q=n_0}^{n_1} \left[\hat{S}(\tau_q) (1 - B_q) \Delta_{ijk}\{A_q\} + A_q (1 - B_q) \Delta_{ijk}\{\hat{S}(\tau_q)\} - A_q \hat{S}(\tau_q) \Delta_{ijk}\{B_q\} \right].$$

Taking each component in turn, the deviates for A_q are

$$\Delta_{ijk}\{A_q\} = T_q^{-1} \Delta_{ijk}\{T_q^{(1)}\} - \frac{T_q^{(1)}}{T_q^2} \sum_{i=1}^M \Delta_{ijk}\{T_q^{(i)}\}$$

where $T_q^{(i)} = \hat{\lambda}_0^{(i)}(\tau_q) \exp(\hat{\beta}^{(i)'} \mathbf{x}^{(i)})$ and $T_q = \sum_{i=1}^M T_q^{(i)}$, with deviates

$$\Delta_{ijk}\{T_q^{(1)}\} = \exp(\hat{\beta}^{(1)'} \mathbf{x}^{(1)}) \Delta_{ijk}\{\hat{\lambda}_0^{(1)}(\tau_q)\} + \mathbf{x}^{(1)} T_{1q} \Delta_{ijk}\{\hat{\beta}^{(1)}\}.$$

The deviates for B_q are

$$\Delta_{ijk}\{B_q\} = - \sum_{m=1}^M \left[\exp(\hat{\beta}^{(m)'} \mathbf{x}^{(m)}) (\tau_{1q} - \tau_{0q}) B_q \left(\mathbf{x}^{(m)} \hat{\lambda}_0^{(m)}(\tau_q) \Delta_{ijk}\{\hat{\beta}^{(m)'}\} + \Delta_{ijk}\{\hat{\lambda}_0^{(m)}(\tau_q)\} \right) \right]$$

For $q > n_0$, we note that $\hat{S}(\tau_q) = \prod_{l=n_0}^{q-1} B_l$ so that

$$\Delta_{ijk}\{\hat{S}(\tau_q)\} = \hat{S}(\tau_q) \sum_{l=n_0}^{q-1} B_l^{-1} \Delta_{ijk}\{B_l\},$$

and $\Delta_{ijk}\{\hat{S}(\tau_q)\}$ is zero when $q = n_0$.

The Taylor deviates for A_q , B_q and $\hat{S}(\tau_q)$ are each functions of $\hat{\lambda}_0^{(m)}$ and $\hat{\beta}^{(m)}$. For $\hat{\lambda}_0^{(m)}$, we have

$$\Delta_{ijk}\{\hat{\lambda}_0^{(m)}(\tau_q)\} = D^{(m)}(\tau_q)^{-1} \left[w_{ijk} \delta_{ijk}^{(m)}(t_{ijk}) I(\tau_{0q} \leq t_{ijk} < \tau_{1q}) - \hat{\lambda}_0^{(m)}(\tau_q) \Delta_{ijk}\{D^{(m)}(\tau_q)\} \right]$$

where

$$\Delta_{ijk}\{D^{(m)}(\tau_q)\} = \mathcal{A}_{ijk}(\tau_q) w_{ijk} \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}) + \left[\sum_{i,j,k} \mathbf{x}_{ijk}^{(m)} \mathcal{A}_{ijk}(\tau_q) w_{ijk} \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}) \right] \Delta_{ijk}\{\hat{\beta}^{(m)}\}. \quad (14)$$

with $\mathcal{A}_{ijk}(\tau_q)$ defined in Equation (6). The Taylor deviates for each $\hat{\beta}^{(m)}$ are

$$\Delta_{ijk}\{\hat{\beta}^{(m)}\} = \mathcal{H}(\hat{\beta}^{(m)})^{-1} w_{ijk} \delta_{ijk}^{(m)}(t_{ijk}) \{\mathbf{x}_{ijk}^{(m)} - \bar{\mathbf{H}}(\hat{\beta}^{(m)}, t_{ijk})\} \quad (15)$$

