HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2016 August 07.

Published in final edited form as:

Biometrics. 2016 June ; 72(2): 546–553. doi:10.1111/biom.12435.

Multivariate Piecewise Exponential Survival Modeling

Yan Li[1,*], **Orestis A. Panagiotou**[2], **Amanda Black**[2], **Dandan Liao**[3], and **Sholom Wacholder**[2]

[1]Joint Program in Survey Methodology, University of Maryland at College Park, Maryland 20742, U.S.A.

[2]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, Maryland 20892, U.S.A.

[3]Measurement, Statistics & Evaluation, College of Education, University of Maryland at College Park, Maryland 20742, U.S.A.

Summary

In this article, we develop a piecewise Poisson regression method to analyze survival data from complex sample surveys involving cluster-correlated, differential selection probabilities, and longitudinal responses, to conveniently draw inference on absolute risks in time intervals that are prespecified by investigators. Extensive simulations evaluate the developed methods with extensions to multiple covariates under various complex sample designs, including stratified sampling, sampling with selection probability proportional to a measure of size (PPS), and a multi-stage cluster sampling. We applied our methods to a study of mortality in men diagnosed with prostate cancer in the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial to investigate whether a biomarker available from biospecimens collected near time of diagnosis stratifies subsequent risk of death. Poisson regression coefficients and absolute risks of mortality (and the corresponding 95% confidence intervals) for prespecified age intervals by biomarker levels are estimated. We conclude with a brief discussion of the motivation, methods, and findings of the study.

Keywords

Absolute risks; Complex sampling designs; Marginal prediction; PLCO; Poisson regression; Probability proportional to a measure of size (PPS)

1. Introduction

Lack of standard easy-to-use software has hindered the development and application of case-cohort design (Prentice, 1986), despite practical advantages (Wacholder et al., 1992) and efficiency that are comparable to the standard nested case-control design.

* yli6@umd.edu.

6. Supplementary Materials

The R codes for implementing multivariate piecewise exponential survival modeling are available with this paper at the *Biometrics* website on Wiley Online Library.

Essentially, in a case-cohort design, biomarker levels are measured only in a *subcohort* or a random sample of individuals from a cohort, and in all the cases. The efficiency loss from case-cohort designs is small, but the cost-savings from measuring biomarkers only on the subcohort and on cases can be very large when the cases comprise a small fraction of the cohort. The savings arise from collecting or measuring expensive, individual data for members of the sample instead of the entire cohort.

Because all the covariates are available for cases and a random sample of the entire cohort, case-cohort studies allow estimation of any parameter that can be estimated from the full cohort. One particular advantage for biomarker studies in clinical epidemiology is that absolute risks of disease are easily available, unlike standard Cox proportional hazards modeling. In particular, case-cohort designs allow Poisson regression that provides estimates of the absolute risk with the additional benefit of allowing for multiple complex time variables (age, time since first exposure or randomization, time exposed, etc.) (Wacholder, 1991). Poisson regression is also a reasonable alternative to fitting proportional hazards models for estimates of hazard ratios or risk ratios (Breslow et al., 1983).

Li et al. (2012) developed a piecewise-exponential approach where Poisson regression model parameters are estimated from pseudo-likelihood and the corresponding variances are derived by Taylor linearization methods. The simple piecewise exponential assumption allows efficient computation, even with time-varying exposures. In addition, the estimates of covariances retain the computational efficiency and the flexibility of Poisson regression methods. Methods by Li et al., (2012), however, were developed for the situation when the failure rate for each time interval is modeled only by a single categorical covariate. In this article, we extend their methods to a more typical, but more complex, problem of multiple covariates, both categorical and continuous, and emphasize the modeling of absolute survival rates in time intervals that are specified by the investigators. In addition, extensive simulations evaluate the extensions to multi-covariates under various complex sample designs, including stratified sampling, sampling with selection probability proportional to a measure of size (PPS), and a multi-stage cluster sampling.

This work was motivated by a study of mortality in men diagnosed with prostate cancer in the Prostate, Lung, Colorectal, and Ovarian (PLCO) cancer screening trial. The goal of the study was to evaluate whether a hypothesized biomarker available from biospecimens collected near time of diagnosis stratifies subsequent risk of prostate cancer death. In our sampling plan, all men who died of prostate cancer (cases) are selected with certainty and a subcohort of men diagnosed with prostate cancer is selected with stratified simple random sampling (SSRS) from the intervention arm of PLCO. The proposed piecewise Poisson regression method is applied to evaluate the prognostic value of a biomarker of interest among men diagnosed with prostate cancer. Poisson regression coefficients and absolute risks of mortality (and the corresponding 95% confidence intervals) for each of three prespecified age intervals by biomarker levels are estimated.

In Section 2, we describe the methodology. The performance of the proposed methods is evaluated using simulation studies with various sampling designs in Section 3 and illustrated

through application to the case-cohort data with SSRS from PLCO in Section 4. We conclude with a brief discussion in Section 5.

