Lifetime Data Anal (2009) 15:24–40
DOI 10.1007/s10985-008-9095-z

# Weighted analyses for cohort sampling designs

**Robert J. Gray**

Received: 12 January 2008 / Accepted: 24 July 2008 / Published online: 19 August 2008

**Abstract** Weighted analysis methods are considered for cohort sampling designs that allow subsampling of both cases and non-cases, but with cases generally sampled more intensively. The methods fit into the general framework for the analysis of survey sampling designs considered by Lin (Biometrika 87:37–47, 2000). Details are given for applying the general methodology in this setting. In addition to considering proportional hazards regression, methods for evaluating the representativeness of the sample and for estimating event-free probabilities are given. In a small simulation study, the one-sample cumulative hazard estimator and its variance estimator were found to be nearly unbiased, but the true coverage probabilities of confidence intervals computed from these sometimes deviated significantly from the nominal levels. Methods for cross-validation and for bootstrap resampling, which take into account the dependencies in the sample, are also considered.

**Keywords** Subsampling · Horvitz-Thompson estimators · Cumulative hazard estimator · Cross-validation · Case–control study · Case–cohort design

## 1 Introduction

Clinical trials often involve collection of tissue from the patients enrolled in the trials. The tissue can be used to investigate whether biomarkers are useful for predicting clinical outcomes or for predicting benefit from specific treatments. For various reasons, including costs of biomarker evaluation and lack of availability of tissue or consent

R. J. Gray (✉)
Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
44 Binney Street, Boston, MA 02115, USA
e-mail: gray@jimmy.harvard.edu

for use of tissue, the biomarkers are often evaluated on only a subset of the patients on the study.

If the clinical outcomes of interest are event times, and the event rate is low, then it is well-known that a more efficient design is obtained by sampling subjects with observed events (cases) more intensively than cases with censored event outcomes (non-cases). Case-cohort sampling (Prentice 1986), with the cohort consisting of the patients enrolled on the study, is one attractive option for such a biomarker study. However, the classic case-cohort design assumes all cases, plus a random subcohort, are included in the sample. For biomarker studies in clinical trials, it is generally not possible to include all of the cases, due to lack of availability of analyzable samples. Thus an extension of the case-cohort design is needed, where the sampling fractions for both cases and non-cases are $<100\%$. Such designs fit into the general survey sampling framework considered by Lin (2000). An alternative approach would be to regard the cohort as the patients on the study with analyzable tissue available, but this is often a vaguely defined set, since additional tissue might be obtained during the course of the biomarker project, and determining whether the tissue is evaluable often takes considerable effort, and thus may not be done on all tissue samples.

Cohort sampling has traditionally been studied in the context of epidemiologic cohort studies. The issues which limit evaluation of biomarkers to a subset of the cases in the clinical setting would often occur in cohort studies, too, so the methods discussed here will also be useful in these settings, although the terminology of patients enrolled on clinical trial will continue to be used in the following.

There are two levels of sampling in such projects. The set of patients enrolled on the study will be referred to here as the *cohort*. The subset of the cohort analyzed for biomarkers will be referred to as the *sample*. The larger set of patients meeting the criteria for entry on the clinical trial will be referred to as the *population*. The terms “study population” and “target population” could also be used for the cohort and the population. The population is assumed to be much larger than the cohort enrolled on the study, so that it can essentially be assumed to be infinite. Throughout, the focus is on drawing inferences on effects in the population, which is often referred to as “superpopulation” inference in the survey sampling literature (see e.g. Lin 2000); the problem of finite sample inferences on effects in the cohort of patients on the study is generally not of interest. Since the cohort is not a random sample from the population, the question of whether the patients on the clinical trial are a representative sample from the population should always be of concern. However, this issue cannot be directly addressed without additional data on the population, and it will not be considered further here.