where $\mathcal{H}(\hat{\beta}^{(m)})$ is the second partial derivative of the pseudo-likelihood,

$$\begin{aligned} \mathcal{H}(\hat{\beta}^{(m)}) &= - \left[\sum_{i,j,k} \mathbf{x}_{ijk}^{(m)} \mathbf{x}_{ijk}^{(m)'} h_{ijk}^{(m)} \right] + \bar{\mathbf{H}}(\hat{\beta}^{(m)}, t_{ijk}) \bar{\mathbf{H}}(\hat{\beta}^{(m)}, t_{ijk})', \\ h_{ijk}^{(m)} &= w_{ijk} y_{ijk}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}) / \sum_{i,j,k} w_{ijk} y_{ijk}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}), \end{aligned}$$

and $\bar{\mathbf{H}}(\hat{\beta}^{(m)}, t_{ijk})$ is defined in Equation (3). Thus, the deviates for $\hat{\beta}^{(m)}$ are equivalent to the per-observation update in a Newton-Raphson optimization algorithm where the objective function is the weighted pseudo-likelihood of the Cox regression model.

Derivatives and Taylor deviates for the semiparametric hazard function

Denote the $N^{(m)}$ ordered observed event times occurring within $[t_0, t_1]$ for the m th cause as $u_1^{(m)} < u_2^{(m)} < \dots < u_{N^{(m)}}^{(m)}$. In terms of these event times, Equation (1) becomes

$$\hat{\pi}(t_0, t_1; \mathbf{x}) = \sum_{i=1}^{N^{(1)}} \exp(\hat{\beta}^{(1)'} \mathbf{x}^{(1)}) \hat{\lambda}_0^{(1)}(u_i^{(1)}) \prod_{j=1}^M \left(\frac{\hat{S}_0^{(j)}(u_i^{(1)})}{\hat{S}_0^{(j)}(u_1^{(1)})} \right)^{\exp(\hat{\beta}^{(j)'} \mathbf{x}^{(j)})} = \sum_{i=1}^{N^{(1)}} \hat{\pi}(u_i^{(1)}). \quad (16)$$

As with the piecewise model, we determine the derivative and deviates for each component of (16). For the $\hat{\beta}^{(j)}$, the deviate is

$$\frac{\partial \hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{\beta}^{(j)}} = \mathbf{x}^{(j)} \left[\hat{\pi}(t_0, t_1; \mathbf{x}) + \exp(\hat{\beta}^{(j)'} \mathbf{x}^{(j)}) \sum_{i=1}^{N^{(1)}} \log \left(\hat{S}_0^{(j)}(u_i^{(1)}) / \hat{S}_0^{(j)}(u_1^{(1)}) \right) \hat{\pi}(u_i^{(1)}) \right],$$

when $j = 1$ and

$$\frac{\partial \hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{\beta}^{(j)}} = \mathbf{x}^{(j)} \exp(\hat{\beta}^{(j)'} \mathbf{x}^{(j)}) \sum_{i=1}^{N^{(1)}} \log \left(\hat{S}_0^{(j)}(u_i^{(1)}) / \hat{S}_0^{(j)}(u_1^{(1)}) \right) \hat{\pi}(u_i^{(1)})$$

for competing causes. The Taylor deviates for each $\hat{\beta}^{(j)}$ are the same as given by Equation (15) of the piecewise model.