2. Methods

Let the follow-up time be divided into I disjoint time intervals, $i = 1, 2, \ldots, I$, and $\boldsymbol{x}$ be a p-vector of covariates, including both continuous and/or categorical covariates. Let N be the size of the total cohort. Suppose we have exposure data for a sample size d of observed cases and a subcohort of size n from the entire cohort. Denote t_{im} and t_{ik} the time at risk in the ith interval for the mth and kth subjects from the case sample and from the subcohort, respectively, for $m = 1, \ldots, d$ and $k = 1, \ldots, n$. We write the loglikelihood function as

$$\tilde{L}=\sum_{i=1}^{I}\left[\sum_{m=1}^{d} w_m^{(1)}\left\{\xi_{im} log\left(r_{im}\right)-r_{im}t_{im}\right\}-\sum_{k=1}^{n}\left(1-\delta_k\right)w_k^{(0)}r_{ik}t_{ik}\right],$$

where ξ_{im} denotes a disease indicator that is equal to 1 if the case m experiences the event in the time interval i. δ_k is an indicator variable that is equal to 1 if the kth subject in the subcohort becomes a case and equal to 0 otherwise. $w_m^{(1)}$ and $w_k^{(0)}$ are the sample weights of the mth and kth subjects selected from the case sample and from the subcohort, respectively; these weights are inverse of the selection probability. Absolute risk of failure at time interval i for individual j, $r_{ij} = r_{ij}(\boldsymbol{x}_{ij}, \boldsymbol{\theta}_i)$ with $\boldsymbol{\theta}_i = (\alpha_i, \boldsymbol{\beta})$, is a function of a p-vector of regression covariates $\boldsymbol{x}_{ij}$, α_i the intercept parameter for time interval i, and $\boldsymbol{\beta}$ a p-vector of unknown parameters. We assume r_{ij} is twice differentiable with respect to $\boldsymbol{\theta}_i$. Denote $r_{ij}^{\boldsymbol{\theta}}$ and $r_{ij}^{\boldsymbol{\theta\theta}}$ the first and second derivatives of r_{ij} with respect to the model parameter $\boldsymbol{\theta}$. As a function of the regression covariates, r_{ij} can be quite general and allows either a relative or absolute hazard form. Solve the associated pseudo-score equation

$$\tilde{U}_{\beta}=\sum_{i=1}^{I}\left\{\sum_{m=1}^{d} w_m^{(1)}\frac{\boldsymbol{r}_{im}^{\boldsymbol{\theta}}}{r_{im}}\left(\xi_{im}-r_{im}t_{im}\right)-\sum_{k=1}^{n}\left(1-\delta_k\right)w_k^{(0)}\boldsymbol{r}_{ik}^{\boldsymbol{\theta}}t_{ik}\right\}$$

for $\boldsymbol{\theta}$ and the solution $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$ (refer to Binder (1983) for conditions for consistency). A convenient way of approximating the variance for nonlinear estimators involves calculating the Taylor deviate of the estimator for each observation. Following Shah (2004), we derived the Taylor deviate by differentiating the weighted nonlinear estimator with respect to its weights.

For case m', we have

$$\boldsymbol{z}_{m'}=\frac{\partial\tilde{\boldsymbol{\theta}}}{\partial w_{m'}^{(1)}}=-\left(\tilde{\boldsymbol{I}}_{\boldsymbol{\theta\theta}}\right)^{-1}\sum_{i=1}^{I}\left\{\xi_{im'} log\left(r_{im'}\right)-r_{im'}t_{im'}\right\} \quad (1)$$

and for control k', we have

$$\boldsymbol{z}_{k'}=\frac{\partial\tilde{\boldsymbol{\theta}}}{\partial w_{k'}^{(0)}}=-\left(\tilde{\boldsymbol{I}}_{\boldsymbol{\theta\theta}}\right)^{-1}\sum_{i=1}^{I}\left\{\left(1-\delta_{k'}\right)r_{ik'}t_{ik'}\right\}, \quad (2)$$

where

$$\tilde{\boldsymbol{I}}_{\boldsymbol{\theta\theta}}=\sum_{i=1}^{I}\left[\sum_{m=1}^{d}w_m^{(1)}\left\{\frac{\boldsymbol{r}_{im}^{\boldsymbol{\theta\theta}}r_{im}-\boldsymbol{r}_{im}^{\boldsymbol{\theta}}\boldsymbol{r}_{im}^{\boldsymbol{\theta}}}{(r_{im})^2}\xi_{im}-\boldsymbol{r}_{im}^{\boldsymbol{\theta\theta}}t_{im}\right\}-\sum_{k=1}^{n}\left(1-\delta_k\right)w_k^{(0)}\boldsymbol{r}_{im}^{\boldsymbol{\theta\theta}}t_{ik}\right].$$