The motivation for this paper came from a genomic project analyzing RNA expression of 371 genes in tumor tissue obtained from paraffin blocks from patients enrolled on an adjuvant breast cancer trial E2197 conducted by the Eastern Cooperative Oncology Group (Goldstein et al. 2008). The expense of the assays limited the genomic project to a subset of the 2952 patients enrolled on the clinical trial. The primary endpoint for the genomic study was time to disease recurrence, but there was interest in also examining other endpoints (such as disease-free survival and overall survival), and plans include updating analyses with additional follow-up at later times, so a more flexible design than time-matched nested case–control sampling (e.g. Prentice and Breslow 1978) was necessary. As many recurrences as possible and roughly 3.5 times as many non-recurrences were sampled from within strata defined by baseline clinical factors. Due to lack of consent, lack of tissue submission, and factors affecting evaluability of the assay, the final genomic sample included only 179 of 363 recurrences.

As discussed, for example, by Lin (2000) and Breslow and Wellner (2007) (and in additional references in those papers), one approach to estimation in such sampling designs is to use generalized Horvitz-Thompson estimators, where contributions are weighted inversely proportional to the sampling fractions in each group. Such estimators, while generally not fully efficient, often have good properties in practice. This paper provides details of applying this approach in the specific setting here, especially with regard to variance estimation. Although the sampling scheme here is similar to the "exclusive" (Rodrigues and Kirkwood 1990) or "cumulative" (Rothman and Greenland 1998) case–control design, with controls sampled from the non-cases at the end of the study period, the weighted analysis methods allow consistent estimates of hazard ratios and cumulative event probabilities in the population, provided there are no biases related to when the non-cases are sampled. In particular, the weighted analysis methods can also be applied to the classic case-cohort design (Chen and Lo 1999; Borgan et al. 2000).

The paper is organized as follows. In the following section, the basic data structure and results on estimating a population mean are reviewed. Comparisons of means between subpopulations and comparisons of the estimates of population parameters based on subjects in the sample and on those not in the sample are also considered. The latter problem is important because the sampling of subjects with observed events is generally not random, so the representativeness within the cohort of the resulting sample is of concern. Next, specifics of Lin (2000) general results on estimators for the proportional hazards model regression parameters and event-free probabilities are given in Sect. 3. Specialization to estimation of event-free probabilities in the one-sample case is also considered and is examined in a simulation study. Finally, an increasingly important issue for analyzing high dimensional biomarker data is validation of results (e.g. Dupuy and Simon 2007). Given the expense of the assays and limited availability of tissue, it is often impractical to have completely separate training and validation samples. Thus, some form of internal validation is often necessary. Some issues related to applying cross-validation and bootstrap methods in this context are considered in Sect. 4. Additional discussion is given in Sect. 5.

## 2 General notation and basic results

Stratified sampling is assumed, with separate random sampling from the subjects with observed events (cases) and those with censored event times (non-cases) within each stratum. Let $M_j$ be the number of subjects in stratum $j$ in the full cohort, $j = 1, \ldots, J$. In general, a '+' subscript will be used to denote summation over that subscript, so $M_+ = \sum_{j=1}^{J} M_j$ ($J = 8$ and $M_+ = 2952$ for E2197). The case/non-case status is determined by the primary event of interest at the time the sample is drawn, but other types of events may be analyzed and additional follow-up may be collected on the primary endpoint by the time of analysis. Thus in general, let $\delta_{ij} = 1$ for subjects with the primary event at the time the sample is drawn (cases) and $\delta_{ij} = 0$ for non-cases, let $T_{ij}$ be the event or censoring time of interest for a particular analysis, $\gamma_{ij}$ be the indicator of whether $T_{ij}$ is an event ($\gamma_{ij} = 1$) or censoring time ($\gamma_{ij} = 0$), where $\gamma_{ij}$ may or may not be the same as $\delta_{ij}$, $x_{ij}$ be a vector of baseline covariates, and $Z_{ij}$ be the value of some generic feature of interest (which could be a function of $T_{ij}$, $\gamma_{ij}$ and $x_{ij}$), for subject $i$ in stratum $j$. Let $D_j = \delta_{+j}$ be the number of potential cases in stratum $j$, so $M_j - D_j$ is the corresponding number of potential non-cases. Also, let $S_{ij} = 1$ if subject $i$ in stratum $j$ in the full cohort is sampled and $S_{ij} = 0$ otherwise, and set $d_j = \sum_i S_{ij}\delta_{ij}$ and $m_j = S_{+j}$, so $m_j$ is the total number of subjects, $d_j$ is the number of cases and $m_j - d_j$ is the number of non-cases sampled from stratum $j$.