The derivatives for the baseline hazard components are

$$\frac{\partial \hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{\lambda}_0^{(1)}(u_i^{(1)})} = \hat{\lambda}_0^{(1)}(u_i^{(1)})^{-1} \hat{\pi}(u_i^{(1)}).$$

The Taylor deviates for the baseline hazard of cause m at observed event time t are

$$\Delta_{ijk}\{\hat{\lambda}_0^{(m)}(t)\} = \frac{\partial \hat{\lambda}_0^{(m)}(t)}{\partial N^{(m)}(t)} \Delta_{ijk}\{N^{(m)}(t)\} + \frac{\partial \hat{\lambda}_0^{(m)}(t)}{\partial Y^{(m)}(t)} \Delta_{ijk}\{Y^{(m)}(t)\},$$

where

$$N^{(m)}(t) = \sum_{i,j,k} w_{ijk} y_{ijk}^{(m)}(t) \delta_{ijk}^{(m)}(t)$$

and

$$Y^{(m)}(t) = \sum_{i,j,k} w_{ijk} y_{ijk}^{(m)}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}).$$

In terms of these quantities, the Taylor deviates are

$$\Delta_{ijk}\{\hat{\lambda}_0^{(m)}(t)\} = \{Y^{(m)}(t)\}^{-1} (w_{ijk} y_{ijk}^{(m)}(t) \delta_{ijk}^{(m)}(t) - \hat{\lambda}_0^{(m)}(t) \Delta_{ijk}\{Y^{(m)}(t)\})$$

with

$$\Delta_{ijk}\{Y^{(m)}(t)\} = w_{ijk} y_{ijk}^{(m)}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}) + \left[\sum_{i,j,k} \mathbf{x}_{ijk} w_{ijk} y_{ijk}^{(m)}(t) \exp(\hat{\beta}^{(m)'} \mathbf{x}_{ijk}^{(m)}) \right] \Delta_{ijk}\{\hat{\beta}^{(m)}\}.$$

We note that the hazard deviates for the piecewise and semiparametric model in Equation (14) are equivalent when each interval of the piecewise model contains exactly one observed event time.

The final components are the survival functions. The derivatives for each $\hat{S}_0^{(m)}(u_j^{(1)})$ are

$$\frac{\partial \hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{S}_0^{(m)}(u_j^{(1)})} = \exp(\hat{\beta}^{(m)'} \mathbf{x}^{(m)}) \hat{S}_0^{(m)}(u_j^{(1)})^{-1} \hat{\pi}(u_j^{(1)}).$$

From the semiparametric estimate of Equation (7), the Taylor deviates for the baseline survival up to time $u_j^{(1)}$ for the m th risk type are

$$\Delta_{ijk}\{\hat{S}_0^{(m)}(u_j^{(1)})\} = -\hat{S}_0^{(m)}(u_j^{(1)}) \sum_{u_q^{(m)} \leq u_j^{(1)}} \Delta_{ijk}\{\hat{\lambda}_0^{(m)}(u_q^{(m)})\}.$$

Combining these results, the expression for the Taylor deviates of $\hat{\pi}(t_0, t_1; \mathbf{x})$ are

$$\begin{aligned} & \Delta_{ijk}\{\hat{\pi}(t_0, t_1; \mathbf{x})\} \\ &= \sum_{m=1}^M \frac{\hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{\beta}^{(m)}} \Delta_{ijk}\{\hat{\beta}^{(m)}\} + \sum_{q=1}^{N^{(1)}} \frac{\hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{\lambda}_0^{(1)}(u_q^{(1)})} \Delta_{ijk}\{\hat{\lambda}_0^{(1)}(u_q^{(1)})\} \\ &+ \sum_{q=1}^{N^{(1)}} \sum_{m=1}^M \frac{\hat{\pi}(t_0, t_1; \mathbf{x})}{\partial \hat{S}_0^{(m)}(u_q^{(1)})} \Delta_{ijk}\{\hat{S}_0^{(m)}(u_q^{(1)})\}. \end{aligned}$$

