Improving the Efficiency of Relative-Risk Estimation in Case-Cohort Studies

Michal KULICH and D. Y. LIN

The case-cohort design is a common means of reducing the cost of covariate measurements in large failure-time studies. Under this design, complete covariate data are collected only on the cases (i.e., the subjects whose failure times are uncensored) and on a subcohort randomly selected from the whole cohort. In many applications, certain covariates are readily measured on all cohort members, and surrogate measurements of the expensive covariates also may be available. The existing relative-risk estimators for the case-cohort design disregard the covariate data collected outside the case-cohort sample and thus incur loss of efficiency. To make better use of the available data, we develop a class of weighted estimators with general time-varying weights that are related to a class of estimators proposed by Robins, Rotnitzky, and Zhao. The estimators are shown to be consistent and asymptotically normal under appropriate conditions. We identify the estimator within this class that maximizes efficiency, numerical studies demonstrate that the efficiency gains of the proposed estimator over the existing ones can be substantial in realistic settings. We also study the estimation of the cumulative hazard function. An illustration with data taken from Wilms' tumor studies is provided.

KEY WORDS: Case-control study; Measurement error; Missing data; Proportional hazards; Survival data; Two-phase design.

1. INTRODUCTION

Clinical and epidemiological cohort studies are routinely conducted to assess the effects of possibly time-dependent covariates on a failure time. For large studies, the assembly of the covariate histories on all cohort members can be prohibitively expensive. The cost can be substantially reduced by using the so-called "case-cohort" design (Prentice 1986). Under this design, the covariate histories are ascertained only for the cases (i.e., the subjects who experience the event of interest during the follow-up period) and for a relatively small subcohort that is a random sample from the original cohort. The case-cohort design has been applied in cancer research (e.g., Mark et al. 2000; Zeegers, Goldbohm, and van den Brandt 2001), heart disease research (Folsom, Aleksic, Catellier, Juneja, and Wu 2002), and HIV research (Nokta et al. 2002). This design has played an increasingly important role in genetic studies (e.g., Ensrud et al. 1999; Rasmussen et al. 2001) because it avoids the high cost associated with genotyping a large number of subjects.

The case-cohort design is a form of two-phase sampling. At the first phase, the study cohort is randomly sampled from a general population; at the second phase, the subcohort is randomly selected from the study cohort. In many applications, certain variables (e.g., treatment assignment, age, gender, and error-prone versions of the expensive true covariates) are observed on all of the subjects in the cohort. Such data are referred to as the *first-phase covariate data*. At the second phase of the case-cohort sampling, complete covariate histories (including all of the expensive covariates not measured at the first phase) are assembled for the cases and the subcohort. These data are known as the *second-phase covariate data*.

The Cox (1972) proportional hazards model is the basis for most methods used to study relative risks in failure-time studies. Most of the existing relative-risk estimators under the case-cohort design are based on modifications of the full-data partial likelihood score function by weighting the contributions from the cases and subcohort members with the inverses of their true or estimated sampling probabilities, ignoring the first-phase covariate data (Prentice 1986; Self and Prentice 1988; Kalbfleisch and Lawless 1988). The only method that utilizes some of the first-phase information is stratification on the first-phase covariates (Borgan, Langholz, Samuelsen, Goldstein, and Pogoda 2000). Estimators with time-varying weights have been proposed as an alternative means of improving the efficiency of the case-cohort estimation (Barlow 1994; Borgan et al. 2000). Simulation studies suggest that the stratified estimator II with time-varying weights proposed by Borgan et al. (2000), referred to as the BII estimator throughout the present article, is the most efficient among the existing estimators. The asymptotic theory for this type of estimators has not been previously established. With the exception of Self and Prentice (1988), no authors studied the estimation of the cumulative hazard function under the case-cohort design.

In this article we seek to improve the efficiency of the relative-risk estimation for case-cohort studies by making fuller use of the available first-phase covariate data. We also aim to fill in the gap in the existing theory of the case-cohort estimation. In the next section we describe the fundamentals of case-cohort sampling and present a unified estimation framework encompassing the existing weighted estimators. In Section 3 we develop a general class of weighted estimators by incorporating arbitrary stochastic processes as time-varying weights into the empirical sampling probabilities, which are in turn used to weight the contributions of the cases and subcohort members to the partial likelihood score function. The resulting doubly weighted (DW) estimators are proven to be consistent and asymptotically normal under appropriate conditions. The efficiency of the DW estimator depends on the choice of the time-varying weights. A by product of Section 3 is establishment of the theoretical properties of the BII estimator and other existing estimators with time-varying weights. In Section 3 we also propose and study a related class of estimators for the cumulative hazard function.

Data arising from a case-cohort study can be viewed as a special type of a missing-data problem. In Section 4 we show

Michal Kulich is Assistant Professor, Department of Probability and Mathematical Statistics, Charles University, Praha, Czech Republic (E-mail: *kulich@karlin.mff.cuni.cz*). D. Y. Lin is Dennis Gillings Distinguished Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *lin@bios.unc.edu*). This work was supported by National Institutes of Health grants 5R01-CA040644-17 and R01-CA82659. The authors thank the National Wilms' Tumor Study Group for providing the data. They are thank the editor, an associate editor, and three referees for their reviews and comments.

Journal of the American Statistical Association
September 2004, Vol. 99, No. 467, Theory and Methods
DOI 10.1198/016214504000000584

that the class of DW estimators is asymptotically equivalent to a class of augmented estimators considered by Robins, Rotnitzky, and Zhao (1994) and specify a time-varying weight that yields an asymptotically efficient estimator within this class. We propose to combine this estimator with the BII estimator through an optimal linear combination as a means of implementing the efficient DW estimator to achieve good numerical properties in finite samples. In Section 5 we report the results of our simulation studies, which demonstrate that the proposed estimator reliably estimates the relative-risk parameter and that its efficiency gains over the existing estimators can be substantial in practical situations. In Section 6 we illustrate the proposed methods with an example, and in Section 7 we give some concluding remarks. Most of the technical details are relegated to the Appendix.

2. A GENERAL FRAMEWORK FOR CASE–COHORT ESTIMATORS

2.1 Model Assumptions and Definition of Case-Cohort Sampling

Let T be the failure time, let C be a potential censoring time, and let $\mathbf{Z}(t)$ be an m-vector of covariate processes. Suppose that T is conditionally independent of C given $\mathbf{Z}(\cdot)$ and that the conditional distribution of T given $\mathbf{Z}(\cdot)$ follows the Cox (1972) proportional hazards model

$$\lambda(t|\mathbf{Z}) = \lambda_0(t)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(t)\},$$

where $\lambda(t|\mathbf{Z})$ is the conditional hazard for failure given the covariate history up to time t, $\boldsymbol{\beta}_0$ is a vector-valued parameter, and $\lambda_0(t)$ is an unspecified baseline hazard function.

Define $X = \min(T, C)$, $\Delta = I(T \le C)$, $N(t) = I(X \le t, \Delta = 1)$, and $Y(t) = I(X \ge t)$. Suppose that the support of C is bounded above by $\tau > 0$ such that $\Pr(Y(\tau) = 1) > 0$. A subject whose failure time is observed (i.e., $\Delta = 1$) is called a case, and a censored subject (i.e., $\Delta = 0$) is referred to as a control.

Consider a cohort of n subjects who can be divided into K mutually exclusive strata based on a discrete random variable V. In practice, V represents some of the first-phase information. We require that V affects the failure time only through the covariates; that is, T is independent of V given $\mathbf{Z}(\cdot)$. Let the selection of a subject into the subcohort be indicated by a binary random variable ξ that is conditionally independent of $(T, C, \mathbf{Z}(\cdot))$ given V. Sampling of subcohort subjects may be done prospectively or retrospectively. The stratum variable V may include any information available at the time of sampling, including failure status and censored failure time. For each $k = 1, \ldots, K$, let $\Pr(\xi = 1|V = k) = \alpha_k$, where $\alpha_k > 0$. We do not require that $\alpha_j \ne \alpha_k$ for all $j \ne k$; in fact, V may define a finer stratification than that used for sampling the subcohort. Let E denote the expectation over the joint distribution of $(T, C, \mathbf{Z}(\cdot), V, \xi)$ and let E_k and var_k denote the expectation and variance within the kth stratum, that is, conditionally on $V = k$.

Throughout this article we assume that an independent realization of the quintuple $(T, C, \mathbf{Z}(\cdot), V, \xi)$ is attached to each subject. The independence structure implies that the subcohort is selected by Bernoulli sampling and the subcohort size is random in each stratum. We let n_k denote the number of subjects in the kth stratum and let $q_k \equiv \Pr(V = k)$ denote the limiting proportion of subjects in the kth stratum. The subjects are indexed by the subscript pairs $\{ki\}$, where k denotes the stratum number and i indexes subjects within strata. For nonstratified sampling (i.e., $K = 1$), we drop the stratum index k and use a single subscript i to identify subjects.

Under the case-cohort design, complete observations $(X_{ki}, \Delta_{ki}, \mathbf{Z}_{ki}(t), 0 \le t \le \tau, V_{ki}, \xi_{ki} \equiv 1)$ are available for all subcohort subjects, and at least $(X_{ki}, \Delta_{ki} \equiv 1, \mathbf{Z}_{ki}(X_{ki}))$ are observed for the cases. Different case-cohort estimators make different assumptions on what additional data are available for the nonsubcohort members. The information that may or may not be observed includes complete covariate histories of the cases, at-risk histories outside the subcohort, inexpensive covariates, and surrogates that do not enter the model but can be used to predict the expensive covariates.

2.2 Principles of Parameter Estimation With Case-Cohort Data

With full data, $\boldsymbol{\beta}_0$ would be estimated by $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$, the root of the partial likelihood (Cox 1972) score function

$$\mathbf{U}_{\mathrm{F}}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{Z}_i(t) - \overline{\mathbf{Z}}_{\mathrm{F}}(t, \boldsymbol{\beta})\}\, dN_i(t), \tag{1}$$

where

$$\overline{\mathbf{Z}}_{\mathrm{F}}(t, \boldsymbol{\beta}) = \mathbf{S}_{\mathrm{F}}^{(1)}(t, \boldsymbol{\beta})/S_{\mathrm{F}}^{(0)}(t, \boldsymbol{\beta}), \tag{2}$$

$$\mathbf{S}_{\mathrm{F}}^{(1)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)\}Y_i(t),$$

and

$$S_{\mathrm{F}}^{(0)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)\}Y_i(t).$$

Only the cases contribute to the summation in (1); the controls affect $\mathbf{U}_{\mathrm{F}}$ only through the at-risk covariate average $\overline{\mathbf{Z}}_{\mathrm{F}}$.

Under the case-cohort design, (1) cannot be calculated, because $\overline{\mathbf{Z}}_{\mathrm{F}}$ involves unobserved data. Virtually all existing case-cohort estimators are based on pseudoscores parallel to (1), with $\overline{\mathbf{Z}}_{\mathrm{F}}$ replaced by an approximation $\overline{\mathbf{Z}}_{\mathrm{C}}$,

$$\mathbf{U}_{\mathrm{C}}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^{\tau}\{\mathbf{Z}_{ki}(t) - \overline{\mathbf{Z}}_{\mathrm{C}}(t, \boldsymbol{\beta})\}\, dN_{ki}(t). \tag{3}$$

We have switched to double indices $\{ki\}$ to reflect the potential stratification. The case-cohort at-risk average is defined as $\overline{\mathbf{Z}}_{\mathrm{C}}(t, \boldsymbol{\beta}) = \mathbf{S}_{\mathrm{C}}^{(1)}(t, \boldsymbol{\beta})/S_{\mathrm{C}}^{(0)}(t, \boldsymbol{\beta})$, where

$$\begin{aligned}\mathbf{S}_{\mathrm{C}}^{(1)}(t, \boldsymbol{\beta}) &= n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\varrho_{ki}(t)\mathbf{Z}_{ki}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t),\\ S_{\mathrm{C}}^{(0)}(t, \boldsymbol{\beta}) &= n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\varrho_{ki}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t).\end{aligned} \tag{4}$$

The potentially time-varying weight $\varrho_{ki}(t)$ eliminates subjects with incomplete data from the estimation by setting $\varrho_{ki} = 0$ whenever $\Delta_{ki} = \xi_{ki} = 0$. The second-phase subjects have positive $\varrho_{ki}(t)$, usually equal to the inverses of their true or estimated sampling probabilities. Various proposals for $\varrho_{ki}(t)$ have been published, yielding different case-cohort estimators.

The existing case-cohort estimators follow two different approaches that differ in the way the cases are treated. The first approach includes the cases in $\overline{\mathbf{Z}}_{\mathrm{C}}$ only at their failure times unless they happen to be in the subcohort. Thus the cases are sampled with probability 1 at the failure time only, but not before, and the subcohort is considered a sample from all study subjects regardless of the failure status. We refer to these estimators as the *N-estimators*. The original estimator proposed by Prentice (1986) belongs to this group; it is obtained by setting $\varrho_i(t) = \xi_i/\alpha$ for $t < T_i$ and $\varrho_i(T_i) = 1/\alpha$. (Only a case can be evaluated at the failure time T_i.) The α's cancel out in the numerator and denominator of $\overline{\mathbf{Z}}_{\mathrm{C}}$; we include them to emphasize the inverse sampling probability interpretation of ϱ_i. Self and Prentice (1988) considered a slightly modified estimator with $\varrho_i(t) = \xi_i/\alpha$ for all t; other authors have suggested using $\varrho_i(T_i) = 1$ rather than $\varrho_i(T_i) = 1/\alpha$.