Following Li et al. (2012), the variance of $\tilde{\boldsymbol{\theta}}$ can then be approximated by the variance of the weighted sum of

$$\sum_{m=1}^{N}w_m^{(1)}z_m+\sum_{k=1}^{n}w_k^{(0)}z_k. \quad (3)$$

The estimate of the variance–covariance matrix of this weighted sum depends on the sample design for the selection of the cases and the subcohort. Assume cases are selected by simple random sampling, but allow for the subcohort to be selected with a complex design, involving stratification and/or clustering. Suppose the entire cohort is partitioned into primary sampling units (PSUs) with similar PSUs grouped into strata (for $h = 1, \ldots, H$). Within stratum h, I_h PSUs are selected at the first stage and then at the second stage, that is, within selected PSU l (for $l = 1, \ldots, I_h$) in stratum h, n_{hl} individuals are selected such that the subcohort has the total sample size $n=\sum_{h=1}^{H}n_h$ with $n_h=\sum_{l=1}^{I_h}n_{hl}$. We select the subcohort sample with possibly unequal probabilities of selections of the PSUs and/or secondary sampling units. Following Cochran (1977), the variance estimate is

$$\operatorname{var}\left(\tilde{\boldsymbol{\theta}}\right)=\frac{N}{N-1}\sum_{m=1}^{N}\left(z^{1m}-\bar{z}^{1}\left(z^{1m}-\bar{z}^{1}\right)^{T}+\sum_{h=1}^{H}\frac{I_h}{I_h-1}\sum_{l=1}^{I_h}\left(z_0^{(hl)}-\bar{z}_0^{(h)}\right)\left(z_0^{(hl)}-\bar{z}_0^{(h)}\right)^{T}, \quad (4)$$

where $z^{1m}=w_m^{(1)}z_m$, $\bar{z}^{1}=\frac{1}{N}\sum_{m=1}^{N}z^{1m}$, the weighted PSU total $z_0^{(hl)}=\sum_{k=1}^{n_{hl}}w_{hlk}^{(0)}z_k$ and $\bar{z}_0^{(h)}=\frac{1}{I_h}\sum_{l=1}^{I_h}z_0^{(hl)}$. If all cases are sampled and the subcohort is selected by stratified simple random sampling (SSRS), then w's for the sample of cases are equal to one and the weighted PSU totals within strata are just the weighted individuals with $I_h = n_h$ and $z_0^{(hl)}=w_{hlk}^{(0)}z_k$.

Under multistage cluster sampling in large-scale population-based surveys, sampling of PSU's at the first stage is often approximated by sampling with replacement due to small selection probabilities, even though the PSUs are selected without replacement (Rao and Rust, 2009). The I_h PSU totals of weighted Taylor deviates in stratum h, i.e., $z_0^{(hl)}$ for $l = 1, \ldots, I_h$, are then independent random variables, and their variances can be estimated by sample variance, i.e., the average of squared deviation of $z_0^{(hl)}$ from their stratum mean. Thus, the variability from sampling within PSU's is accounted for in this variance estimation (Korn and Graubard, 1999).

3. Simulations

We conducted simulations to evaluate the performance of the proposed estimator and its variance estimator under various sampling designs that reflect typical epidemiologic studies. We assumed that exposed and unexposed individuals enter the cohort at a certain [the same or different age at entry] age and are followed for up to 15 years over three 5-year time intervals. Two sets of disease rates over three time-intervals are specified (λ_1, λ_2, λ_3) = (90, 150, 200) per 100,000 person-years and $(\lambda_1^*, \lambda_2^*, \lambda_3^*) = (1475.7, 2256.1, 3497.7)$ per 100,000 person-years, representing the 2000–2004 annual US mortality rates for all causes at ages 60–64, 65–69, and 70–74, respectively, according the Surveillance, Epidemiology, and End Results (SEER) Program (http://seer.cancer.gov/).

The entire cohort of size $N = 100,000$ is generated with two covariates (x_1 and x_2) and two sets of failure time, where $x_1 \sim binomial(N, \pi)$ with $\pi = \Pr(x_1 = 1) = 0.3$ and $x_2 \sim N(0\ 1)$; one set failure time t is generated from the piecewise exponential distribution with the rate in the ith time interval

$$r_i = \exp\left(\alpha_i + \beta_1 x_1 + \beta_2 x_2\right),$$

where $\alpha_i = \ln(5\lambda_i/100,000)$ for time interval $i = 1, 2, 3$ and $\beta_1 = \beta_2 = \ln(5) = 1.61$; while the other set of failure time s is generated for competing causes with mortality rates of

$$r_i^* = \exp\left(\alpha_i^*\right)$$

with $\alpha_i^* = \ln\left(5\lambda_i^*/100,000\right)$ for $i = 1, 2, 3$. With the two sets of failure time, the disease indicator at each time interval can be inferred by $d_i = 1$ if $t_i < 5$ and $t_i < s_i$; 0 otherwise; and $d_{i'} = 0$ if $d_i = 1$ for any $i < i'$.

For simple illustration purposes, we sample all the cases. For the selection of the subcohort, three one-stage sample designs (simple random sampling [SRS], selection probability proportional to the size [PPS], stratified SRS [SSRS]) are applied. For PPS, we sample a subcohort with selection probability proportional to the size of the function of the failure time, i.e., $\text{size}_j = \frac{1}{\sqrt[4]{t_j}}$, and therefore the inclusion probability $p_j = \frac{n \times \text{size}_j}{\sum_{j=1}^{N} \text{size}_j}$ for $j = 1, \ldots, N$. For SSRS, we form four strata of equal sizes, with strata defined by the value of x_2,

increasing from stratum 1 to 4. The final sample weights for selected units are the inverse of the selection probabilities. Note that SRS is noninformative sampling, while the PPS and SSRS designs depend on the failure time t or the covariate x_2 that predicts the failure time. We are expecting that the analysis without considering the designs of PPS and SSRS will produce biased estimates, whereas the proposed estimators, taking the design features into account, are design-consistent under various sampling designs. To make fair comparison among the three sample designs, we set the selection probability for SRS and PPS to be 0.01 and selection probabilities for SSRS to be 0.016, 0.0064, 0.0128, and 0.0192, respectively, for the four strata so that all the three designs select subcohorts of the same sample size of 1000.