References

- Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat.* 1978;701–726.
- Benichou J, Gail MH. Estimates of absolute cause-specific risk in cohort studies. *Biometrics.* 1990; 46(3):813–826. [PubMed: 2242416]
- Benichou J, Gail MH. Methods of inference for estimates of absolute risk derived from population-based case-control studies. *Biometrics.* 1995; 51(1):182–194. [PubMed: 7766773]
- Binder D. Fitting Cox's proportional hazards models from survey data. *Biometrika.* 1992; 79(1):139–147.
- Breslow N. Covariance analysis of censored survival data. *Biometrics.* 1974; 30(1):89–99. [PubMed: 4813387]
- Cai T, Hyndman RJ, Wand MP. Mixed model-based hazard estimation. *J Comput Graph Stat.* 2002; 11(4):784–798.
- Cox C, Rothwell S, Madans J, Finucane F, Freid V, Kleinman J, Barbano H, Feldman J. Plan and operation of the NHANES I Epidemiologic Followup Study, 1987. *Vital Health Stat Ser 1.* 1992; 27:1–190. [PubMed: 1413562]
- Demnati A, Rao JNK. Linearization variance estimators for model parameters from complex survey data. *Surv Methodol.* 2010; 36(2):193–201.
- Deville J. Variance estimation for complex statistics and estimators: linearization and residual techniques. *Surv Methodol.* 1999; 25(2):193–204.
- Engel A, Murphy R, Maurer K, Collins E. Plan and operation of the HANES I augmentation survey of adults 25–74 years United States, 1974–1975. *Vital Health Stat Ser 1.* 1978; 14:1–110. [PubMed: 695333]
- Ezzati T, Massey J, Waksberg J, Chu A, Maurer K. Sample design: Third national health and nutrition examination survey. *Vital Health Stat Ser 2.* 1992; 113:1–35. [PubMed: 1413563]
- Fine J, Gray R. A proportional hazards model for the subdistribution of a competing risk. *J Am Stat Assoc.* 1999;496–509.
- Graubard B, Korn E. Inference for superpopulation parameters using sample surveys. *Statistical Science.* 2002; 17(1):73–96.
- Graubard BI, Fears TR. Standard errors for attributable risk for simple and complex sample designs. *Biometrics.* 2005; 61(3):847–855. [PubMed: 16135037]
- Gray RJ. Weighted analyses for cohort sampling designs. *Lifetime Data Anal.* 2009; 15(1):24–40. [PubMed: 18712477]
- Hampel FR. Influence curve and its role in robust estimation. *J Am Stat Assoc.* 1974; 69(346):383–393.
- Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Stat Med.* 1988; 7(1–2):149–160. [PubMed: 3353602]

- Kish L, Frankel MR. Inference from complex samples. *J R Stat Soc Series B Stat Methodol*. 1974; 36(1):1–22.
- Korn EL, Graubard BI. Examples of differing weighted and unweighted estimates from a sample survey. *Am Stat*. 1995; 49(3):291–295.
- Korn, EL.; Graubard, BI. *Wiley Series in Probability and Statistics*. New York: Wiley; 1999. *Analysis of Health Surveys*.
- Langholz B, Borgan O. Estimation of absolute risk from nested case-control data. *Biometrics*. 1997; 53(2):767–77,499. [PubMed: 9192463]
- Langholz B, Jiao J. Computational methods for case-cohort studies. *Comput Stat Data Anal*. 2007; 51(8):3737–3748.
- Lin D. On fitting Cox's proportional hazards models to survey data. *Biometrika*. 2000; 87(1):37–47.
- Lin D, Wei L. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc*. 1989;1074–1078.
- Lumley T. survey: analysis of complex survey samples. R package version 3.26. 2011
- Lumley TS. Analysis of complex survey samples. *J Stat Softw*. 2004; 9(1):1–19.
- McDowell A, Engel A, Massey J, Maurer K. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980. *Vital Health Stat Ser 1*. 1981; 15:1–114. [PubMed: 7344293]
- Patterson B, Dayton C, Graubard B. Latent class analysis of complex sample survey data. *J Am Stat Assoc*. 2002; 97(459):721–741.
- Preston, D.; Lubin, JH.; Pierce, D.; McConney, ME. *Epicure user's guide*. Seattle, WA: 1993.
- Rao JNK, Scott AJ. On simple adjustments to chi-square tests with sample survey data. *Ann Stat*. 1987; 15(1):385–397.
- Reid N, Crepeau H. Influence functions for proportional hazards regression. *Biometrika*. 1985; 72(1): 1–9.
- Särndal, CE.; Swensson, B.; Wretman, J. *Springer Series in Statistics*. New York: Springer-Verlag; 1992. *Model Assisted Survey Sampling*.
- Shah, B. *Joint Statistical Meetings. ASA*; 2002. *Calculus of Taylor deviations*.
- Shen Y, Cheng SC. Confidence bands for cumulative incidence curves under the additive risk model. *Biometrics*. 1999; 55(4):1093–1100. [PubMed: 11315053]
- Williams R. Product-limit survival functions with correlated survival times. *Lifetime Data Anal*. 1995; 1(2):171–186. [PubMed: 9385099]
- Woodruff RS. Simple method for approximating variance of a complicated estimate. *J Am Stat Assoc*. 1971; 66(334):411–414.