The general (stratified) N-estimator has weights

$$\varrho_{ki}(t) = \xi_{ki}/\widehat{\alpha}_k(t), \qquad t < T_{ki} \qquad \text{and} \qquad \varrho_{ki}(T_{ki}) = 1, \quad (5)$$

where $\widehat{\alpha}_k(t)$ is a possibly time-varying estimator of α_k. Using an estimated rather than the known true sampling probability can actually improve efficiency (cf. Robins et al. 1994). The simplest way to estimate α_k is to use the empirical proportion of the sampled subjects (Borgan et al. 2000, est. I). A time-varying weight can be obtained by calculating the proportion of the sampled subjects among those who remain at risk at a given time point (Barlow 1994; Borgan et al. 2000, est. I TV).

The second approach is to include the cases in $\overline{\mathbf{Z}}_{\mathrm{C}}$ and weight them by 1 throughout their entire at-risk periods. We call the estimators under this approach the *D-estimators*. This approach treats the cases quite distinctly from the subcohort. Given the failure status, the probability that a case is included in the case-cohort sample is 1; the original sampling probabilities α_k apply to the controls only. The sampling indicators ξ_{ki} are no longer relevant for the cases. Thus, once the failure status is known, one can form a separate stratum consisting of the cases and consider the whole stratum sampled with probability 1, whereas the subcohort is sampled with probabilities $\alpha_1, \ldots, \alpha_K$ from the controls classified into the remaining K strata. This shows that, conditional on failure status, the analysis of the case-cohort design is similar to that of the case-control design whether or not subcohort sampling is done retrospectively.

The first D-estimator, defined by the weights $\varrho_i(t) = \Delta_i + (1 - \Delta_i)\xi_i/\alpha$, was proposed by Kalbfleisch and Lawless (1988) for $K = 1$. A general stratified D-estimator is defined by the weights

$$\varrho_{ki}(t) = \Delta_{ki} + (1 - \Delta_{ki})\xi_{ki}/\widehat{\alpha}_k(t). \quad (6)$$

Again, empirical sampling proportions can be substituted for $\widehat{\alpha}_k(t)$. Because of the separation between the cases and the controls, $\widehat{\alpha}_k(t)$ should be evaluated from the controls only. Examples of constant and time-varying weights of this type have been given by Chen and Lo (1999), Borgan et al. (2000, est. II), and Chen (2001). Specifically, the BII estimator (called estimator II with time-varying weights by Borgan et al.) arises by setting $\widehat{\alpha}_k(t) = \sum_i \xi_{ki}(1 - \Delta_{ki})Y_{ki}(t)/\sum_i (1 - \Delta_{ki})Y_{ki}(t)$, which is the proportion of the sampled controls among those who remain at risk at the time t.

There are two principal differences between the N- and D-estimators. Unlike the D-estimators, the N-estimators generate predictable weights, which facilitates the use of the martingale theory in deriving their theoretical properties. However, this is no longer a critical issue, because other tools that do not require predictability have become available. More importantly, the D-estimators require the retrospective assessment of complete covariate histories for the cases, which improves efficiency but may not be always feasible. In this article we focus on the D-estimators, although we indicate how to adapt our results to the N-estimators.

3. A GENERAL DOUBLY WEIGHTED ESTIMATOR

In this section we generalize the D-estimators defined by (3), (4), and (6). Here the estimated sampling probabilities are weighted by an arbitrary random processes, and each component of $\overline{\mathbf{Z}}_{\mathrm{C}}$ uses a separate probability estimate with a potentially different weighting process. These extensions lead to a class of DW estimators, which includes the BII estimator and other estimators mentioned in the previous section as special cases. We establish the asymptotic properties of DW estimators and study the estimation of the cumulative baseline hazard function.

3.1 Definition of the Doubly Weighted Estimator

Let $\mathbf{A}_{ki}(t)$ be a diagonal matrix with m potentially different random processes on the diagonal. Consider the following estimators of the subcohort sampling probabilities:

$$\widehat{\boldsymbol{\alpha}}_k(t) = \left\{\sum_{i=1}^{n_k}(1 - \Delta_{ki})\mathbf{A}_{ki}(t)\right\}^{-1}\left\{\sum_{i=1}^{n_k}\xi_{ki}(1 - \Delta_{ki})\mathbf{A}_{ki}(t)\right\}.$$

We have m estimators of α_k on the diagonal of $\widehat{\boldsymbol{\alpha}}_k(t)$, which is a matrix. Each estimator can be interpreted as an empirical sampling proportion based on the controls, with the contribution of each control weighted by a component of $\mathbf{A}_{ki}(t)$. We modify (6) slightly to reflect the current matrix structure,

$$\boldsymbol{\varrho}_{ki}(t) = \Delta_{ki}\mathbf{I}_m + (1 - \Delta_{ki})\xi_{ki}\widehat{\boldsymbol{\alpha}}_k^{-1}(t),$$

where $\mathbf{I}_m$ is an $m \times m$ identity matrix. We call $\mathbf{A}_{ki}(t)$ the *second-level weight* to distinguish it from the first-level weight $\boldsymbol{\varrho}_{ki}(t)$.

The at-risk covariate average is estimated by

$$\overline{\mathbf{Z}}_{\mathrm{DW}}(t, \boldsymbol{\beta}) \equiv \left\{\mathbf{S}_{\mathrm{DW}}^{(0)}(t, \boldsymbol{\beta})\right\}^{-1}\mathbf{S}_{\mathrm{DW}}^{(1)}(t, \boldsymbol{\beta}),$$

where

$$\mathbf{S}_{\mathrm{DW}}^{(1)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\boldsymbol{\varrho}_{ki}(t)\mathbf{Z}_{ki}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t)$$

and

$$\mathbf{S}_{\mathrm{DW}}^{(0)}(t, \boldsymbol{\beta}) = n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\boldsymbol{\varrho}_{ki}(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t).$$

Note that $\mathbf{S}_{\mathrm{DW}}^{(1)}$ is an m-vector, whereas $\mathbf{S}_{\mathrm{DW}}^{(0)}$ is a diagonal $m \times m$ matrix. The pseudoscore is defined as

$$\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}) = \sum_{k=1}^{K}\sum_{i=1}^{n_k}\int_0^{\tau}\{\mathbf{Z}_{ki}(t) - \overline{\mathbf{Z}}_{\mathrm{DW}}(t, \boldsymbol{\beta})\}\,dN_{ki}(t), \quad (7)$$

and the DW estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ is the solution to the equation $\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}) = \mathbf{0}$.

The BII estimator can be obtained by placing m copies of $Y_{ki}(t)$ on the diagonal of $\mathbf{A}_{ki}(t)$. In the sequel, the pseudoscore of the BII estimator will be denoted by $\mathbf{U}_{\mathrm{B}}$, the associated at-risk covariate average by $\overline{\mathbf{Z}}_{\mathrm{B}}$, and the estimator itself by $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$. Borgan et al.'s estimator II with fixed weights is also a special case: Simply set $\mathbf{A}_{ki}(t) = \mathbf{I}_m$.

3.2 Properties of the Doubly Weighted Estimator

In addition to the usual regularity conditions for the Cox regression (Andersen and Gill 1982), we assume the following:

Condition 1.

(a) For each component Z_j of $\mathbf{Z}(t)$, $\operatorname{var}\int_0^\tau |dV_j(t)| < \infty$, where $V_j(t) = Z_j(t)\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(t)\}$. For each component A_j of $\mathbf{A}(t)$, $\operatorname{var}\int_0^\tau |dA_j(t)| < \infty$.

(b) The matrix $\mathbf{A}_{ki}(t)$ is independent of ξ_{ki} given stratum k.

(c) The absolute values of the diagonal elements of $\boldsymbol{\mu}_k(t) \equiv E_k(1-\Delta_{ki})\mathbf{A}_{ki}(t)$ are bounded away from 0 for all $t \in [0,\tau]$ and all $k = 1,\ldots,K$.

Thus we require that the total variations of the second-level weights and of certain transformations of covariate processes have finite second moments, that the second-level weight be independent of the subcohort sampling indicator, and that its expectation over the controls in a stratum does not cross 0 in $[0,\tau]$. We relax Condition 1(c) in Section 3.4.

Andersen and Gill (1982) showed that there exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ and a vector of deterministic functions $\overline{\mathbf{z}}(t,\boldsymbol{\beta})$ such that $\overline{\mathbf{Z}}_{\mathrm{F}}(t,\boldsymbol{\beta})$ converges to $\overline{\mathbf{z}}(t,\boldsymbol{\beta})$ in probability uniformly in $t \in [0,\tau]$ and $\boldsymbol{\beta} \in \mathcal{B}$, provided that $\Pr(Y(\tau) > 0) > 0$. We show in Appendix, Section A.2, that under Condition 1, $\overline{\mathbf{Z}}_{\mathrm{DW}}$ converges to the same limit. The following approximation of the pseudoscore at $\boldsymbol{\beta}_0$, proven in Section A.3, is the key to understanding the asymptotic properties of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$.

Theorem 1. Under Condition 1,

$$\begin{aligned}\frac{1}{\sqrt{n}}&\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0)\\ &= \frac{1}{\sqrt{n}}\mathbf{U}_{\mathrm{F}}(\boldsymbol{\beta}_0)\\ &\quad + \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(1-\Delta_{ki})\left(1-\frac{\xi_{ki}}{\alpha_k}\right)\\ &\quad \times \int_0^\tau \left\{\mathbf{R}_{ki}(t) - \boldsymbol{\mu}_k^{-1}(t)\mathbf{A}_{ki}(t)\boldsymbol{\psi}_k(t)\right\} d\Lambda_0(t) + o_P(1),\end{aligned}$$

where $\mathbf{R}_{ki}(t) = \{\mathbf{Z}_{ki}(t) - \overline{\mathbf{z}}(t,\boldsymbol{\beta}_0)\}\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t)$, $\boldsymbol{\psi}_k(t) = E_k(1-\Delta_{ki})\mathbf{R}_{ki}(t)$, and $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$.

Theorem 1 implies that the DW pseudoscore can be approximated by the partial likelihood score plus a sum of iid mean-0 random vectors. Let $\mathcal{I}_{\mathrm{F}}$ be the limiting partial likelihood information matrix, that is,

$$\mathcal{I}_{\mathrm{F}} = \int_0^\tau \left\{\frac{\mathbf{s}^{(2)}(t,\boldsymbol{\beta}_0)}{s^{(0)}(t,\boldsymbol{\beta}_0)} - \overline{\mathbf{z}}^{\otimes 2}(t,\boldsymbol{\beta}_0)\right\} s^{(0)}(t,\boldsymbol{\beta}_0)\,d\Lambda_0(t),$$

where $\mathbf{s}^{(l)}(t,\boldsymbol{\beta}) = E\{\mathbf{Z}(t)^{\otimes l}\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)\}Y(t)\}$ for $l = 0,1,2$, and $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for any vector $\mathbf{a}$.

Theorem 2. Under Condition 1, $n^{-1/2}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathcal{I}_{\mathrm{F}} + \boldsymbol{\Sigma}_{\mathrm{DW}})$ and

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{DW}} - \boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathcal{I}_{\mathrm{F}}^{-1} + \mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{DW}}\mathcal{I}_{\mathrm{F}}^{-1}),$$

where

$$\begin{aligned}\boldsymbol{\Sigma}_{\mathrm{DW}} = &\sum_{k=1}^{K} q_k \frac{1-\alpha_k}{\alpha_k}\\ &\times E_k\Bigg[(1-\Delta_{ki})\\ &\times \int_0^\tau \left\{\mathbf{R}_{ki}(t) - \boldsymbol{\mu}_k^{-1}(t)\mathbf{A}_{ki}(t)\boldsymbol{\psi}_k(t)\right\} d\Lambda_0(t)\Bigg]^{\otimes 2}. \quad (8)\end{aligned}$$

A proof is given in Appendix, Section A.4. The asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ is that of the full-data partial likelihood estimator plus an extra term due to case-cohort sampling. Importantly, the variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ depends on the choice of the second-level weights $\mathbf{A}_{ki}$, but the estimator is consistent no matter how the $\mathbf{A}_{ki}$'s are selected.

Remark 1. The asymptotic distribution of a DW N-estimator with first-level weights $\boldsymbol{\varrho}_{ki}(t) = \xi_{ki}\widehat{\boldsymbol{\alpha}}_k^{-1}(t)$ for $t < T_{ki}$ and $\boldsymbol{\varrho}_{ki}(T_{ki}) = \mathbf{I}_m$ can be obtained from Theorems 1 and 2 by replacing each occurrence of $(1-\Delta_{ki})$ in the extra pseudoscore term, $\boldsymbol{\psi}_k(t)$, $\boldsymbol{\mu}_k(t)$, and $\boldsymbol{\Sigma}_{\mathrm{DW}}$, with 1.