Table 1 shows the relative biases of the model parameter estimates, their empirical variance estimates, and the ratio of linearization to empirical variance estimates from 1000 simulation runs. We observe the following: (i) the parameter estimates are approximately unbiased across SRS, SSRS, and PPS sampling designs with relative biases close to zero; (ii) compared to SRS and PPS, SSRS design is more efficient and produces smallest empirical variance estimates; and (iii) the proposed Taylor linearization variance estimates approximate the true variances well with the ratios of Taylor linearization to empirical variance estimates consistently close to the value of one.

In addition to the one-stage sampling designs described above (SRS, PPS, and SSRS), we also apply two-stage sampling to select the subcohort to study the clustering effect on the performance of the proposed estimators and their variances. To introduce the clustering effect, we sort the population by the value of the risk factor of x_1, and then sequentially group the population into $M = 1000$ clusters of each cluster size of $N_l = 100$. As such, the population intracluster correlation $\rho(x_1) > 0$. At the first stage of sampling, $m = 40$ clusters are randomly selected by PPS sampling, where the size for cluster l, $\overline{\text{size}}_l$ for $l = 1, \ldots, M$, is defined by the cluster mean of the size value for each individual defined before, i.e., $\overline{\text{size}}_l = \frac{1}{N_l}\sum_{k=1}^{N_l} \text{size}_{lk}$, the function of the failure time. At the second stage of sampling, stratified SRS is used to sample individuals within sampled clusters, where four equalsized strata have the value of x_2 increasing from stratum 1 to stratum 4, and the selection probabilities are, respectively, 0.04, 0.16, 0.32, 0.48. As a result, the subcohort size is about $n = 1000$.

We include a variance estimator (denoted by *var*), for comparison purposes, that considers differential weighting effect but ignores the clustering effect, which is computed as the sum of the variance of weighted Taylor deviates in cases and the variance of weighted Taylor deviates in the sampled subcohort. Recall that the proposed linearization variance estimators account for both the differential weighting and clustering effects.

Table 2 shows the simulation results across the two multistage sample designs, with and without intracluster correlation of the risk factor x_1. We observe that the variance estimates that ignore the clustering effect considerably underestimate the true (empirical) variances when population intracluster correlation $\rho(x_1) > 0$ with the variance ratios smaller than 1 (ranging from 0.312 to 0.768), while the proposed linearization variance estimators are

consistently close to empirical estimates whenever the population intracluster correlation $\rho(x_1) = 0$ or $\rho(x_1) > 0$.

The robustness of the proposed methods with various disease rates at the three time intervals with different population size ($N = 5000$) was also evaluated; the results showed a similar pattern (Table 3).

4. Application to PLCO Data

The PLCO trial, sponsored by the National Cancer Institute, is a multicenter randomized trial aimed at evaluating the effectiveness of prostate, lung, colorectal, and ovarian cancer screening modalities on cancer-specific mortality. The design and rationale are described in detail elsewhere (Prorok et al., 2000). In brief, 154,952 subjects (49.5% men and 50.5% women) aged 55–74 years were enrolled at 10 screening centers between September 1993 and July 2001. 77,469 subjects (men $n = 38, 340$) randomized to the intervention arm underwent regular tests, including a prostate-specific antigen (PSA) test and digital rectal exam at baseline and annually thereafter for 3 years, followed by an additional 2 years of screening with PSA alone. Men with abnormal screening test results were referred to their personal physicians for a diagnostic evaluation. Cancer incidence and deaths were ascertained primarily through the Annual Study Update form, supplemented by periodic linkage to the National Death Index. All prostate cancer diagnoses and deaths were verified by ascertainment of medical records and death certificates. Information on the diagnosis of prostate cancer, including date of diagnosis, stage, and grade (Gleason score), were abstracted by trained medical record specialists. Participants were followed-up for cancer incidence and death through December 31st, 2009.

4.1. *Aim*

The aim of the current study, chosen to illustrate the application and performance of the piecewise Poisson method described in this article, was to evaluate the prognostic value for mortality risk of insulin-like growth factor (IGF)-1 among men diagnosed with prostate cancer.

4.2. *Inclusion/Exclusion Criteria*

Blood specimens were only available from men randomized to the intervention (screening) arm. As such, men from the control arm were ineligible for this study. Men were also considered ineligible if they were missing important covariates (PSA, Gleason score, or clinical stage), had no prediagnostic serum, or if their closest blood draw to diagnosis exceeded 7 years.

4.3. *Case Definition*

Cases were men in the intervention arm who were diagnosed with prostate cancer and subsequently died of prostate cancer. A total of 111 deaths due to prostate cancer were validated and selected with certainty, giving weights equal to one.