Table 1

Leading causes of death by gender in NHANES I/II.

Cardiovascular		Cancer		Other	
Cause	No. (%)	Cause	No. (%)	Cause	No. (%)
Men					
Myocardial infarction	647 (35.8)	Lung	295 (32.6)	Chronic l ower respiratory	140 (14.7)
Other ischemic heart disease	575 (31.8)	Prostate	121 (13.4)	Pneumonia	133 (14.0)
Cerebrovascular disease	229 (12.7)	Colon	87 (9.6)	Emphysema	54 (5.7)
Other heart disease	127 (7.0)	Pancreas	44 (4.9)	Diabetes	50 (5.2)
Heart failure	47 (2.6)	Leukemia	42 (4.6)	Liver disease	46 (4.8)
Hypertensive heart disease	45 (2.5)	Esophageal	32 (3.5)	Unspecified disease	46 (4.8)
Aneurysm	37 (2.0)	Stomach	29 (3.2)	Suicide	30 (3.1)
Atherosclerosis	32 (1.8)	Kidney	27 (3.0)	Septicemia	29 (3.0)
Diseases of the pericardium	15(0.8)	Non-Hodgkin's lymphoma	26 (2.9)	Alzheimer's	19 (2.0)
Hypertension and renal disease	15(0.8)	Bladder	25 (2.8)	Nephritis	14 (1.5)
Total	1,805		906		953
Women					
Other ischemic heart disease	444 (29.7)	Breast	131 (18.3)	Diabetes	83 (11.5)
Myocardial infarction	418 (27.9)	Lung	115 (16.1)	Pneumonia	83 (11.5)
Cerebrovascular disease	312 (20.9)	Colon	109 (15.3)	Chronic lower respiratory	54 (7.5)
Other heart disease	111 (7.4)	Ovarian	36 (5.0)	Septicemia	31 (4.3)
Heart failure	54 (3.6)	Pancreatic	35 (4.9)	Liver disease	30 (4.1)
Hypertensive heart disease	35 (2.3)	Non-Hodgkin's lymphoma	25 (3.5)	Unspecified disease	23 (3.2)
Atherosclerosis	28 (1.9)	Leukemia	22 (3.1)	Emphysema	20 (2.8)
Diseases of the pericardium	21 (1.4)	Brain/CNS	21 (2.9)	Nephritis	18 (2.5)
Diseases of arteries	17 (1.1)	Multiple myeloma	21 (2.9)	Kidney infection	16 (2.2)
Aneurysm	15 (1.0)	Stomach	19 (2.7)	Nutritional deficiency	15 (2.1)
Total	1,492		714		723

Table 2

Selected risk factors for gender- and cause-specific weighted Cox regression analyses in NHANES I/II.