3.3 Estimation of Λ_0 and the Asymptotic Variance of $\hat{\boldsymbol{\beta}}_{\mathrm{DW}}$

We propose estimating the cumulative baseline hazard $\Lambda_0(t)$ by a DW estimator defined as

$$\widehat{\Lambda}_{\mathrm{DW}}(t) = n^{-1}\int_0^t \left\{S_\Lambda^{(0)}(u,\widehat{\boldsymbol{\beta}}_\Lambda)\right\}^{-1}\sum_{k,i} dN_{ki}(u),$$

where $\widehat{\boldsymbol{\beta}}_\Lambda$ is any case-cohort estimator satisfying $\sqrt{n}(\widehat{\boldsymbol{\beta}}_\Lambda - \boldsymbol{\beta}_0) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathcal{I}_{\mathrm{F}}^{-1} + \mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_\Lambda\mathcal{I}_{\mathrm{F}}^{-1})$, $S_\Lambda^{(0)}(t,\boldsymbol{\beta}) = n^{-1} \times \sum_{k,i}\varrho_{ki}^\Lambda(t)\exp\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t)$, and

$$\varrho_{ki}^\Lambda(t) = \Delta_{ki} + (1-\Delta_{ki})\frac{\sum_{k,i}(1-\Delta_{ki})A_{ki}^\Lambda(t)}{\sum_{k,i}\xi_{ki}(1-\Delta_{ki})A_{ki}^\Lambda(t)}.$$

Here the second-level weight $A_{ki}^\Lambda(t)$ is a single random process. The following theorem establishes the weak convergence of $\widehat{\Lambda}_{\mathrm{DW}}$.

Theorem 3. Suppose that Condition 1 holds for $A_{ki}^\Lambda(t)$. Then $\sqrt{n}(\widehat{\Lambda}_{\mathrm{DW}}(t) - \Lambda_0(t))$ converges weakly to a mean-0 Gaussian process.

The theorem is proven and the covariance function of the limiting process is given in Appendix, Section A.5.

The limiting variance $\mathcal{I}_{\mathrm{F}}$ of the partial likelihood score can be consistently estimated by

$$\widehat{\mathcal{I}}_{\mathrm{F}} \equiv n^{-1}\int_0^\tau \left\{\frac{\mathbf{S}_{\mathrm{B}}^{(2)}(t,\widehat{\boldsymbol{\beta}}_{\mathrm{DW}})}{S_{\mathrm{B}}^{(0)}(t,\widehat{\boldsymbol{\beta}}_{\mathrm{DW}})} - \overline{\mathbf{Z}}_{\mathrm{B}}^{\otimes 2}(t,\widehat{\boldsymbol{\beta}}_{\mathrm{DW}})\right\}\sum_{k,i} dN_{ki}(t),$$

where $\mathbf{S}_{\mathrm{B}}^{(2)}$, $S_{\mathrm{B}}^{(0)}$, and $\overline{\mathbf{Z}}_{\mathrm{B}}$ are estimators of $\mathbf{s}^{(2)}$, $s^{(0)}$, and $\overline{\mathbf{z}}$, based on the BII estimator. The extra pseudoscore variance, $\boldsymbol{\Sigma}_{\mathrm{DW}}$, can

be consistently estimated by replacing all of the unknown quantities by their empirical counterparts. The estimator takes the form

$$\widehat{\boldsymbol{\Sigma}}_{\mathrm{DW}} = \frac{1}{n}\sum_{k=1}^{K}\frac{n_k(n_k-\widetilde{n}_k)}{\widetilde{n}_k^2} \times \sum_{i=1}^{n_k}\xi_{ki}(1-\Delta_{ki}) \times \left[\int_0^\tau \left\{\widehat{\mathbf{R}}_{ki}(t) - \widehat{\boldsymbol{\mu}}_k^{-1}(t)\mathbf{A}_{ki}(t)\widehat{\boldsymbol{\psi}}_k(t)\right\} d\widehat{\Lambda}_{\mathrm{DW}}(t)\right]^{\otimes 2},$$

where $\widetilde{n}_k = \sum_{i=1}^{n_k}\xi_{ki}$, $\widehat{\boldsymbol{\psi}}_k(t) = \widetilde{n}_k^{-1}\sum_{i=1}^{n_k}\xi_{ki}(1-\Delta_{ki})\widehat{\mathbf{R}}_{ki}(t)$, $\widehat{\boldsymbol{\mu}}_k(t) = n_k^{-1}\sum_{i=1}^{n_k}(1-\Delta_{ki})\mathbf{A}_{ki}(t)$, and $\widehat{\mathbf{R}}_{ki}(t) = \{\mathbf{Z}_{ki}(t) - \overline{\mathbf{Z}}_{\mathrm{DW}}(t,\widehat{\boldsymbol{\beta}}_{\mathrm{DW}})\}\exp\{\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}^{\mathrm{T}}\mathbf{Z}_{ki}(t)\}Y_{ki}(t)$.

3.4 Generalization to Arbitrary Weights

The theory presented in Section 3.2 requires Condition 1(c), which rules out sign changes in the means of the second-level weights. This condition is not guaranteed to hold for a general $\mathbf{A}_{ki}(t)$. We relax it by dynamic stratification on the sign of $\mathbf{A}_{ki}(t)$. Let $\boldsymbol{\gamma}_{ki}^+(t) = I(\mathbf{A}_{ki}(t) > 0)$, $\boldsymbol{\gamma}_{ki}^-(t) = I(\mathbf{A}_{ki}(t) < 0)$, $\mathbf{A}_{ki}^+(t) = \boldsymbol{\gamma}_{ki}^+(t)\mathbf{A}_{ki}(t)$, and $\mathbf{A}_{ki}^-(t) = -\boldsymbol{\gamma}_{ki}^-(t)\mathbf{A}_{ki}(t)$. Assume that for every $t \in [0,\tau]$, the second-level weights are nonzero with a positive probability. Incorporate the contributions of the positive parts of the weights into $\widehat{\boldsymbol{\alpha}}_k^+(t) = \{\sum(1-\Delta_{ki})\mathbf{A}_{ki}^+(t)\}^{-1} \times \{\sum\xi_{ki}(1-\Delta_{ki})\mathbf{A}_{ki}^+(t)\}$ and the contributions of the negative parts into $\widehat{\boldsymbol{\alpha}}_k^-(t)$, defined analogously. Set

$$\boldsymbol{\varrho}_{ki}(t) = \Delta_{ki}\mathbf{I}_m + (1-\Delta_{ki})\xi_{ki} \times\left[\boldsymbol{\gamma}_{ki}^+(t)\{\widehat{\boldsymbol{\alpha}}_k^+(t)\}^{-1} + \boldsymbol{\gamma}_{ki}^-(t)\{\widehat{\boldsymbol{\alpha}}_k^-(t)\}^{-1}\right]$$

and calculate the DW estimator in the ususal way. For each t, the diagonal matrices of indicators $\boldsymbol{\gamma}_{ki}^+$ and $\boldsymbol{\gamma}_{ki}^-$ classify the observations according to the sign of each component of $\mathbf{A}_{ki}$ into an upper or a lower stratum. An observation may contribute to the upper stratum for some covariates and to the lower stratum for others. Condition 1(c) is guaranteed to hold for $\mathbf{A}_{ki}^+$ and $\mathbf{A}_{ki}^-$ and Theorems 1 and 2 apply in a slightly modified form. The integral $\int(\mathbf{R}_{ki} - \boldsymbol{\mu}_k^{-1}\mathbf{A}_{ki}\boldsymbol{\psi}_k)\,d\Lambda_0$ in Theorem 1 is replaced by

$$\int_0^\tau \left[\mathbf{R}_{ki}(t) - \boldsymbol{\gamma}_{ki}^+(t)\{\boldsymbol{\mu}_k^+(t)\}^{-1}\mathbf{A}_{ki}^+(t)\boldsymbol{\psi}_k^+(t) - \boldsymbol{\gamma}_{ki}^-(t)\{\boldsymbol{\mu}_k^-(t)\}^{-1}\mathbf{A}_{ki}^-(t)\boldsymbol{\psi}_k^-(t)\right] d\Lambda_0(t), \quad (9)$$

where $\boldsymbol{\mu}_k^+(t)$ is a diagonal matrix of $E_k\{(1-\Delta_{ki})A_{kij}(t)|A_{kij}(t) > 0\} > 0$, $\boldsymbol{\psi}_k^+(t)$ is a diagonal matrix of $E_k\{(1-\Delta_{ki})R_{kij}(t)|A_{kij}(t) > 0\}$, and $\boldsymbol{\mu}_k^-$ and $\boldsymbol{\psi}_k^-$ are defined analogously. The index $j = 1,\ldots,m$ runs over the components of the covariate vector. The asymptotic distribution of the dynamically stratified DW estimator is as shown in Theorem 2, with the integral in (8) replaced by (9).

4. IMPLEMENTATION OF THE DOUBLY WEIGHTED ESTIMATOR

In Sections 3.1 and 3.4 we introduced a class of DW estimators with arbitrary second-level weights. In this section we use the theory presented in Section 3.2 and the results of Robins et al. (1994) to identify the efficient estimator within the DW class. We propose to combine this estimator with the BII estimator in an adaptive manner so as to implement the efficient DW estimator in practice. To simplify the notation, we drop the argument t from $\mathbf{Z}(t)$. All of the results can be easily extended to time-varying covariates.

4.1 The Efficient Doubly Weighted Estimator

The dynamically stratified DW estimators of Section 3.4 are consistent and asymptotically normal under mild conditions. However, their efficiency depends on the choice of $\mathbf{A}_{ki}$, which needs to be evaluated on every subject in the original cohort. These weights can incorporate all information observed during the first phase, that is, the stratum variable V, the failure indicator Δ, the censoring time C, the observed components of $\mathbf{Z}$, and surrogates for the unobserved components of $\mathbf{Z}$. We denote all of the variables observed during the first phase by $\mathbf{W}$.

Robins et al. (1994), hereafter referred to as RRZ, considered the general problem of regression models with missing covariates. They introduced a class of estimators with estimating equations $\sum\mathbf{D}_i(\boldsymbol{\beta},\mathbf{h}) = \mathbf{0}$, where $\mathbf{D}_i$ are iid terms of the form

$$\frac{\eta}{\pi}\boldsymbol{\phi} - \frac{\eta-\pi}{\pi}\mathbf{h}, \quad (10)$$

η is a binary indicator of fully observed covariates, $\pi = \Pr(\eta = 1)$, $\boldsymbol{\phi}$ is a mean-0 estimating function evaluable on subjects with complete data, and $\mathbf{h}$ is an arbitrary function of observed data. RRZ showed that for a given estimation function $\boldsymbol{\phi}$, the optimal choice of $\mathbf{h}$ is $\mathbf{h} = E(\boldsymbol{\phi}|\text{observed data})$. An estimator that achieves the semiparametric efficiency bound can be obtained by projecting $\boldsymbol{\phi}$ onto the orthogonal complement of the tangent space for all nuisance parameters (Bickel, Klaassen, Ritov, and Wellner 1993). In cases where the projection is difficult to calculate, such as the case-cohort design, RRZ recommended taking the usual full-data score function as $\boldsymbol{\phi}$ and augmenting the estimator by the optimal $\mathbf{h}$. This yields the estimator that is efficient within the class (10) restricted to the full-data $\boldsymbol{\phi}$. We call this the efficient augmented estimator.

Translating the foregoing theory in our case-cohort notation with no stratification and all indices dropped, we get $\boldsymbol{\phi} = \int\{\mathbf{Z} - \overline{\mathbf{z}}\}\,dN - \int\mathbf{R}\,d\Lambda_0$, $\eta = 1-(1-\Delta)(1-\xi)$, and $\pi = \Delta + (1-\Delta)\alpha$. Write $\mathbf{h} = \int\mathbf{Q}\,d\Lambda_0$, where $\mathbf{Q}$ is an arbitrary process based on observed data. Then (10) can be written as

$$\int\{\mathbf{Z}-\overline{\mathbf{z}}\}\,dN - \rho\int\mathbf{R}\,d\Lambda_0 - (1-\rho)\int\mathbf{Q}\,d\Lambda_0, \quad (11)$$

where $\rho = \Delta + (1-\Delta)\xi/\alpha$. In contrast, the DW estimators of Section 3.1 have the expansion

$$\int\{\mathbf{Z}-\overline{\mathbf{z}}\}\,dN - \rho\int\mathbf{R}\,d\Lambda_0 - (1-\rho)\int\mathbf{A}\boldsymbol{\mu}^{-1}\boldsymbol{\psi}\,d\Lambda_0. \quad (12)$$

Unless $\boldsymbol{\psi} = \mathbf{0}$ on a subset of $[0,\tau]$ with a positive Lebesgue measure, every estimator satisfying (11) for some $\mathbf{Q}$ also satisfies (12) for some $\mathbf{A}$ and vice versa. Thus the class of DW estimators is essentially the same as the class of augmented RRZ estimators. For stratified sampling, the foregoing arguments can be applied within each stratum to yield the same conclusion.