The following assumptions are made:

1. Within each stratum, the observations $(T_{ij}, \gamma_{ij}, \delta_{ij}, x_{ij}, Z_{ij})$, $i = 1, \ldots, M_j$ are independent and identically distributed, and the strata are also independent.
2. The sampling is performed independently for cases and non-cases, and conditional on stratum and case status, subjects in the cohort are equally likely to be sampled. This implies that $Z_{ij}$ and $S_{ij}$ are independent conditional on stratum and the $\delta_{ij}$.
3. The values of $d_j$ and $m_j$ are fixed by design, or alternatively, are independent of other data values (ie covariates and event times) so the analysis can be conditioned on the actual number sampled.

Since fixed numbers of cases and non-cases are sampled from the finite cohort within each stratum, $S_{1j}, \ldots, S_{M_j,j}$ are not independent.

The sampling weights for the weighted estimators are defined by $w_{ij} = w_j(\delta_{ij})$, where

$$w_j(\delta) = \delta D_j/d_j + (1-\delta)(M_j - D_j)/(m_j - d_j),$$

which are the inverse sampling fractions within strata and case status groups.

Now consider estimating the marginal mean $\mu_Z = \mathrm{E}(Z_{ij})$, which is the mean in the population the cohort is drawn from. A straightforward estimator is

$$\hat{\mu}_Z = \sum_{j=1}^{J}\sum_{i=1}^{M_j} S_{ij} w_{ij} Z_{ij} \Big/ \sum_{j=1}^{J}\sum_{i=1}^{M_j} S_{ij} w_{ij} = \sum_{ij} S_{ij} w_{ij} Z_{ij}/M_+, \tag{1}$$

which has the form of a Horvitz-Thompson estimator. A basic introduction to Horvitz-Thompson estimators and finite population inference was given by Overton and Stehman (1995). Here we focus on superpopulation inference, as considered for example in Lin (2000).

Now $M_+\hat{\mu}_Z = Z_{++} + \tilde{Z}$, where $\tilde{Z} = \sum_{ij}\{S_{ij} w_j(\delta_{ij}) - 1\} Z_{ij}$, and as in Lin (2000), $Z_{++}$ and $\tilde{Z}$ are asymptotically independent, so $\mathrm{Var}(M_+\hat{\mu}_Z) \doteq \mathrm{Var}(Z_{++}) + \mathrm{Var}(\tilde{Z})$. The following proposition gives the main result, which will be used repeatedly in subsequent sections.

**Proposition 1** *Under the sampling assumptions here,*

$$\begin{aligned}\mathrm{Var}(\tilde{Z}) \doteq \sum_j \mathrm{E}_\delta[\{w_j(1)-1\}D_j \mathrm{Var}(Z_{ij}|\delta_{ij}=1) \\ +\{w_j(0)-1\}(M_j-D_j)\mathrm{Var}(Z_{ij}|\delta_{ij}=0)]. \qquad (2)\end{aligned}$$