4.4. *Subcohort Sampling*

After exclusions, 3259 men in the intervention arm who were diagnosed with prostate cancer remained eligible for inclusion. Measuring the biomarkers of interest in all of these men would have been cost-prohibitive, immensely time-consuming, and could result in further depletion of a precious specimen resource. Therefore, from these eligible men we selected a subcohort using stratified random sampling (SSRS) to evaluate the prognostic value of IGF-1 on prostate cancer mortality. Before sampling, we stratified the entire cohort by whether a case was *prevalent*, (i.e., diagnosed within 12 months of first screen), or *incident* (i.e., diagnosed more than 12 months after first screen), and by four categories of risk of prostate cancer mortality (low risk: PSA < 10 ng/ml and Gleason 2–6 and clinical stage T1/T2a; intermediate-risk: clinical stage T2b/T2c or PSA 10–20 ng/ml or Gleason score 7; high-risk: PSA > 20 ng/ml or Gleason score 8–10, or clinical stage T3a; very high: T3b-T4 or metastatic: Any T, N1, or Any T, Any N, M1) defined by the 2012 National Comprehensive Cancer Network guidelines for the risk stratification of men diagnosed with prostate cancer. Within these eight strata, we randomly selected ~20% of the men within low-, intermediate- and high-risk strata and oversampled the very-high-risk strata with respective selection probabilities 69 and 92% for prevalent and incident cases. Overall, a subcohort of 594 men diagnosed with prostate cancer was selected for the analysis. The weight for each subcohort control participant is the inverse of the selection probability.

4.5. *Covariates and Follow-Up Time*

Body weight and height, race, history of diabetes, and other comorbidities were collected at baseline on a self-administered lifestyle questionnaire, while the IGF-1 was measured at the time of subcohort selection using blood samples closest to the time of prostate cancer diagnosis.

Follow-up time for the men in the subcohort was determined from the date of prostate cancer diagnosis to the exit date of prostate cancer death, death due to other causes, or the end of follow-up (December 31, 2009), whichever came first. Survival time was created by computing the follow-up time in each of the three time intervals (<75, 75 to <80, ≥80 years of age). The time between age of diagnosis and the end of the follow-up was up to 13 years (with median follow-up time 7 years).

4.6. *Real Data Application*

We used a modified Cox regression method (Binder, 1983) suitable for analysis of complex survey data to estimate hazard ratios (HR). Absolute risks (AR) of failure in the *i*th time interval were estimated using the developed piecewise-Poisson regression with failure rate defined by

$$r_i = \exp\left(\alpha_i + \beta_1 logIGF1 + \beta_2 BMI + \beta_3 \text{smoke} + \beta_4 \text{fam} + \beta_5 \text{diabetes} + \beta_6 \text{Damico}\right),$$

where α_i is the intercept corresponding to the *i*th age interval; logIGF1 is log-transformed levels of IGF-1 (in ng/mL); BMI is the body mass index (<25, 25–30, ≥30 in kg/m^2); smoke is the baseline smoking status (never, current, former smoker); fam is the family history (no

or yes); refers to whether the subject was diagnosed with diabetes or not, and Damico is the risk of prostate cancer mortality (low, medium, high, or very high/metastatic) according to the D'Amico risk classification.

The results for this example are reported in Table 4. As expected, the risk of mortality for cases diagnosed as prostate cancer tends to increase as age increases. Cox regression and piecewise-Poisson regression analyses show the similar results in terms of the statistical significance of the estimated regression coefficients, except that a marginal increase in mortality risk was found for BMI ≥ 30 kg/m^2 compared with BMI < 25 kg/m^2 (with P-value = 0.081) using the piecewise-Poisson regression method. The risk of prostate cancer mortality is not associated with logIGF1 levels (P-value = 0.614). With the estimated Poisson regression model coefficients $\hat{\boldsymbol{\theta}}=\left(\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\beta}_1,\ldots,\hat{\beta}_6\right)^T$, the marginal prediction ($\boldsymbol{P_M}$) of absolute risks for the three age intervals of 55–74, 75–79, and 80–85 years of old can be estimated by (Korn and Graubard, 1999):

$$\boldsymbol{P}_M=\frac{\sum_{k=1}^{n} w_k \exp\left(\boldsymbol{x_k}\hat{\boldsymbol{\theta}}\right)}{\sum_{k=1}^{n} w_k,} \qquad (5)$$

where $\boldsymbol{x_k}$ is a 3×3 diagonal matrix, concatenated with a 3-row matrix of each row the same covariate vector $\boldsymbol{x}$ for the kth observation. The variance of $\boldsymbol{P_M}$ are estimated by first deriving the Taylor deviate of $\boldsymbol{P_M}$ for the uth observation ($u = 1, \ldots, n$):

$$\begin{aligned} z_u &= \frac{\partial}{\partial w_u}\boldsymbol{P}_M \\ &= \frac{\left\{\exp\left(\boldsymbol{x}_u\hat{\boldsymbol{\theta}}\right)+\left(\frac{\partial}{\partial w_u}\hat{\boldsymbol{\theta}}\right)\sum_{k=1}^{n} w_k\boldsymbol{x_k}\exp\left(\boldsymbol{x_k}\hat{\boldsymbol{\theta}}\right)\right\}-\boldsymbol{P}_M}{\sum_{k=1}^{n} w_k}, \end{aligned}$$