The efficient DW estimator is obtained by setting $\mathbf{A}_{ki}(t) = \mathrm{diag}[E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}]$. With this weight, $\boldsymbol{\mu}_k(t) = E_k(1-\Delta_{ki}) \times \mathbf{A}_{ki}(t) = \mathrm{diag}\{E_k(1-\Delta_{ki})\mathbf{R}_{ki}(t)\} = \mathrm{diag}\{\boldsymbol{\psi}_k(t)\}$, and hence $\boldsymbol{\mu}_k^{-1}$

and $\boldsymbol{\psi}_k$ cancel out each other. The extra-score variance of the efficient DW estimator is $\boldsymbol{\Sigma}_{\mathrm{DW}} = \sum_{k=1}^{K} q_k \frac{1-\alpha_k}{\alpha_k} E_k\{(1-\Delta_{ki}) \int_0^\tau [\mathbf{R}_{ki}(t) - E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}]\, d\Lambda_0(t)\}^{\otimes 2}$. The dynamically stratified version of the efficient DW estimator has the same asymptotic variance and avoids the problem with 0 in both $\boldsymbol{\mu}_k$ and $\boldsymbol{\psi}_k$.

The cumulative baseline hazard estimator that minimizes the asymptotic variance of $\sqrt{n}(\widehat{\Lambda}_{\mathrm{DW}}(t) - \Lambda_0(t))$ for every $t \in [0, \tau]$ (over the class of estimators proposed in Sec. 3.3) can be found in the same way. In particular, it follows from the asymptotic expansions in Appendix, Section A.5, that the optimal $\widehat{\Lambda}_{\mathrm{DW}}$ should use $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ as $\widehat{\boldsymbol{\beta}}_{\Lambda}$ and $A_{ki}^{\Lambda}(t) \equiv E_k[\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_{ki}\} Y_{ki}(t)|\mathbf{W}_{ki}]$ as the second-level weight.

As shown earlier, the efficient DW estimator is asymptotically equivalent to the efficient augmented estimator described by RRZ. We recognized the connection of the DW estimators with the RRZ class of estimators only after our results were mostly developed. A direct implementation of the efficient augmented estimator with survival data is not straightforward; one cannot simply substitute an estimate of the efficient $\mathbf{Q} = E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}$ into (11) because of the unknown functions $\overline{\mathbf{z}}$ and Λ_0 involved in (11). If these functions are replaced by consistent estimators (Wang and Chen 2001), a nonnegligible variability is added to the pseudoscore, and efficiency is lost. It is possible, though, that one could obtain the efficient augmented estimator through a representation of (10) as $\eta\hat{\pi}^{-1}\boldsymbol{\phi}$, where $\hat{\pi}$ is an estimate of the subcohort selection probability based on a correctly specified logistic regression model (RRZ, sec. 6.4).

4.2 Calculating the Efficient Doubly Weighted Estimator

The efficient second-level weight $\mathbf{A}_{ki}(t)$ has the elements

$$E_k\big[\mathbf{Z}_{ki}\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_{ki}\}Y_{ki}(t)|\mathbf{W}_{ki}\big] - \overline{\mathbf{z}}(t,\boldsymbol{\beta}_0)E_k\big[\exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_{ki}\}Y_{ki}(t)|\mathbf{W}_{ki}\big]$$

on the diagonal. These elements involve $\boldsymbol{\beta}_0$, $\overline{\mathbf{z}}$, and two unknown conditional expectations. We show in Appendix, Section A.6, that the unknown parameters in $\mathbf{A}_{ki}(t)$ can be replaced by any consistent estimators without affecting the asymptotic variance of the pseudoscore. (This would not be true for estimating parameters in the pseudoscore itself.) Thus we can replace $\boldsymbol{\beta}_0$ by, for example, the BII estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$, and replace $\overline{\mathbf{z}}$ by $\overline{\mathbf{Z}}_{\mathrm{B}}$. The conditional expectations can be estimated from the second-phase sample.

If Y_{ki} and all components of $\mathbf{Z}_{ki}$ but one are observed during the first phase, then we only need to estimate $E_k[Z_l \exp\{\beta_l Z_l\}|\mathbf{W}]$ and $E_k[\exp\{\beta_l Z_l\}|\mathbf{W}]$, where Z_l is the covariate not observed during the first phase. Unless Z_l is binary, it is not obvious how this can be done in general. Instead, we suggest using approximations to the conditional expectations. The simplest method is to specify a model for the mean of Z_l given $\mathbf{W}$ and to approximate the conditional expectations by plugging in the fitted values $\widehat{Z}_l$ for Z_l. In fact, this approximates $\exp(\beta_l Z_l)$ and $Z_l \exp(\beta_l Z_l)$ linearly by the first-order Taylor expansions around $\widehat{Z}_l$. One could also use the second-order expansions, which augment the plug-in approach by incorporating the residual variance of $\widehat{Z}_l$ and allow for heteroscedasticity.

In our experience, using the second-order expansion or even estimating the exact conditional expectations from the true distribution of $\mathbf{Z}$ given $\mathbf{W}$ did not confer a substantial advantage over the plug-in approach. Thus we recommend constructing $\widehat{\mathbf{Z}}$ such that for fully observed covariates, $\widehat{Z}_l = Z_l$, and for second-phase covariates, $\widehat{Z}_l$ is the fitted value from a rich model regressing Z_l on all the first-phase variables. Then we set

$$\mathbf{A}_{ki}(t) \equiv \mathrm{diag}\big[\{\widehat{\mathbf{Z}}_{ki} - \overline{\mathbf{Z}}_{\mathrm{B}}(t, \widehat{\boldsymbol{\beta}}_{\mathrm{B}})\}\exp\{\widehat{\boldsymbol{\beta}}_{\mathrm{B}}^{\mathrm{T}}\widehat{\mathbf{Z}}_{ki}\}Y_{ki}(t)\big] \qquad (13)$$

and implement the DW estimator as suggested in Section 3.4.

4.3 The Combined Doubly Weighted Estimator: Justifications and Properties

Although it has appealing asymptotic properties, the proposed efficient DW estimator may not always perform well in finite sample sizes. First, the asymptotic theory assumes that the number of subcohort controls is large in every stratum throughout $[0, \tau]$. In practice, the subcohort usually thins out toward the end of the study, affecting the performance of the estimator. Second, the proposed DW estimator is efficient only if the model for $\mathbf{Z}$ given $\mathbf{W}$ is correct. A misspecified model may seriously reduce the efficiency of the DW estimator. For these reasons, we propose that the efficient DW estimator be combined with another consistent estimator that does not use the entire first-phase data (e.g., the BII estimator).

Take any diagonal $m \times m$ matrix $\boldsymbol{\Omega}$ and define a combined pseudoscore

$$\mathbf{U}_{\mathrm{CW}}(\boldsymbol{\beta}) = \boldsymbol{\Omega}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}) + (\mathbf{I}_m - \boldsymbol{\Omega})\mathbf{U}_{\mathrm{B}}(\boldsymbol{\beta}).$$

The combined doubly weighted (CDW) estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{CDW}}$, is the solution to $\mathbf{U}_{\mathrm{CW}}(\boldsymbol{\beta}) = \mathbf{0}$. It follows from Theorems 1 and 2 that this estimator is consistent and asymptotically normal. The asymptotic variance of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{CDW}} - \boldsymbol{\beta}_0)$ is $\mathcal{I}_{\mathrm{F}}^{-1} + \mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{CW}}(\boldsymbol{\Omega})\mathcal{I}_{\mathrm{F}}^{-1}$, where

$$\boldsymbol{\Sigma}_{\mathrm{CW}}(\boldsymbol{\Omega}) = \boldsymbol{\Omega}\boldsymbol{\Sigma}_{\mathrm{DW}}\boldsymbol{\Omega}^{\mathrm{T}} + \boldsymbol{\Omega}\boldsymbol{\Sigma}_{\mathrm{DB}}(\mathbf{I}_m - \boldsymbol{\Omega})^{\mathrm{T}} + (\mathbf{I}_m - \boldsymbol{\Omega})\boldsymbol{\Sigma}_{\mathrm{DB}}^{\mathrm{T}}\boldsymbol{\Omega}^{\mathrm{T}} + (\mathbf{I}_m - \boldsymbol{\Omega})\boldsymbol{\Sigma}_{\mathrm{B}}(\mathbf{I}_m - \boldsymbol{\Omega})^{\mathrm{T}},$$

$\boldsymbol{\Sigma}_{\mathrm{B}}$ is the asymptotic variance of $\mathbf{U}_{\mathrm{B}}$, and

$$\boldsymbol{\Sigma}_{\mathrm{DB}} = \sum_{k=1}^{K} q_k \frac{1-\alpha_k}{\alpha_k} \times E_k\Bigg[(1-\Delta_{ki}) \times \int_0^\tau \big\{\mathbf{R}_{ki}(t) - \boldsymbol{\mu}_k^{-1}(t)\mathbf{A}_{ki}(t)\boldsymbol{\psi}_k(t)\big\}\, d\Lambda_0(t) \times \int_0^\tau \left\{\mathbf{R}_{ki}(t) - \frac{Y_{ki}(t)}{E_k\{(1-\Delta_{ki})Y_{ki}(t)\}}\boldsymbol{\psi}_k(t)\right\}^{\mathrm{T}} d\Lambda_0(t)\Bigg].$$

The asymptotic variance can be easily minimized over diagonal $\boldsymbol{\Omega}$ to obtain the optimal $\boldsymbol{\Omega}_0 \equiv \mathrm{diag}(\omega_1, \ldots, \omega_m)$, where $\omega_j = (\sigma_{\mathrm{B}}^{j,j} - \sigma_{\mathrm{DB}}^{j,j})/(\sigma_{\mathrm{B}}^{j,j} + \sigma_{\mathrm{DW}}^{j,j} - 2\sigma_{\mathrm{DB}}^{j,j})$, and $\sigma_X^{j,j}$ is the jth diagonal element of the matrix $\boldsymbol{\Sigma}_X$. The CDW estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{CDW}}$ that we propose uses a combination matrix $\widehat{\boldsymbol{\Omega}}_0$, which is a consistent estimator of $\boldsymbol{\Omega}_0$ obtained by replacing all unknown quantities with their empirical counterparts. By Slutsky's theorem, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{CDW}}$ is $\mathcal{I}_{\mathrm{F}}^{-1} + \mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{CW}}(\boldsymbol{\Omega}_0)\mathcal{I}_{\mathrm{F}}^{-1}$.

The optimal combination matrix $\widehat{\boldsymbol{\Omega}}_0$ ensures that the asymptotic variance of each component of the CDW estimator is no larger than that of the BII estimator alone or the DW estimator alone. Thus the CDW estimator is protected against a deterioration of efficiency below that of the BII estimator due to an incorrect model for $\mathbf{Z}$ given $\mathbf{W}$ and generally improves the performance of the DW estimator in finite samples.

The efficient augmented RRZ estimator and the CDW estimator have the same limiting distributions if the model for the distribution of $\mathbf{Z}$ given $\mathbf{W}$ is correctly specified. If this model is incorrect, then the variance of both estimators will increase, but the relative ordering is unclear. Both estimators have an upper bound for the asymptotic variance: for RRZ, the variance is never larger than that for the estimator that weights by the true sampling probabilities (RRZ, prop. 6.1). For CDW, the variance is never larger than that for the stratified BII estimator (which is more efficient than the estimator using the true sampling probabilities). This suggests that the CDW estimator may be more efficient than the RRZ estimator when the model for missing covariates is badly misspecified.

The CDW estimator can be calculated according to the following algorithm:

1. Obtain the covariate predictions $\widehat{\mathbf{Z}}_{ki}$.
2. Find $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$ and calculate $\overline{\mathbf{Z}}_{\mathrm{B}}$.
3. Evaluate all $\mathbf{A}_{ki}(t)$.
4. Estimate $\boldsymbol{\Sigma}_{\mathrm{DW}}$, $\boldsymbol{\Sigma}_{\mathrm{B}}$, and $\boldsymbol{\Sigma}_{\mathrm{DB}}$ at $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$ and evaluate $\widehat{\boldsymbol{\Omega}}_0$.
5. Iteratively solve $\mathbf{U}_{\mathrm{CW}}(\boldsymbol{\beta}) = \mathbf{0}$ starting at $\widehat{\boldsymbol{\beta}}_{\mathrm{B}}$ based on the $\mathbf{A}_{ki}$'s and $\widehat{\boldsymbol{\Omega}}_0$ calculated at steps 3 and 4.

5. SIMULATION STUDIES

We conducted extensive simulation studies to evaluate the finite-sample properties of the CDW estimator. For comparison, we also evaluated the nonstratified Self–Prentice estimator (SP), the stratified BII estimator, and, as a benchmark, the full-data Cox estimator (F). The DW part of the CDW estimator pertains to a dynamically stratified DW estimator with estimated plug-in weights $\mathbf{A}_{ki}$ given by (13). Once the number of controls at risk in a stratum drops below 5, the weights $\mathbf{A}_{ki}$ in the stratum are kept constant for all the subsequent failure times. This adjustment prevents aberrations in the weights as the risk sets get small.