Risk factor	Women			Men		
	CVD ^a	Cancer	Other	CVD	Cancer	Other
Race						
White	1.00 (referent)				1.00 (referent)	
Black	1.27 (1.05, 1.55)				1.10 (0.84, 1.43)	
Other	0.96 (0.44, 2.07)				0.35 (0.14, 0.85)	
Annual income						
<10,000				1.42 (1.23, 1.64)	1.29 (1.07, 1.56)	1.99 (1.59, 2.46)
10,000				1.00 (referent)	1.00 (referent)	1.00 (referent)
Marital status						
Divorced or separated	1.28 (0.99, 1.66)				1.56 (1.14, 2.13)	1.70 (1.26, 2.29)
Other	1.00 (referent)			1.00 (referent)	1.00 (referent)	
BMI ^b						
<18	1.00 (referent)		4.34 (3.01, 6.25)	1.00 (referent)	1.91 (1.04, 3.49)	3.21 (2.22, 4.65)
18–30					1.00 (referent)	
>30	1.46 (1.25, 1.71)		1.00 (referent)	1.35 (1.12, 1.62)	1.27 (0.99, 1.63)	1.00 (referent)
Smoking status						
Never	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)
Former	1.28 (1.03, 1.60)	1.06 (0.78, 1.43)	1.34 (1.00, 1.81)	1.09 (0.92, 1.28)	1.15 (0.89, 1.49)	1.21 (0.95, 1.56)
Current	2.08 (1.74, 2.48)	1.87 (1.48, 2.368)	2.34 (1.83, 3.01)	1.90 (1.61, 2.25)	2.37 (1.84, 3.07)	2.11 (1.66, 2.67)
Unknown	2.35 (1.79, 3.08)	1.64 (1.06, 2.53)	2.54 (1.67, 3.87)	4.05 (3.03, 5.40)	2.48 (1.65, 3.73)	3.55 (2.46, 5.12)
Hypertensive (160/90)						
No	1.00 (referent)		1.00 (referent)	1.00 (referent)		
Yes	1.97 (1.66, 2.33)		1.55 (1.18, 2.04)	1.92 (1.60, 2.29)		
Alcohol consumption						
<1 drinks per week	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)	1.00 (referent)
1–7 drinks per week	0.84 (0.68, 1.03)	1.26 (1.00, 1.59)	0.74 (0.56, 0.98)	0.76 (0.65, 0.89)	1.00 (0.82, 1.22)	0.97 (0.79, 1.19)
> 7 drinks per week	0.70 (0.49, 1.01)	0.99 (0.65, 1.52)	0.95 (0.62, 1.45)	0.83 (0.69, 0.99)	1.29 (1.02, 1.63)	1.60 (1.27, 2.01)
Serum cholesterol (mg), Mean (SD)	1.002 (1.000, 1.003)			1.003 (1.002, 1.005)		0.998 (0.996, 1.000)

Risk factor	Women			Men		
	CVD ^a	Cancer	Other	CVD	Cancer	Other
Diabetic						
No	1.00 (referent)		1.00 (referent)	1.00 (referent)		1.00 (referent)
Yes	2.30 (1.81, 2.92)		1.97 (1.40, 2.77)	1.89 (1.47, 2.44)		2.20 (1.59, 3.05)
Unknown	1.26 (1.08, 1.48)		1.06 (0.83, 1.35)	0.89 (0.76, 1.04)		1.20 (0.96, 1.49)

^aCardiovascular disease

^bBody mass index

Table 3

Estimates of ρ and ϕ for the absolute risk of cancer death between the ages of 50 and 80 years for women in NHANES I/II.