where $\frac{\partial}{\partial w_u}\hat{\boldsymbol{\theta}}$ is Taylor deviate of $\hat{\boldsymbol{\theta}}$, derived in (1) and (2). Then, the variance of $\boldsymbol{P_M}$, denoted by var($\boldsymbol{P_M}$), can be approximated by the variance of the weighted sum of z_u along the same line as described in (3) and (4) in the method section. Marginal predictive margins with their $95\%\text{CI}=\boldsymbol{P}_M \mp 1.96\sqrt{\text{var}\left(\boldsymbol{P}_M\right)}$ for the three age intervals of 55–74, 75–79, and 80–85 years old are, respectively, 0.46% (0.33–0.59%), 0.69% (0.42–0.96%), and 1.39% (0.65–2.13%). Table 5 further presents marginal predictive margins ($\boldsymbol{P_M}$) and the corresponding 95% CI's by age intervals and biomarker levels. Here the $\boldsymbol{P_M}$'s are estimated using (5) but changing the value of logIGF1 in $\boldsymbol{x_k}$ to be the weighted mean of logIGF1 within the four groups formed by 25th, 50th, and 75th quartiles of logIGF, i.e., 4.31, 4.78, 5.01, and 5.3 (log-ng/mL), respectively. It can be observed that marginal predicted absolute risk P_M increases as age increases, but decreases as the biomarker level logIGF1 increases. The similar pattern that the risk of prostate cancer mortality decreases as the logIGF1 level increases, although not statistically significant, has been previously reported in Cao et al.

(2014). In general, the role of IGF-1 in the etiology and progression of prostate cancer has been controversial and the available literature is likely to suffer from multiple reporting biases (Panagiotou and Ioannidis, 2012).

5. Discussion

The methods in this article overcome an obstacle to the use of an economical case-cohort design. The Poisson regression analysis method described here has good performance for analyzing case-cohort studies. The absolute risk, which is very important for clinical practice and decision-making, is conveniently estimated, as in the data example. Data analysts familiar with the basics of R programming and generalized linear models should be able to use the software available at the *Biometrics* website.

Li et al., (2012) focused on categorical covariates thereby reducing the number of terms in the likelihood down to the number of categories of the covariates. In this article, the focus is on estimating the absolute risk for individuals which is more easily addressed by changing the likelihood from category-based to individual-based. In addition, this methodological change provides more flexibility for incorporating time-dependent covariates and facilitating multi-covariate modeling of both continuous and categorical covariates. In the PLCO data example, we now further estimate predictive margins (PM) of absolute risks (i.e., covariate-adjusted absolute risks), the corresponding variances, and 95% C.I. by age intervals and biomarker levels. The variances of PM are estimated using Taylor linearization methods that account for differential weighting and multistage clustering effects in complex designs.

To the best of our knowledge, no existing software is available to analyze data from complex sample surveys involving cluster-correlated, differential selection probabilities, and longitudinal responses, to conveniently draw inference on absolute risks in time intervals that are prespecified by investigators. For example, the **loglink** procedure in the survey software SUDAAN fits log-linear regression models to count data and provides estimates of exponentiated linear functions of the regression coefficients. Expected value of the response is related to the covariates, however, by assuming common absolute risks over time with the OFFSET variable included for the correction of unequal time interval lengths. By contrast, the developed method estimates absolute risks that vary among prespecified time intervals.

The developed methods can be applied to case-cohort studies with general complex sampling designs. As shown with simulations, the proposed methods consistently produce approximately unbiased estimates under various commonly used sample designs, such as stratified simple random sampling, proportion proportional to sizes (defined by outcome and/or risk factors), or even stratified multistage cluster sampling. Moreover, the proposed methods are robust to the random effects induced from the number of subjects that are originally selected in the subcohort from the entire cohort at the baseline but become cases during the follow up.

In analysis of the PLCO data on mortality from prostate cancer, results in terms of the estimates, corresponding standard errors, and statistical significance of the regression coefficients were similar for piecewise Poisson regression and Cox proportional hazards

regression analyses. This result is consistent with our expectation. Essentially, piecewise Poisson regression for modeling survival data assumes a constant absolute risk of failure within the prespecified time intervals for each individual. That is, over the same time interval, individual i has a constant absolute risk of failure, although different from the absolute risks of failure in other time intervals. On the other hand, Cox proportional hazard regression uses a single baseline hazard, which is a function of the follow-up time metric, and therefore the absolute risk changes over the follow-up time metric, and of course within time intervals. If the assumption holds for piecewise Poisson regression, i.e., absolute risk of failure within time intervals is approximately constant, we would expect both methods to produce similar results in estimating relative risks (approximated by relative hazards for Cox regression). Under the situation in which each cell of the cross-classification of person-time and events includes a single event, piecewise Poisson regression method is essentially equivalent to the Cox proportional hazards regression (Loomie et al., 2005). In addition, the proposed piecewise Poisson method provides convenient estimation of absolute risks of failure as well as potentially greater computational efficiency, which can be useful in applications to case-cohort studies.