We report here two sets of studies involving three covariates, a binary Z_1 with $\Pr(Z_1 = 1) = p$, $Z_2 \sim \mathrm{N}(0, .5^2)$, and $\log(Z_3) \sim \mathrm{N}(cz_2, .5^2)$ conditional on $Z_2 = z_2$. The failure times are exponentially distributed, and the censoring times are uniform. The study cohort consists of 3,000 subjects. The subcohort was drawn from the whole cohort regardless of failure status. Stratified sampling was done so that subcohort subjects were about equally distributed between the predefined strata. A total of 1,000 simulation runs were generated for each setting.

First, we set $p = .5$ and $c = .2$, and assumed that Z_1 and Z_3 were observed at the first phase, while Z_2 was observed only at the second phase. A surrogate $\widetilde{Z}_2 \equiv Z_2 + \varepsilon$ was available for every subject, where ε was normal with mean 0, independent of Z_2. The correlation between Z_2 and $\widetilde{Z}_2$ was equal to either .71 or .93. Eight strata were defined based on Z_1 and on the medians of $\widetilde{Z}_2$ and Z_3. The estimated covariate $\widehat{Z}_2$ in (13) was obtained from a linear model regressing Z_2 on Z_1, $\widetilde{Z}_2$, $\log(Z_3)$, and the censoring time C, based on subcohort controls. The generated datasets included on average about 300 cases and 300 subcohort controls.

Table 1. Summary Statistics for the Simulation Studies With a Continuous Surrogate Covariate

	$\mathrm{corr}(Z_2, \widetilde{Z}_2) = .93$					$\mathrm{corr}(Z_2, \widetilde{Z}_2) = .71$				
Method	*EST*	*SE*	*SEE*	*CP*	*RE*	*EST*	*SE*	*SEE*	*CP*	*RE*
$\beta_1 = .3$										
F	.304	.117	.117	.95	1.00	.299	.120	.117	.94	1.00
S–P	.313	.198	.188	.94	.39	.306	.192	.188	.96	.39
BII	.308	.132	.131	.95	.80	.301	.139	.135	.94	.75
CDW	.304	.123	.121	.94	.93	.300	.132	.128	.94	.83
$\beta_2 = 1.2$										
F	1.197	.122	.120	.95	1.00	1.203	.121	.120	.95	1.00
S–P	1.230	.220	.206	.93	.34	1.242	.210	.207	.95	.33
BII	1.220	.194	.176	.93	.46	1.238	.191	.181	.94	.44
CDW	1.187	.148	.135	.92	.78	1.225	.189	.161	.91	.55
$\beta_3 = .2$										
F	.204	.082	.081	.96	1.00	.206	.083	.081	.95	1.00
S–P	.232	.160	.146	.94	.31	.233	.158	.147	.93	.30
BII	.218	.129	.121	.93	.45	.223	.132	.122	.92	.44
CDW	.189	.094	.086	.93	.88	.190	.101	.091	.93	.79

NOTE: EST, SE, SEE, CP, and RE represent the sampling mean of the estimator, sampling standard error of the estimator, sampling mean of the standard error estimator, coverage probability of the 95% Wald-type confidence interval, and the estimated efficiency of the estimator relative to the full-data estimator. The first-phase data consist of censoring time, Z_1, Z_3, and a surrogate, $\widetilde{Z}_2$, for Z_2. The subcohort includes 300 controls.

The results, given in Table 1, show that the CDW estimator can have much higher efficiency than the other case-cohort estimators. The efficiency gain of the CDW estimator depends on whether the covariate is binary or continuous and on the information about the covariate contained in the first phase data. Parameter estimates for fully observed continuous covariates show impressive efficiency gains. Estimates for fully observed binary covariates show smaller efficiency gains compared with BII, but have the highest overall efficiency. The efficiency gain of CDW for incompletely observed continuous covariates ranges from substantial to negligible, depending on the quality of the surrogate information.

In this set of studies, all case-cohort estimators exhibit a small bias, and standard errors tend to be underestimated. The bias arises with all case-cohort estimators when the subcohort proportion is lower than .1 and increases as the proportion tends to 0.

In the second set of studies, we let $c = .6$ and $\mathrm{logit}(p) = -1.5 + .8Z_2 - .50 \log Z_3$. Unconditionally, $\Pr(Z_1 = 1) = .19$. Covariates Z_2 and Z_3 were observed at the first phase, but Z_1 was observed only at the second phase. We generated two surrogates for Z_1, one with sensitivity .9 and specificity .7 and the other with sensitivity and specificity both equal to .6. Eight strata were defined based on a surrogate for Z_1, and the medians of Z_2 and Z_3. The generated datasets included on average about 260 cases and 250 or 500 subcohort controls. The plug-in weights in the CDW estimator used fitted values from a logistic regression model for Z_1 with Z_2, $\log Z_3$, the surrogate for Z_1, and the censoring time as covariates. The logistic model was fitted to data on subcohort controls.

As shown in Table 2, the efficiency gain of CDW over BII was very substantial for both β_2 and β_3, regardless of the sensitivity and specificity of the surrogate for Z_1. In contrast, almost no efficiency was gained for β_1. This is not surprising, because

Table 2. Estimated Efficiencies of BII and CDW Estimators Relative to the Full-Data Partial Likelihood Estimator Under the Model $\lambda(t|\mathbf{Z}) = Z_1 + .8Z_2 - .5Z_3$ With Binary Z_1 and Binary Surrogates W_1 or W_2 for Z_1

		$N_C = 250$		$N_C = 500$	
Parameter	*Method*	W_1	W_2	W_1	W_2
β_1	BII	.42	.35	.61	.54
	CDW	.45	.36	.63	.54
β_2	BII	.50	.48	.68	.66
	CDW	.83	.78	.91	.87
β_3	BII	.58	.55	.75	.73
	CDW	.84	.79	.92	.89

NOTE: N_C is the mean number of subcohort controls. The first-phase data consist of censoring time, Z_2, Z_3, and surrogates W_1 or W_2 for Z_1. (W_1 has sensitivity .9 and specificity .7; W_2 has sensitivity and specificity .6.)

stratification on a surrogate for β_1 already accounts for most of the information on β_1 contained in the first-phase data.

6. WILMS' TUMOR STUDIES

To illustrate the use of the CDW estimator in practice, we analyzed data collected in two randomized studies in Wilms' tumor patients. Wilms' tumor is a rare kidney cancer occurring in young children. Factors that affect survival and relapse include the histological type of the tumor, classified as favorable versus unfavorable, stage (I–IV), age at diagnosis, and tumor diameter. The National Wilms' Tumor Study Group (NWTSG) conducted several randomized studies to test different treatments in Wilms' tumor patients. We used data on 3,915 subjects participating in two of the NWTSG trials (D'Angio et al. 1989; Green et al. 1998) to evaluate the joint effect of histological type and other covariates on relapse-free survival.

In the NWTSG studies, histological type was assessed in two ways. Pathologists at the individual sites analyzed a tumor sample and determined a preliminary "local" histological type. Each sample was then sent to a central facility, where an experienced pathologist reevaluated it. This reevaluation was an expensive and time-consuming process. The central assessment can be considered the "true" histological type, and the local assessment can be considered an imprecise surrogate. About 11% patients had unfavorable central histology. The sensitivity of unfavorable local histology was 74%, and the specificity was 98%. In the NWTSG studies, central histology was evaluated for all patients. If a case-control design had been used, however, the cost of central histology assessments would have been dramatically reduced.

We took advantage of the full data to investigate different methods of parameter estimation under case-cohort sampling. We pretended that true histology was evaluated only on the cases and the subcohort, while all of the other covariates, including local histology, were available for the whole cohort. The case-cohort design was in fact used in recent NWTSG studies, although it was set up differently than it was in our example.

Before simulating the case-cohort design, we built a model for relapse-free survival based on the full data. Because the effect of age at diagnosis was nonmonotone, we included a continuous piecewise-linear age effect in the linear predictor. Thus there were two separate age effects, one for age up to 1 year and one for age 1 year and older. The final model contained eight parameters: one for histology (unfavorable vs. favorable), two for age at diagnosis, one for stage (III–IV vs. I–II), one for tumor diameter (in cm), two for the interaction of histology and age, and one for the interaction of stage and tumor diameter. All the terms included in the model were highly significant.

We also built a logistic model to predict true histology from local histology, stage (IV vs. I–III), age at diagnosis (over 10 years vs. under 10 years), and study (indicating in which of the two studies the subjects were participating). There were six parameters in this model: the intercept, one parameter for each covariate, and one parameter for the interaction of local histology and stage. As expected, the estimated parameter for local histology had the largest absolute value (not considering the intercept). The other parameter estimates suggested that institutional pathologists were more likely to misclassify favorable histology in stage IV patients and unfavorable histology in patients over age 10 years. In practice, of course, both models would be built based on the case-cohort data. However, we focus here on the performance of the estimators rather than on the issues related to model selection under the case-cohort design. The fact that we built the logistic model on the full data rather than on the case-cohort data generated at each simulation is unlikely to have had any effect on the CDW estimator. The predictions of true histology are driven almost entirely by local histology and its interactions, which were so highly significant that they would be included even if the model were built on a small subsample of the data.

The subjects were divided into eight strata according to local histology, stage (stage III–IV vs. I–II), and age (1 year or older vs. younger than 1 year). All 260 control subjects from the five smallest strata were always included in the subcohort; we sampled about 120 control subjects from two larger strata with 400–1,000 control subjects and about 160 subjects from the largest stratum, which included over 1,600 controls. The subcohort was drawn 1,000 times and included on average 662 control subjects. The number of deaths and relapses in the cohort was 669. Overall, about one-third of the study subjects were in the second-phase sample.

The probability of true unfavorable histology given institutional histology and other covariates was estimated from the logistic model and substituted for true unfavorable histology in the second-level weight. The logistic model was fitted using the cases and the subcohort subjects only.

Estimated standard errors, given in Table 3, show that the CDW estimator exhibits large efficiency gains over the BII estimator for the main effects of histology, age, tumor diameter, and stage, and for the diameter–stage interaction. The table also displays square roots of empirical mean squared errors (SMSE) centered at the full-data estimate, which capture both extra variability of the case-cohort estimators and their potential biases. For the completely observed covariates, the SMSEs of the CDW estimator are tiny, which means that it achieves almost full efficiency. This is not the case for the BII estimator. The efficiency gains achieved by the CDW estimator for the age–histology interaction parameters are more modest. Both are slightly more biased than is the case with the BII estimator. However, the SMSE for the interaction with age under 1 year still favors the CDW estimator over the BII estimator. The other interaction is estimated with nearly equal standard errors and SMSEs by the two estimators. Overall, however, the CDW estimator works

Table 3. Analysis of the Wilms' Tumor Data

	Full data		BII			CDW		
Parameter	Estimate	SE	Estimate[a]	SE[b]	SMSE[c]	Estimate[a]	SE[b]	SMSE[c]
UH	4.041	.413	4.043	.452	.187	4.047	.432	.137
UH∗Age1	−2.634	.464	−2.645	.533	.280	−2.651	.512	.242
UH∗Age2	−.058	.034	−.053	.052	.045	−.049	.051	.046
Age1	−.661	.326	−.673	.363	.162	−.659	.331	.044
Age2	.104	.017	.106	.026	.021	.102	.018	.007
Stage	1.346	.244	1.343	.333	.227	1.345	.255	.126
Diam	.069	.014	.070	.020	.015	.070	.015	.007
Stage∗Diam	−.076	.019	−.076	.028	.020	−.076	.020	.011

NOTE: UH, unfavorable true histology; Age1, slope for <1 year; Age2, slope for ≥ 1 year.

[a]Mean estimated parameter over 1,000 simulated subcohorts.

[b]Mean estimated standard error over 1,000 simulated subcohorts.

[c]Square root of the empirical MSE conditional on the full-data estimate.

well for this problem. It achieves high efficiency in most parameters even though true histology is evaluated for just one-third of the whole cohort.

7. DISCUSSION

The proposed CDW estimator is asymptotically efficient within the broad DW class of case-cohort estimators provided that the model for $E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}$ in the second-level weights is correct. In particular, the CDW estimator is more efficient than the estimators proposed by Chen and Lo (1999), Borgan et al. (2000), and Chen (2001) and is asymptotically equivalent to the efficient augmented estimator of Robins et al. (1994). The first-order approximation to $E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}$ that we advocated for practical use entails only a minor efficiency loss.

For completely observed continuous covariates, the efficiency gain of the CDW estimator over other stratified estimators is substantial; for completely observed binary covariates, the gain is smaller but still appreciable. The efficiency for incompletely observed covariates depends on the ability of the first-phase data to predict the true values of the covariate. It is not surprising that the CDW estimator gains more efficiency for continuous covariates. If all fully observed covariates are binary, then the estimator proposed by Chen and Lo (1999) and Borgan et al. (2000, est. II with constant weights) is efficient within the DW class provided that it is stratified on all combinations of the binary covariates. If the censoring time is also available at the first phase, and censoring is independent of the covariates, then the BII estimator is efficient within the DW class. These results follow from Section 4.1.