Smoking status	BMI	$\pi(\%)$	SE	ρ (CVD%, Other%)	ϕ CVD% Other%
Overall				(94.7, 5.3)	34.3 17.3
Never		8.1	1.5	(57.7, 42.3)	
Former		8.2	2.0	(56.6, 43.4)	
Current	<18	12.3	2.8	(52.2, 47.8)	23.0 22.6
Unknown		10.6	4.1	(50.0, 50.0)	
Never		8.9	1.5	(96.6, 4.4)	
Former		9.2	2.0	(96.1, 3.9)	
Current	18-30	14.8	2.8	(95.3, 4.7)	35.6 16.2
Unknown		13.0	4.2	(94.9, 5.1)	
Never		8.7	1.6	(98.0, 2.0)	
Former		9.0	2.2	(98.2, 1.8)	
Current	>30	14.3	3.4	(97.7, 2.3)	32.7 20.1
Unknown		12.5	4.5	(97.5, 2.5)	

Table 4

Absolute risk of cause-specific death (%) between ages 65–75 years for referent male by smoking status.

Smoking status	Piecewise exponential			Semiparametric		
	$\hat{\pi}(\%)$	Influence	Standard Error	$\hat{\pi}(\%)$	Influence	Standard Error
Cardiovascular						
Never	23.1	5.5	5.6	23.2	5.7	5.4
Former	29.2	7.8	7.6	29.3	7.3	7.2
Current	37.6	9.9	10.0	37.8	9.2	9.3
Unknown	55.8	11.1	11.1	55.7	11.0	11.2
Cancer						
Never	6.8	2.5	2.7	6.9	2.6	2.8
Former	8.0	3.3	2.9	8.2	3.4	3.3
Current	12.5	4.5	4.4	12.7	4.6	4.6
Unknown	8.6	6.4	6.6	8.7	6.7	6.8

Table 5

Expected ten-year cause-specific mortality for NHANES I/II US male population aged 35–75 under different smoking intervention scenarios. ‘Cessation’ considers the change in deaths if current smokers were former smokers. ‘No exposure’ considers the change in deaths if current and former smokers were never smokers.

Cause of death	No intervention	Smoking intervention, N (Prevented deaths)	
		Cessation	No exposure
Cancer	3,088,427	2,610,800 (477,627)	1,868,455 (1,219,972)
Cardiovascular	5,173,242	4,757,214 (416,028)	3,814,235 (1,359,007)
Other	2,785,291	2,548,402 (236,889)	2,109,647 (675,643)
Total	11,046,960	9,916,416 (1,130,544)	7,792,337 (3,254,623)

Table 6

Ten-year projected number of deaths due to cardiovascular disease based on NHANES III (1988–2006).

Race	Age	Sample			Population		
		Obs.	Exp.	Obs./Exp.	Obs.	No. (SE)	Obs./Exp.
Women							
White	30–64	25	29.6	0.8	479,373	575,403 (7296)	0.8
	65–69	24	20.7	1.1	409,120	374,209 (6664)	1.1
	70–75	64	73.5	0.9	649,266	779,090 (10761)	0.8
Black	30–64	42	32.0	1.3	159,468	127,665 (1955)	1.2
	65–69	26	20.7	1.3	74,490	71,539 (1749)	1.0
	70–75	27	28.9	0.9	97,390	101,891 (2672)	1.0
Other	30–64	23	22.4	1.0	118,353	86599 (4536)	1.4
	65–69	20	18.8	1.1	28,876	44,375 (4833)	0.7
	70–75	23	22.1	1.0	31,150	53,447 (5519)	0.6
Men							
White	30–64	42	56.2	0.7	750,983	1,134,933 (9581)	0.7
	65–69	39	38.8	1.0	616,549	567,883 (6664)	1.1
	70–75	81	76.4	1.1	827,236	742,220 (8467)	1.1
Black	30–64	61	44.1	1.4	234,389	158,099 (1457)	1.5
	65–69	22	23.8	0.9	52,724	58,917 (932)	0.9
	70–75	31	33.4	0.9	75,094	83,021 (1479)	0.9
Other	30–64	40	49.5	0.8	65,032	186,429 (3115)	0.3
	65–69	21	27.7	0.8	41,398	47,124 (1688)	0.9
	70–75	25	28.3	0.9	50,701	49,534 (2042)	1.0