The case-cohort design with Poisson regression may overcome the reluctance of some molecular epidemiologists to apply the economical practice of using the same subcohort measurements with several case series. Of course, when different case series are studied at different times, using preexisting samples may be untenable when strong batch effects are suspected. For well-standardized assays, especially in studies with limited amounts of biospecimens available, routine estimates of absolute risk measures, including positive and (complement of) negative predictive values, are available using the case-cohort design. In addition, the full panoply of case-cohort analysis, such as flexibility regarding different time scales (age, time since specimen collection, time after diagnosis of prostate cancer, etc.) is available (Preston, 2005) as well as reduce costs about biomarker measurement. Now that we have established that the Poisson regression approach works well for case-cohort data as well as for full-cohort studies (Breslow et al., 1983), the economy of the case-cohort design is restricted only by possibility of differential error, when assays in different batches or with sampling designs are different in the subcohort and case biospecimens.

References

Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. Journal of the American Statistical Association. 1983; 78:1–12.

Binder DA. On variances of asymptotically normal estimators from complex surveys. International Statistics Review. 1983; 51:279–292.

Cao Y, Lindström S, Schumacher F, Stevens VL, Albanes D, et al. Insulin-like growth factor pathway genetic polymorphisms, circulating IGF1 and IGFBP3, and prostate cancer survival. Journal of National Cancer Institute. 2014; 106:dju085.

Korn, EL.; Graubard, BI. Analysis of Health Surveys. John Wiley & Sons, Inc; Hoboken, NJ: 1999.

Li Y, Gail MH, Preston DL, Graubard BI, Lubin JH. Piecewise exponential survival times and analysis of case-cohort data. Statistics in Medicine. 2012; 31:1361–1368. [PubMed: 22415661]

Loomis D, Richardson DB, Elliott L. Poisson regression analysis of ungrouped data. Occupational and Environmental Medicine. 2005; 62:325–329. [PubMed: 15837854]

Panagiotou OA, Ioannidis JP. Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. Journal of Clinical Epidemiology. 2012; 65:740–7. [PubMed: 22537426]

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11.

Preston DL. Poisson Regression in Epidemiology. Encyclopedia of Biostatistics. 2005:6.

Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. Controlled Clinical Trials. 2000; 21:273S–309S. [PubMed: 11189684]

Rust KF, Rao JN. Variance estimation for complex surveys using replication techniques. Statistical Methods in Medical Research. 1996; 5:283–310. [PubMed: 8931197]

Shah BV. Comment on 'linearization variance estimators for survey data' by Demnati A and Rao JNK. Survey Methodology. 2004; 30:29. Available at http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=12-001-X&chropg=1&lang=eng.

Surveillance, Epidemiology and End Results (SEER). Program (www.seer.cancer.gov) SEER*Stat Database: Mortality- All COD, Total U.S. (1969–2007) Katrina/Rita Population Adjustment - Linked To County Attributes – Total U.S., 1969–2007 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released May 2010. Underlying mortality data provided by NCHS (www.cdc.gov/nchs). Date of access: March 29, 2011.

Wacholder S. Practical considerations in choosing between the case-cohort and nested case-control designs. Epidemiology. 1991; 2:155–158. [PubMed: 1932316]

Table 1

Simulation results from piecewise Poisson regression analysis under various one-stage sample designs

	α_1	α_2	α_3	β_1	β_2
True parameter values	−5.404	−4.893	−4.605	1.609	1.609
	Simple random sampling: SRS				
Relative bias ($\times 10^2$)	0.860	0.196	−1.370	3.071	1.534
Empirical variance estimates ($\times 10^3$	2.808	2.639	6.957	6.278	1.438
Ratio of linearization to empirical variance estimates	0.892	0.933	0.930	0.931	0.957
	Probability proportional to size of failure time: PPS(t)				
Relative bias ($\times 10^2$)	1.785	1.839	1.646	6.816	2.008
Empirical variance estimates ($\times 10^3$)	3.431	3.312	9.645	7.913	1.797
Ratio of linearization to empirical variance estimates	0.938	0.976	0.920	0.928	0.933
	Stratified SRS with strata defined by risk factors: SSRS(x_2)				
Relative bias ($\times 10^2$)	−0.060	−0.003	−0.414	−1.543	0.896
Empirical variance estimates ($\times 10^3$)	2.403	2.228	5.815	4.940	1.173
Ratio of linearization to empirical variance estimates	0.971	1.009	1.019	1.015	1.087

Table 2

Simulation results from piecewise Poisson regression analysis under various two-stage sample designs

	α_1	α_2	α_3	β_1	β_2
True parameter values	−5.404	−4.893	−4.605	1.609	1.609
	PPS(t)+SSRS(x_2) with intracluster correlation $\rho(x_1) = 0$				
Relative bias ($\times 10^2$)	0.078	−0.017	−0.131	0.269	0.134
Empirical variance estimates ($\times 10^3$)	2.494	2.490	5.795	5.266	1.121
Ratio of *var*[a] to empirical variance estimates	0.937	0.916	1.029	0.953	1.149
Ratio of linearization to empirical variance estimates	0.949	0.961	1.055	0.960	1.028
	PPS(t)+SSRS(x_2) with intracluster correlation $\rho(x_1) > 0$				
Relative bias ($\times 10^2$)	−0.066	−0.159	−0.291	0.552	−0.235
Empirical variance estimates ($\times 10^3$)	5.212	4.493	7.776	36.986	1.661
Ratio of *var*[a] to empirical variance estimates	0.443	0.509	0.766	0.132	0.768
Ratio of linearization to empirical variance estimates	1.063	1.065	1.038	1.053	1.028

[a] *var* considers differential weighting effect but ignores the clustering effect, computed as the sum of the variance of weighted Taylor deviates in cases and the variance in sampled subcohort.