Although the CDW estimator is efficient within the DW class, it does not reach the semiparametric efficiency bound. Efficient estimation in case-cohort design has been studied by Nan, Emond, and Wellner (2004) and by Nan (unpublished data). The manuscript by Nan implements the fully efficient estimator when all covariates and surrogates are discrete. Nan's estimator requires estimation of the conditional distributions of censoring times given covariates. As a result, it can handle only a small number of discrete covariates.

The idea of adaptively combining the DW and BII pseudoscores achieves two different objectives. First, it ensures that the efficiency of the CDW estimator is never smaller than that of the BII estimator, even if the model for $E_k\{\mathbf{R}_{ki}(t)|\mathbf{W}_{ki}\}$ is seriously misspecified and/or the first-phase data are of poor quality. Second, it is a reliable method for calculating the DW estimator in finite samples. The asymptotic theory for the DW estimator assumes that the number of subcohort controls who are at risk in a given stratum increases to infinity at each failure time. In practice, the risk sets can get very small at the largest failure times, which leads to computational problems and unreliable approximations of the distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ by its limit in law. Note that large sample sizes at time $t = 0$ do not preclude problems with small risk sets at t close to τ.

There are several ways to reduce the susceptibility of the DW estimator to small risk sets in finite samples. One is to artificially censor the data when the risk sets are still sufficiently large. However, this method may severely reduce efficiency. Another possibility is to stabilize the weights once the risk sets get too small. This may help, but it will not resolve the problem entirely. Among the approaches that we tested, the combined estimator with weights stabilized over small risk sets performed best in finite samples. In our simulation studies, the combination matrix put on average more than .95 of the weight on the DW component, as would be expected in cases where the DW estimator is efficient, yet the CDW estimator had smaller bias and better confidence interval coverage that the DW estimator alone. Given the additional advantage of protection against badly misspecified models for missing covariates, we recommend the CDW estimator. Our simulation studies show that it performs well whenever the number of failures is at least 200, the subcohort sampling fraction is at least 5–10%, and each stratum contains at least 30 controls (unless the whole stratum is sampled).

Throughout this article, we have assumed that the subcohort is selected by Bernoulli sampling. In practice, this is often done by simple random sampling of a fixed number of subjects in each stratum. However, the key asymptotic results of Theorems 1 and 2 apply to simple random sampling without any modification. This follows from the fact that the integrals in (8) have mean 0 conditionally on the stratum. The tightness required in the proofs follows from example 3.6.14 of van der Vaart and Wellner (1996). The results can be also extended to the general multiplicative intensity model (Andersen and Gill 1982) with arbitrary counting processes $N(\cdot)$ (including recurrent events) and general at-risk processes $Y(t)$ (including left truncation), as well as to multiple endpoints and/or competing risks. As pointed out by Prentice (1986), an attractive feature of the case-cohort design is that the same subcohort can be used for several different endpoints.

The sampling and stratification schemes were set in Section 2.1 and kept fixed throughout the rest of the article. An important open question is how to choose strata and set sampling probabilities to maximize efficiency.

APPENDIX: PROOFS

To simplify notation, we sometimes drop the time argument of random processes. For the same reason, the proofs presented in Sections A.1–A.3 pertain to a single covariate. Their extensions to multiple covariates are straightforward. The following proposition is used repeatedly.

Proposition A.1. Let $B_i(t)$, $i = 1, \ldots, n$, be independent and identically distributed real-valued random processes on $[0, \tau]$ with $EB_i(t) \equiv \mu_B(t)$, $\operatorname{var} B_i(0) < \infty$, and $\operatorname{var} B_i(\tau) < \infty$. Suppose that almost all paths of $B_i(t)$ have finite variation. Then $n^{-1/2}\sum_i\{B_i(t) - \mu_B(t)\}$ converges weakly in $\ell^\infty[0, \tau]$ to a mean-0 Gaussian process, and $n^{-1}\sum_i B_i(t)$ converges in probability to $\mu_B(t)$ uniformly in t.

Proof. Suppose first that the $B_i(t)$'s have nondecreasing sample paths. Then, by example 2.11.16 of van der Vaart and Wellner (1996), $n^{-1/2}\sum_i\{B_i(t) - \mu_B(t)\}$ converges weakly to a mean-0 Gaussian process. In the general case, almost every path $b(t)$ of $B_i(t)$ can be written as $b^+(t) - b^-(t)$, where b^+ and b^- are nondecreasing in t. Hence $B_i(t) = B_i^+(t) - B_i^-(t)$, where $B_i^+(t)$ and $B_i^-(t)$ meet the conditions of example 2.11.16. This implies that they are jointly tight. The joint finite-dimensional convergence of the normalized $B_i^+(t)$ and $B_i^-(t)$ follows from the multivariate central limit theorem.

A.1 Asymptotic Expansion of Weights

We first investigate the properties of the time-varying sampling probability estimator $\widehat{\alpha}_k(t)$ under Condition 1. For a given k, define

$$\mathbf{B}_n(t) \equiv n_k^{-1}\sum_{i=1}^{n_k}\begin{pmatrix}\xi_{ki}(1-\Delta_{ki})A_{ki}(t)\\ (1-\Delta_{ki})A_{ki}(t)\end{pmatrix}$$

and

$$\boldsymbol{\mu}(t) \equiv E\mathbf{B}_n(t) = \mu_k(t)\begin{pmatrix}\alpha_k\\ 1\end{pmatrix}.$$

Note that we used Condition 1(b) to calculate $\boldsymbol{\mu}(t)$. By Condition 1(a) and Proposition A.1, $\sqrt{n_k}\{\mathbf{B}_n(t) - \boldsymbol{\mu}(t)\}$ converges weakly to a bivariate mean-0 Gaussian process. This implies that $\sup_{t\in[0,\tau]}|\mathbf{B}_n(t) - \boldsymbol{\mu}(t)| \xrightarrow{P} 0$.

We can write $\widehat{\alpha}_k(t)$ and $\widehat{\alpha}_k^{-1}(t)$ as ratios of the two components of $\mathbf{B}_n(t)$, and write α_k and α_k^{-1} as ratios of the two components of $\boldsymbol{\mu}(t)$. The ratios are continuous transformations as long as $\alpha_k > 0$ and $|\mu_k(t)| > \varepsilon$ for some $\varepsilon > 0$ and all $t \in [0, \tau]$ [Condition 1(c)]. Hence, under these conditions, $\widehat{\alpha}_k(t)$ converges to α_k and $\widehat{\alpha}_k^{-1}(t)$ converges to α_k^{-1}, both uniformly over $t \in [0, \tau]$. It then follows from the functional delta method that, uniformly in t,

$$\begin{aligned}&\sqrt{n}\{\widehat{\alpha}_k^{-1}(t) - \alpha_k^{-1}\}\\ &\quad= \{\alpha_k\mu_k(t)\}^{-1}\frac{1}{\sqrt{n_k}}\\ &\qquad\times\sum_{i=1}^{n_k}(1-\xi_{ki}/\alpha_k)(1-\Delta_{ki})A_{ki}(t) + o_P(1).\end{aligned}\tag{A.1}$$

A.2 Convergence of the At-Risk Average Process

We prove that, under Condition 1, $\sup_{t\in[0,\tau],\beta\in\mathcal{B}}|\overline{Z}_{\mathrm{DW}}(t,\beta) - \overline{z}(t,\beta)| \xrightarrow{P} 0$. It suffices to show that $\sup_{t\in[0,\tau],\beta\in\mathcal{B}}|S_{\mathrm{DW}}^{(l)}(t,\beta) - S_{\mathrm{F}}^{(l)}(t,\beta)| \xrightarrow{P} 0$ uniformly in t for $l = 0, 1$. Clearly,

$$\begin{aligned}&S_{\mathrm{DW}}^{(l)}(t,\beta) - S_{\mathrm{F}}^{(l)}(t,\beta)\\ &\quad= n^{-1}\sum_{k,i}(1-\xi_{ki}/\alpha_k)(1-\Delta_{ki})Z_{ki}^l\exp(\beta Z_{ki})Y_{ki}(t)\\ &\qquad- n^{-1}\sum_{k,i}\{\widehat{\alpha}_k^{-1}(t) - \alpha_k^{-1}\}(1-\Delta_{ki})\xi_{ki}Z_{ki}^l\exp(\beta Z_{ki})Y_{ki}(t).\end{aligned}\tag{A.2}$$

Thus

$$\begin{aligned}&\left|S_{\mathrm{DW}}^{(l)}(t,\beta) - S_{\mathrm{F}}^{(l)}(t,\beta)\right|\\ &\le\left|n^{-1}\sum_{k,i}(1-\xi_{ki}/\alpha_k)(1-\Delta_{ki})Z_{ki}^l\exp(\beta Z_{ki})Y_{ki}(t)\right|\\ &\quad+\sum_k|\widehat{\alpha}_k^{-1}(t) - \alpha_k^{-1}|n^{-1}\sum_i(1-\Delta_{ki})\xi_{ki}|Z_{ki}|^l\exp(\beta Z_{ki})Y_{ki}(t).\end{aligned}$$

Both terms on the right side of the inequality converge to 0 in probability uniformly in t and β. Because $ES_{\mathrm{F}}^{(0)}(t,\beta)$ and $ES_{\mathrm{DW}}^{(0)}(t,\beta)$ are bounded away from 0 on $[0, \tau] \times \mathcal{B}$, the proof is completed.

A.3 Proof of Theorem 1

Note that $U_{\mathrm{DW}}(\beta_0) = U_{\mathrm{F}}(\beta_0) + \sum_{k,i}\int_0^\tau(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})\,dN_{ki}$. We decompose $N_{ki}(t)$ as a sum of a martingale $M_{ki}(t)$ and a compensator $\int_0^t Y_{ki}(s)\exp(\beta_0 Z_{ki})\,d\Lambda_0(s)$ to get

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{k,i}\int_0^\tau(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})\,dN_{ki}\\ &\quad=\int_0^\tau(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})\,d\left\{\frac{1}{\sqrt{n}}\sum_{k,i}M_{ki}\right\}\\ &\qquad+\frac{1}{\sqrt{n}}\int_0^\tau(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})\sum_{k,i}Y_{ki}\exp(\beta_0 Z_{ki})\,d\Lambda_0.\end{aligned}$$

By the martingale central limit theorem, $n^{-1/2}\sum M_{ki}(t)$ converges weakly to a mean-0 Gaussian process. The Skorokhod strong embedding theorem and Helly's second theorem imply that the first term on the right side converges to 0 in probability (Kulich and Lin 2000, app. 1). The integrand of the second term can be written as $(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})S_{\mathrm{F}}^{(0)} = (S_{\mathrm{F}}^{(1)} - S_{\mathrm{DW}}^{(1)}) + \overline{Z}_{\mathrm{DW}}(S_{\mathrm{DW}}^{(0)} - S_{\mathrm{F}}^{(0)})$. Because $\overline{Z}_{\mathrm{DW}}$ converges to $\overline{z}$ uniformly, we have

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{k,i}\int_0^\tau(\overline{Z}_{\mathrm{F}} - \overline{Z}_{\mathrm{DW}})\,dN_{ki}\\ &\quad=\sqrt{n}\int_0^\tau\left(S_{\mathrm{F}}^{(1)} - S_{\mathrm{DW}}^{(1)}\right)d\Lambda_0 + \sqrt{n}\int_0^\tau\left(S_{\mathrm{DW}}^{(0)} - S_{\mathrm{F}}^{(0)}\right)\overline{z}\,d\Lambda_0\\ &\qquad+ o_P(1).\end{aligned}$$

By (A.2),

$$\begin{aligned}&\sqrt{n}\int_0^\tau\left(S_{\mathrm{F}}^{(1)} - S_{\mathrm{DW}}^{(1)}\right)d\Lambda_0\\ &\quad=\frac{1}{\sqrt{n}}\sum_{k,i}\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})\int_0^\tau Z_{ki}\exp(\beta Z_{ki})Y_{ki}\,d\Lambda_0\\ &\qquad-\frac{1}{\sqrt{n}}\sum_{k,i}\xi_{ki}(1-\Delta_{ki})\int_0^\tau(\widehat{\alpha}_k^{-1} - \alpha_k^{-1})Z_{ki}\exp(\beta Z_{ki})Y_{ki}\,d\Lambda_0.\end{aligned}$$

By (A.1), the second term on the right side can be approximated by

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{k,i}\xi_{ki}(1-\Delta_{ki})\\
&\qquad\times\int_0^\tau \frac{Z_{ki}\exp(\beta Z_{ki})Y_{ki}}{\alpha_k\mu_k}\\
&\qquad\qquad\times\left\{\frac{1}{n_k}\sum_j\left(1-\frac{\xi_{kj}}{\alpha_k}\right)(1-\Delta_{kj})A_{kj}\right\}d\Lambda_0\\
&=\frac{1}{\sqrt{n}}\sum_{k,j}\left(1-\frac{\xi_{kj}}{\alpha_k}\right)(1-\Delta_{kj})\\
&\qquad\times\int_0^\tau\frac{A_{kj}}{\mu_k}\left\{\frac{1}{n_k}\sum_i\frac{\xi_{ki}}{\alpha_k}(1-\Delta_{ki})Z_{ki}\exp(\beta Z_{ki})Y_{ki}\right\}d\Lambda_0.
\end{aligned}$$