This result follows from the joint distribution of the $S_{ij}$ given $\boldsymbol{\delta}=(\delta_{11},\ldots,\delta_{JM_J})'$ and from the independence of the $Z_{ij}$ and $S_{ij}$ given $\boldsymbol{\delta}$. Additional details are given in the Appendix. Thus in general, $\mathrm{Var}(\hat{\mu}_Z)$ can be estimated from the sample by

$$M_+^{-1}\left(\hat{V}_Z+\sum_j\left[\{w_j(1)-1\}D_j\hat{V}_{Zj}^1+\{w_j(0)-1\}(M_j-D_j)\hat{V}_{Zj}^0\right]\right), \qquad (3)$$

where $\hat{V}_Z=\sum_{ij}S_{ij}w_{ij}(Z_{ij}-\hat{\mu}_Z)^2/(M_+-1)$ and $\hat{V}_{Zj}^\delta=\sum_i I(\delta_{ij}=\delta)S_{ij}(Z_{ij}-\hat{\mu}_{Zj}^\delta)^2/(m_j^\delta-1)$, $\hat{\mu}_{Zj}^\delta=\sum_i I(\delta_{ij}=\delta)S_{ij}Z_{ij}/m_j^\delta$, and $m_j^\delta=\sum_i S_{ij}I(\delta_{ij}=\delta)$ (the $w_{ij}$ are constant within a stratum by case status combination, so they do not need to be included in the formulas for the $\hat{V}_{Zj}^\delta$). The quantities $\hat{V}_Z$ and $\hat{V}_{Zj}^\delta$ can be replaced by other appropriate estimators in special cases.

One special case is estimating the proportion of the population with some characteristic, such as the proportion of the population with primary tumors $\le 2$ cm for E2197. If $Z_{ij}=1$ for subjects with tumors $\le 2$ cm and $Z_{ij}=0$ for other subjects, then $\hat{\mu}_Z$ estimates the proportion of the population with this characteristic. In this case, $\hat{V}_Z$ can be replaced by $\hat{\mu}_Z(1-\hat{\mu}_Z)$ and $\hat{V}_{Zj}^\delta$ can be replaced by $\hat{\pi}_j^\delta(1-\hat{\pi}_j^\delta)$, where $\hat{\pi}_j^\delta=\sum_i I(\delta_{ij}=\delta)S_{ij}Z_{ij}/m_j^\delta$.

### 2.1 Conditional means

Conditional means and probabilities are often also of interest. For example, it may be of interest to estimate the mean gene expression level within subpopulations defined by baseline characteristics, such as hormone receptor positive and negative subsets or high, intermediate and low tumor grade subsets in E2197. Let $G_{ij}$ be a variable indicating the subpopulation level, $G_{ij}=1,\ldots,g$. Then a weighted estimator for $\mu_Z(l)=\mathrm{E}(Z_{ij}|G_{ij}=l)$ is

$$\hat{\mu}_Z(l)=\sum_{ij}S_{ij}w_{ij}I(G_{ij}=l)Z_{ij}\Big/\sum_{ij}S_{ij}w_{ij}I(G_{ij}=l).$$

Since

$$\begin{aligned}\hat{\mu}_Z(l)-\mu_Z(l)&=\frac{\sum_{ij}S_{ij}w_{ij}I(G_{ij}=l)\{Z_{ij}-\mu_Z(l)\}}{\sum_{ij}S_{ij}w_{ij}I(G_{ij}=l)}\\ &=\frac{\sum_{ij}S_{ij}w_{ij}I(G_{ij}=l)\{Z_{ij}-\mu_Z(l)\}}{M_+P(G_{ij}=l)}+o_p\left(M_+^{1/2}\right)\end{aligned}$$

has the same asymptotic form as (1), with $Z^*_{ij} = I(G_{ij} = l)\{Z_{ij} - \mu_Z(l)\}/P(G_{ij} = l)$ in place of $Z_{ij}$, it follows from Proposition 1 that the asymptotic variance of $\hat{\mu}_Z(l)$ can be estimated analogously to (3), using the estimated quantities

$$\hat{Z}^*_{ij} = I(G_{ij} = l)\{Z_{ij} - \hat{\mu}_Z(l)\}M_+ \Big/ \sum_{ij} S_{ij} w_{ij} I(G_{ij} = l).$$

Since $\hat{\mu}_Z(l)$ is constant for subjects with $I(G_{ij} = l)$, it can be dropped from $\hat{Z}^*_{ij}$ for computing empirical variances, and this process is equivalent to calculating the empirical variances in (3) directly from the $Z_{ij}$, but using just those subjects in the subset with $G_{ij} = l$.