Table 3

Simulation results from piecewise Poisson regression analysis under various mortality rates at the three time intervals

	α_1	α_2	α_3	β_1	β_2
	$(\lambda_1, \lambda_2, \lambda_3) = (90, 150, 200)$				
True parameter values	−5.991	−5.298	−4.893	1.609	1.609
Relative bias ($\times 10^2$)	0.234	0.151	0.041	0.559	0.373
Empirical variance estimates ($\times 10^2$)	1.640	1.370	2.287	1.750	0.517
Ratio of linearization to empirical variance estimates	1.016	0.998	0.976	1.009	1.074
	$(\lambda_1, \lambda_2, \lambda_3) = (50, 150, 200)$				
True parameter values	−5.404	−4.893	−4.605	1.609	1.609
Relative bias ($\times 10^2$)	0.111	0.061	−0.087	0.497	0.124
Empirical variance estimates ($\times 10^2$)	1.154	1.045	2.192	1.522	0.414
Ratio of linearization to empirical variance estimates	0.970	0.961	0.927	0.932	1.080
	$(\lambda_1, \lambda_2, \lambda_3) = (20, 50, 100)$				
True parameter values	−6.908	−5.991	−5.298	1.609	1.609
Relative bias ($\times 10^2$)	0.261	0.217	0.094	0.559	0.559
Empirical variance estimates ($\times 10^2$)	3.124	2.250	2.813	2.667	0.763
Ratio of linearization to empirical variance estimates	0.977	1.040	0.988	0.958	1.048
	$(\lambda_1, \lambda_2, \lambda_3) = (20, 40, 60)$				
True parameter values	−6.908	−6.215	−5.809	1.609	1.609
Relative bias ($\times 10^2$)	0.362	0.209	0.121	0.497	0.870
Empirical variance estimates ($\times 10^2$)	3.602	2.750	3.625	3.014	0.901
Ratio of linearization to empirical variance estimates	0.928	0.991	0.986	0.951	0.986
	$(\lambda_1, \lambda_2, \lambda_3) = (10, 20, 30)$				
True parameter values	−7.601	−6.908	−6.502	1.609	1.609
Relative bias ($\times 10^2$)	0.553	0.391	0.369	0.746	1.243
Empirical variance estimates ($\times 10^2$)	6.027	4.925	5.530	4.462	1.423
Ratio of linearization to empirical variance estimates	0.951	0.965	1.024	0.971	0.933

Table 4

Parameter estimates and their standard errors using Taylor linearization methods

	Piecewise-Poisson regression			Cox regression		
	Estimate	**Standard error**	**P-value**	**Estimate**	**Standard error**	**P-value**
Age (years)						
α_1 [55–75)	−6.318	1.627	<0.001			
α_2 [75–80)	−5.914	1.598	<0.001			
α_3 [80–85]	−5.219	1.636	0.001			
LogIGF1	−0.160	0.317	0.614	−0.235	0.320	0.464
BMI (kg/m^2)						
Overweight (25–30)	0.225	0.286	0.431	0.167	0.285	0.557
Obese (≥30)	0.571	0.327	0.081	0.439	0.319	0.169
Smoking status						
Current smokers	−0.096	0.417	0.817	−0.186	0.418	0.657
Former smokers	0.013	0.242	0.957	−0.008	0.240	0.972
Family history	−0.261	0.409	0.524	−0.368	0.466	0.407
Diabetes	−0.381	0.450	0.397	−0.386	0.424	0.385
D'Amico score						
Medium	0.858	0.361	0.017	0.927	0.349	0.008
High	2.191	0.327	0.000	2.242	0.313	0.000
Very high/metastatic	4.262	0.366	0.000	4.346	0.326	0.000

Table 5

Predictive margins of absolute risks P_M by age intervals and biomarker levels based on the piecewise Poisson regression analysis given in Table 4

	LogIGF1 (in log-ng/mL)			
	4.31	**4.78**	**5.01**	**5.30**
<75 years	0.5% (0.27–0.73%)	0.47% (0.34–0.6%)	0.45% (0.32–0.58%)	0.43% (0.26–0.6%)
75–80 years	0.76% (0.4–1.12%)	0.7% (0.43–0.97%)	0.68% (0.39–0.97%)	0.64% (0.3–0.98%)
>80 years	1.52% (0.55–2.49%)	1.41% (0.65–2.17%)	1.35% (0.61–2.09%)	1.29% (0.51–2.07%)