Thus

$$\begin{aligned}
&\sqrt{n}\int_0^\tau\bigl(S_{\mathrm{F}}^{(1)}-S_{\mathrm{DW}}^{(1)}\bigr)\,d\Lambda_0\\
&=\frac{1}{\sqrt{n}}\sum_{k,i}\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})\\
&\qquad\times\int_0^\tau\Biggl[Z_{ki}\exp(\beta Z_{ki})Y_{ki}\\
&\qquad\qquad-\frac{A_{ki}}{\mu_k}\left\{\frac{1}{n_k}\sum_j\frac{\xi_{kj}}{\alpha_k}(1-\Delta_{kj})Z_{kj}\exp(\beta Z_{kj})Y_{kj}\right\}\Biggr]d\Lambda_0\\
&\qquad+o_P(1).
\end{aligned}$$

Likewise,

$$\begin{aligned}
&\sqrt{n}\int_0^\tau\bigl(S_{\mathrm{DW}}^{(0)}-S_{\mathrm{F}}^{(0)}\bigr)\overline{z}\,d\Lambda_0\\
&=-\frac{1}{\sqrt{n}}\sum_{k,i}\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})\\
&\qquad\times\int_0^\tau\Biggl[\overline{z}\exp(\beta Z_{ki})Y_{ki}\\
&\qquad\qquad-\frac{A_{ki}}{\mu_k}\left\{\frac{1}{n_k}\sum_j\frac{\xi_{kj}}{\alpha_k}(1-\Delta_{kj})\overline{z}\exp(\beta Z_{kj})Y_{kj}\right\}\Biggr]d\Lambda_0\\
&\qquad+o_P(1).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{k,i}\int_0^\tau(\overline{Z}_{\mathrm{F}}-\overline{Z}_{\mathrm{DW}})\,dN_{ki}\\
&\quad=\frac{1}{\sqrt{n}}\sum_{k,i}\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})\int_0^\tau\left(R_{ki}-\frac{A_{ki}}{\mu_k}\widetilde{R}_k\right)d\Lambda_0+o_P(1),
\end{aligned}$$

where $R_{ki}(t)$ is defined in Theorem 1 and

$$\widetilde{R}_k(t)=\frac{1}{n_k\alpha_k}\sum_i\xi_{ki}(1-\Delta_{ki})\{Z_{ki}(t)-\overline{z}(t)\}\exp\{\beta Z_{ki}(t)\}Y_{ki}(t).$$

By Proposition A.1, $\widetilde{R}_k(t)$ converges to $\psi_k(t)\equiv E_k(1-\Delta_{ki})R_{ki}(t)$ in probability uniformly in t. This completes the proof.

A.4 Proof of Theorem 2

We first show that $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ is consistent. This follows from the consistency of $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$ and the fact that the DW pseudoscore approximates the partial likelihood score. To be specific, $|n^{-1}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta})-n^{-1}\mathbf{U}_{\mathrm{F}}(\boldsymbol{\beta})|\le\sup_{t\in[0,\tau],\boldsymbol{\beta}\in\mathcal{B}}|\overline{\mathbf{Z}}_{\mathrm{F}}-\overline{\mathbf{Z}}_{\mathrm{DW}}|$, which converges to 0 in probability uniformly in a neighborhood of $\boldsymbol{\beta}_0$. By the Taylor expansions of $\mathbf{U}_{\mathrm{DW}}(\widehat{\boldsymbol{\beta}}_{\mathrm{DW}})$ and $\mathbf{U}_{\mathrm{F}}(\widehat{\boldsymbol{\beta}}_{\mathrm{F}})$ around $\boldsymbol{\beta}_0$, we have $(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}_{\mathrm{F}})n^{-1}\{\mathbf{D}_{\mathrm{DW}}(\boldsymbol{\beta}_{\mathrm{DW}}^*)-\mathbf{D}_{\mathrm{F}}(\boldsymbol{\beta}_{\mathrm{F}}^*)\}=(\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}-\widehat{\boldsymbol{\beta}}_{\mathrm{F}})n^{-1}\mathbf{D}_{\mathrm{DW}}(\boldsymbol{\beta}_{\mathrm{DW}}^*)$, where $\mathbf{D}_X$ is the derivative of $\mathbf{U}_X$ with respect to $\boldsymbol{\beta}$, and $\boldsymbol{\beta}_X^*$ lies on the line segment between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_X$. Because $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$ is consistent and the $\mathbf{D}_X$'s are bounded in probability, the left side converges to 0. Thus $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ converges in probability to the same limit as $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$, that is, $\boldsymbol{\beta}_0$.

By Theorem 1,

$$\begin{aligned}
&n^{-1/2}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0)\\
&\qquad=n^{-1/2}\mathbf{U}_{\mathrm{F}}(\boldsymbol{\beta}_0)+n^{-1/2}\sum_{k,i}(1-\xi_{ki}/\alpha_k)\boldsymbol{\eta}_{ki}+o_P(1),\quad\text{(A.3)}
\end{aligned}$$

where $\boldsymbol{\eta}_{ki}\equiv(1-\Delta_{ki})\int_0^\tau\{\mathbf{R}_{ki}(t)-\boldsymbol{\mu}_k^{-1}(t)\mathbf{A}_{ki}(t)\boldsymbol{\psi}_k(t)\}\,d\Lambda_0(t)$ are independent and identically distributed. Clearly, $n^{-1/2}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0)$ is asymptotically normal with mean 0. Conditioning on everything but ξ_{ki}, we get $\mathrm{var}_k(1-\xi_{ki}/\alpha_k)\boldsymbol{\eta}_{ki}=E_k\{\boldsymbol{\eta}^{\otimes2}\,\mathrm{var}_k(1-\xi_{ki}/\alpha_k)\}=(1-\alpha_k)/\alpha_kE_k\boldsymbol{\eta}^{\otimes2}$. Thus $n^{-1/2}\sum_{k,i}(1-\xi_{ki}/\alpha_k)\boldsymbol{\eta}_{ki}$ converges to $\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma}_{\mathrm{DW}})$, where $\boldsymbol{\Sigma}_{\mathrm{DW}}=\sum_kq_k(1-\alpha_k)/\alpha_kE_k\boldsymbol{\eta}_{ki}^{\otimes2}$. Furthermore, the two terms on the right side of (A.3) are asymptotically independent, because the (k,i)th contributions to the two terms are uncorrelated. This implies that $n^{-1/2}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0)\xrightarrow{D}\mathrm{N}(\mathbf{0},\mathcal{I}_{\mathrm{F}}+\boldsymbol{\Sigma}_{\mathrm{DW}})$. It then follows from the Taylor expansion and the convergence of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$ and $n^{-1}\mathbf{D}_{\mathrm{DW}}(\boldsymbol{\beta}_0)$ that $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}-\boldsymbol{\beta}_0)=\mathcal{I}_{\mathrm{F}}^{-1}n^{-1/2}\mathbf{U}_{\mathrm{DW}}(\boldsymbol{\beta}_0)+o_P(1)$.

A.5 Estimated Cumulative Baseline Hazard

We assume that $\widehat{\boldsymbol{\beta}}_\Lambda$ is a DW estimator [with weight matrix $\mathbf{A}_{ki}(t)$] that satisfies

$$\begin{aligned}
&\sqrt{n}(\widehat{\boldsymbol{\beta}}_\Lambda-\boldsymbol{\beta}_0)\\
&\qquad=\mathcal{I}_{\mathrm{F}}^{-1}\left\{\frac{1}{\sqrt{n}}\mathbf{U}_{\mathrm{F}}(\boldsymbol{\beta}_0)+\frac{1}{\sqrt{n}}\sum_{k,i}(1-\xi_{ki}/\alpha_k)\boldsymbol{\eta}_{ki}\right\}+o_P(1).
\end{aligned}$$

We show that $\sqrt{n}\{\widehat{\Lambda}_{\mathrm{DW}}(t)-\Lambda_0(t)\}$ converges weakly to a mean-0 Gaussian process whose covariance function at (t,s) is

$$\begin{aligned}
&\mathbf{h}^{\mathrm{T}}(t)(\mathcal{I}_{\mathrm{F}}^{-1}+\mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{DW}}\mathcal{I}_{\mathrm{F}}^{-1})\mathbf{h}(s)\\
&\qquad+\int_0^{\min(t,s)}\frac{d\Lambda_0(u)}{s^{(0)}(u,\boldsymbol{\beta}_0)}+C_1(t,s)-C_2(t,s)-C_2(s,t),
\end{aligned}$$

where $\mathbf{h}(t)=\int_0^t\overline{\mathbf{z}}(u)\,d\Lambda_0(u)$,

$$\begin{aligned}
C_1(t,s)&=\sum_{k=1}^Kq_k\frac{1-\alpha_k}{\alpha_k}\\
&\qquad\times E_k\Biggl[(1-\Delta_{ki})\int_0^t\left\{R_{ki}^\Lambda(u)-\frac{A_{ki}^\Lambda(u)}{\mu_k^\Lambda(u)}\psi_k^\Lambda(u)\right\}\frac{d\Lambda_0(u)}{s^{(0)}(u,\boldsymbol{\beta}_0)}\\
&\qquad\times\int_0^s\left\{R_{ki}^\Lambda(v)-\frac{A_{ki}^\Lambda(v)}{\mu_k^\Lambda(v)}\psi_k^\Lambda(v)\right\}\frac{d\Lambda_0(v)}{s^{(0)}(v,\boldsymbol{\beta}_0)}\Biggr],
\end{aligned}$$

$$\begin{aligned}
C_2(t,s)&=\mathbf{h}^{\mathrm{T}}(t)\mathcal{I}_{\mathrm{F}}^{-1}\sum_{k=1}^Kq_k\frac{1-\alpha_k}{\alpha_k}\\
&\qquad\times E_k\Biggl[(1-\Delta_{ki})\\
&\qquad\times\int_0^t\left\{\mathbf{R}_{ki}(u)-\boldsymbol{\mu}_k^{-1}(u)\mathbf{A}_{ki}(u)\boldsymbol{\psi}_k(u)\right\}d\Lambda_0(u)\\
&\qquad\times\int_0^s\left\{R_{ki}^\Lambda(v)-\frac{A_{ki}^\Lambda(v)}{\mu_k^\Lambda(v)}\psi_k^\Lambda(v)\right\}\frac{d\Lambda_0(v)}{s^{(0)}(v,\boldsymbol{\beta}_0)}\Biggr],
\end{aligned}$$

$R_{ki}^{\Lambda}(u) = \exp\{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}_{ki}(u)\}Y_{ki}(u)$, $\psi_k^{\Lambda}(u) = E_k(1-\Delta_{ki})R_{ki}^{\Lambda}(u)$, and $\mu_k^{\Lambda}(u) = E_k(1-\Delta_{ki})A_{ki}^{\Lambda}(u)$.

We make the decomposition

$$\begin{aligned}\sqrt{n}\{\widehat{\Lambda}_{\mathrm{DW}}(t) - \Lambda_0(t)\} &= \sqrt{n}\int_0^t \big[\{nS_{\Lambda}^{(0)}(u,\widehat{\boldsymbol{\beta}}_{\Lambda})\}^{-1} \\ &\qquad - \{nS_{\Lambda}^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1}\big] d\sum_{k,i} N_{ki}(u) \qquad (A.4)\\ &\quad + \int_0^t \{S_{\Lambda}^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1} d\frac{1}{\sqrt{n}}\sum_{k,i} M_{ki}(u) \qquad (A.5)\\ &\quad + \int_0^t \sqrt{n}\,\frac{S_{\mathrm{F}}^{(0)}(u,\boldsymbol{\beta}_0) - S_{\Lambda}^{(0)}(u,\boldsymbol{\beta}_0)}{S_{\Lambda}^{(0)}(u,\boldsymbol{\beta}_0)}\, d\Lambda_0(u). \qquad (A.6)\end{aligned}$$

By the Taylor series expansion,

$$\begin{aligned}&\{nS_{\Lambda}^{(0)}(\widehat{\boldsymbol{\beta}}_{\Lambda})\}^{-1} - \{nS_{\Lambda}^{(0)}(\boldsymbol{\beta}_0)\}^{-1} \\ &\qquad = -n^{-1}\{S_{\Lambda}^{(0)}(\boldsymbol{\beta}^*)\}^{-2}\mathbf{S}_{\Lambda}^{(1)}(\boldsymbol{\beta}^*)^{\mathrm{T}}(\widehat{\boldsymbol{\beta}}_{\Lambda} - \boldsymbol{\beta}_0),\end{aligned}$$

where $\mathbf{S}_{\Lambda}^{(1)}$ has an obvious definition and $\boldsymbol{\beta}^*$ lies on the line segment between $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_{\Lambda}$. By the consistency of $\widehat{\boldsymbol{\beta}}_{\Lambda}$, the continuity of $S_{\Lambda}^{(0)}(\boldsymbol{\beta})$ and $\mathbf{S}_{\Lambda}^{(1)}(\boldsymbol{\beta})$ and their uniform convergence to $s^{(0)}$ and $\mathbf{s}^{(1)}$, and by the martingale decomposition of $N_{ki}(t)$, (A.4) can be written as $-\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\Lambda} - \boldsymbol{\beta}_0)^{\mathrm{T}}\int_0^t \overline{\mathbf{z}}(u,\boldsymbol{\beta}_0)\,d\Lambda_0(u) + o_P(1)$. Hence (A.4) converges weakly to a mean-0 Gaussian process with covariance function $\mathbf{h}(t)^{\mathrm{T}}(\mathcal{I}_{\mathrm{F}}^{-1} + \mathcal{I}_{\mathrm{F}}^{-1}\boldsymbol{\Sigma}_{\mathrm{DW}}\mathcal{I}_{\mathrm{F}}^{-1})\mathbf{h}(s)$ at (t,s).

By the arguments of Section A.3, the integrand in (A.5) can be replaced by its uniform limit $\{s^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1}$. It then follows from the martingale central limit theorem that (A.5) converges weakly to a mean-0 Gaussian process with covariance function $\int_0^{\min(t,s)}\{s^{(0)}(u,\boldsymbol{\beta}_0)\}^{-1}\,d\Lambda_0(u)$ at (t,s).

With $\widehat{\boldsymbol{\beta}}_{\Lambda}$ replaced by $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$ in (A.4), the sum of (A.4) and (A.5) would be an asymptotic decomposition of the full-data Breslow estimator of $\Lambda_0(t)$; in that case, the two terms would be uncorrelated (Andersen and Gill 1982). However, the difference between $\widehat{\boldsymbol{\beta}}_{\Lambda}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{F}}$ can be approximated by $\mathcal{I}_{\mathrm{F}}^{-1}n^{-1/2}\sum_{k,i}(1-\xi_{ki}/\alpha_k)\boldsymbol{\eta}_{ki}$, which is uncorrelated with $M_{ki}(t)$. Thus (A.4) and (A.5) are uncorrelated in our case as well.

Following the arguments of Section A.3, we can show that (A.6) is asymptotically equivalent to $n^{-1/2}\sum_{k,i}(1-\xi_{ki}/\alpha_k)\eta_{ki}^{\Lambda}(t)$, where

$$\eta_{ki}^{\Lambda}(t) = (1-\Delta_{ki})\int_0^t \left\{R_{ki}^{\Lambda}(u) - \frac{A_{ki}^{\Lambda}(u)}{\mu_k^{\Lambda}(u)}\psi_k^{\Lambda}(u)\right\}\frac{d\Lambda_0(u)}{s^{(0)}(u,\boldsymbol{\beta}_0)}.$$

Because almost all paths of $R_{ki}^{\Lambda}(u)$ and $A_{ki}^{\Lambda}(u)$ have finite variations, (A.6) converges weakly to a mean-0 Gaussian process with covariance function $C_1(t,s)$ at (t,s).

Because both (A.4) and (A.6) involve subcohort sampling indicators, they are not uncorrelated. Using the iid representations of these two terms, we can show that the asymptotic covariance of (A.4) evaluated at t and (A.6) evaluated at s is $-C_2(t,s)$. On the other hand, (A.6) is obviously uncorrelated with (A.5). This completes the proof.

A.6 Weights With Estimated Parameters

In Section 4.3 we claimed that substituting estimated parameters in $\mathbf{A}_{ki}(t)$ does not affect the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\mathrm{DW}}$. In this section we outline a proof for $m = 1$.

Suppose that the second-level weights follow the functional form $A_{ki}(t) = a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))$, where $\mathbf{Q}_{ki}(t)$ are iid processes measured on individual subjects, $\boldsymbol{\theta}_0(t)$ is an unknown parameter, and a is a known real function. Specifically, $\mathbf{Q}_{ki}$ consists of the completely observed covariates, predictors for the expensive covariates, and the at-risk process, and $\boldsymbol{\theta}_0(t)$ includes the true $\boldsymbol{\beta}_0$, $\overline{z}(t)$, and the parameters for predicting the expensive covariate given the cheap covariates. Suppose that $\widehat{\boldsymbol{\theta}}(t)$ is a uniformly consistent estimator of $\boldsymbol{\theta}_0(t)$ such that

$$\begin{aligned}\sup_t\big|a\big(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t)\big) - a\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big) \\ -(\widehat{\boldsymbol{\theta}}(t) - \boldsymbol{\theta}_0(t))g\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big)\big| = o_P(1), \qquad (A.7)\end{aligned}$$

where $g(\mathbf{q},\boldsymbol{\zeta}) = \partial a(\mathbf{q},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. We assume that $g(\mathbf{q},\boldsymbol{\zeta})$ is continuous in $\boldsymbol{\zeta}$, that $g(\mathbf{Q}_{ki}(t),\boldsymbol{\zeta})$ has a finite expectation for all t and $\boldsymbol{\zeta}$, and that almost all paths of $g(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))$ have finite variation.

We need to show that $\sqrt{n_k}\{\widehat{\alpha}_k^{-1}(t,\widehat{\boldsymbol{\theta}}(t)) - \widehat{\alpha}_k^{-1}(t,\boldsymbol{\theta}_0(t))\}$ is $o_P(1)$ uniformly in t, where

$$\widehat{\alpha}_k\big(t,\boldsymbol{\theta}(t)\big) = \frac{n_k^{-1}\sum_{i=1}^{n_k}\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}(t))}{n_k^{-1}\sum_{i=1}^{n_k}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}(t))}.$$

We can write

$$\begin{aligned}&\widehat{\alpha}_k^{-1}\big(t,\widehat{\boldsymbol{\theta}}(t)\big) - \widehat{\alpha}_k^{-1}\big(t,\boldsymbol{\theta}_0(t)\big) \\ &\quad = \frac{n_k^{-1}\sum_i(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))}{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))} \\ &\qquad - \frac{n_k^{-1}\sum_i(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))}{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))}.\end{aligned}$$

Using the identity $xu^{-1} - yv^{-1} = v^{-1}\{(x-y) - xu^{-1}(u-v)\}$, we get

$$\begin{aligned}&\widehat{\alpha}_k^{-1}\big(t,\widehat{\boldsymbol{\theta}}(t)\big) - \widehat{\alpha}_k^{-1}\big(t,\boldsymbol{\theta}_0(t)\big) \\ &\quad = \frac{1}{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))} \\ &\qquad \times\left\{n_k^{-1}\sum_i(1-\Delta_{ki})a\big(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t)\big)\right. \\ &\qquad \left. - n_k^{-1}\sum_i(1-\Delta_{ki})a\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big)\right\} \\ &\qquad - \frac{n_k^{-1}\sum_i(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))}{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))} \\ &\qquad \times\frac{1}{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t))} \\ &\qquad \times\left\{n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a\big(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t)\big)\right. \\ &\qquad \left. - n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big)\right\}.\end{aligned}$$

By consistency of $\widehat{\boldsymbol{\theta}}(t)$ and continuity of $a(t,\cdot)$, $n_k^{-1}\sum_i\xi_{ki}(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))$ converges to $\alpha_k\mu_k(t)$ uniformly in t and $n_k^{-1}\times\sum_i(1-\Delta_{ki})a(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t))$ converges to $\mu_k(t)$ uniformly in t. Thus,

$$\begin{aligned}&\sqrt{n_k}\big\{\widehat{\alpha}_k^{-1}\big(t,\widehat{\boldsymbol{\theta}}(t)\big) - \widehat{\alpha}_k^{-1}\big(t,\boldsymbol{\theta}_0(t)\big)\big\} \\ &\quad = \frac{1}{\alpha_k\mu_k(t)}\frac{1}{\sqrt{n_k}} \\ &\qquad \times\sum_{i=1}^{n_k}\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})\big\{a\big(\mathbf{Q}_{ki}(t),\widehat{\boldsymbol{\theta}}(t)\big) - a\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big)\big\} \\ &\qquad + o_P(1). \qquad (A.8)\end{aligned}$$

In view of (A.7), the right side of (A.8) can be approximated by

$$\frac{\widehat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}_0(t)}{\alpha_k\mu_k(t)}\frac{1}{\sqrt{n_k}}\sum_i\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})g\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big).$$

Disregarding constants, this is a product of $\widehat{\boldsymbol{\theta}}(t)-\boldsymbol{\theta}_0(t)$, which is a uniform $o_P(1)$ term, and

$$\frac{1}{\sqrt{n_k}}\sum_i\left(1-\frac{\xi_{ki}}{\alpha_k}\right)(1-\Delta_{ki})g\big(\mathbf{Q}_{ki}(t),\boldsymbol{\theta}_0(t)\big),$$

which is a normalized sum of smooth independent and identically distributed processes with mean 0 and therefore is uniformly bounded in probability. Hence (A.8) converges to 0 in probability uniformly in t. This completes the proof.

[Received August 2002. Revised April 2004.]

REFERENCES

Andersen, P. K., and Gill, R. (1982), "Cox's Regression Model for Counting Processes: A Large-Sample Study," *The Annals of Statistics*, 10, 1100–1120.

Barlow, W. E. (1994), "Robust Variance Estimation for the Case-Cohort Design," *Biometrics*, 50, 1064–1072.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Inference in Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Borgan, Ø., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000), "Exposure Stratified Case-Cohort Designs," *Lifetime Data Analysis*, 6, 39–58.

Chen, K. (2001), "Generalized Case-Cohort Sampling," *Journal of the Royal Statistical Society*, Ser. B, 63, 791–809.

Chen, K., and Lo, S.-H. (1999), "Case-Cohort and Case-Control Analysis With Cox's Model," *Biometrika*, 86, 755–764.

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 34, 187–220.

D'Angio, G. J., Breslow, N., Beckwith, J. B., et al. (1989), "Treatment of Wilms Tumor: Results of the Third National Wilms Tumor Study," *Cancer*, 64, 349–360.

Ensrud, K. E., Stone, K., Cauley, J. A., et al. (1999), "Vitamin D Receptor Gene Polymorphisms and the Risk of Fractures in Older Women. For the Study of Osteoporotic Fractures Research Group," *Journal of Bone and Mineral Research*, 14, 1637–1645.

Folsom, A. R., Aleksic, N., Catellier, D., Juneja, H. S., and Wu, K. K. (2002), "C-Reactive Protein and Incident Coronary Heart Disease in the Atherosclerosis Risk in Communities (ARIC) Study," *American Heart Journal*, 144, 233–238.

Green, D. M., Breslow, N. E., Beckwith, J. B., et al. (1998), "Comparison Between Single-Dose and Divided-Dose Administration of Dactinomycin and Doxorubicin for Patients With Wilms' Tumor: A Report From the National Wilms' Tumor Study Group," *Journal of Clinical Oncology*, 16, 237–245.

Kalbfleisch, J. D., and Lawless, J. F. (1988), "Likelihood Analysis of Multi-State Models for Disease Incidence and Mortality," *Statistics in Medicine*, 7, 149–160.

Kulich, M., and Lin, D. Y. (2000), "Additive Hazard Regression for Case-Cohort Studies," *Biometrika*, 87, 73–87.

Mark, S. D., Qiao, Y. L., Dawsey, S. M., et al. (2000), "Prospective Study of Serum Selenium Levels and Incident Esophageal and Gastric Cancers," *Journal of the National Cancer Institute*, 92, 1753–1763.

Nan, B., Emond, M., and Wellner, J. A. (2004), "Information Bounds for Cox Regression Models With Missing Data," *The Annals of Statistics*, 32, 723–753.

Nokta, M. A., Holland, F., de Gruttola, V., et al. (2002), "Cytomegalovirus (CMV) Polymerase Chain Reaction Profiles in Individuals With Advanced Human Immunodeficiency Virus Infection: Relationship to CMV Disease," *The Journal of Infectious Diseases*, 185, 1717–1722.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika*, 73, 1–11.

Rasmussen, M. L., Folsom, A. R., Catellier, D. J., Tsai, M. Y., Garg, U., and Eckfeldt, J. H. (2001), "A Prospective Study of Coronary Heart Disease and the Hemochromatosis Gene (HFE) C282Y Mutation: The Atherosclerosis Risk in Communities (ARIC) Study," *Atherosclerosis*, 154, 739–746.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

Self, S. G., and Prentice, R. L. (1988), "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies," *The Annals of Statistics*, 16, 64–81.

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

Wang, C. Y., and Chen, H. Y. (2001), "Augmented Inverse Probability Weighted Estimator for Cox Missing Covariate Regression," *Biometrics*, 57, 414–419.

Zeegers, M. P. A., Goldbohm, R. A., and van den Brandt, P. A. (2001), "Are Retinol, Vitamin C, Vitamin E, Folate and Carotenoids Intake Associated With Bladder Cancer Risk? Results From the Netherlands Cohort Study," *British Journal of Cancer*, 85, 977–983.