A test of the hypothesis $H_0 : \mu_Z(l) - \mu_Z(l') = 0$ or a confidence interval on $\mu_Z(l) - \mu_Z(l')$ will also be of interest in some settings. Proceeding as before, $\hat{\mu}_Z(l) - \hat{\mu}_Z(l') - \{\mu_Z(l) - \mu_Z(l')\} = \sum_{ij} S_{ij} w_{ij} Z^{**}_{ij}/M_+ + o_p(M_+^{1/2})$, where

$$Z^{**}_{ij} = \frac{I(G_{ij} = l)\{Z_{ij} - \mu_Z(l)\}}{P(G_{ij} = l)} - \frac{I(G_{ij} = l')\{Z_{ij} - \mu_Z(l')\}}{P(G_{ij} = l')}$$

(under $H_0$, the $\mu_Z(\cdot)$ terms can be dropped). An estimate of the asymptotic variance can thus be obtained by using the $Z^{**}_{ij}$ in (3), substituting consistent estimates for the unknown quantities in $Z^{**}_{ij}$. Unless levels $l$ and $l'$ of $G$ are contained within separate sampling strata, in general $\hat{\mu}_Z(l)$ and $\hat{\mu}_Z(l')$ will be correlated.

### 2.2 Comparing the sample to the rest of the cohort

As noted in the Introduction, when drawing inferences on effects in the population studied in the clinical trial, the question of whether the sample analyzed in the biomarker study is representative of the full cohort is of interest. It is the estimate of the population distribution obtained using the inverse sampling weights that is relevant, not the raw distribution in the sample. The latter will usually differ from the target population because of different sampling fractions for cases and non-cases. While it is never possible to verify that the relationship of the biomarkers to outcomes obtained from the sample will be similar in the full cohort without evaluating the biomarkers on a random sample from the rest of the cohort, it is possible to compare the estimated population distribution of factors that are measured on the entire cohort in the clinical trial between those in the sample and those not in the sample. If these are similar for known factors, then it will at least be reassuring that the sample is representative with respect to standard factors.

The subjects not in the sample can be thought of as a second sample with selection indicators $1 - S_{ij}$ and sampling weights $w^c_{ij} = w^c_j(\delta_{ij})$, where

$$w^c_j(\delta) = \delta D_j/(D_j - d_j) + (1 - \delta)(M_j - D_j)/(M_j - D_j - m_j + d_j).$$

The quantity $\mathrm{E}(Z_{ij})$ can be estimated from this complementary set with

$$\hat{\mu}_Z^c = \sum_{j=1}^{J}\sum_{i=1}^{M_j}(1-S_{ij})w_{ij}^c Z_{ij}\Big/\sum_{j=1}^{J}\sum_{i=1}^{M_j}(1-S_{ij})w_{ij}^c = \sum_{ij}(1-S_{ij})w_{ij}^c Z_{ij}/M_+.$$

Then since $1/w_{ij} + 1/w_{ij}^c = 1$, so $1 + w_{ij}/w_{ij}^c = w_{ij}$, it follows that

$$\begin{aligned} M_+(\hat{\mu}_Z - \hat{\mu}_Z^c) &= \sum_{ij} Z_{ij}\{S_{ij}w_{ij} - (1-S_{ij})w_{ij}^c\} \\ &= \sum_{ij}\delta_{ij}Z_{ij}w_{ij}^c(S_{ij}w_{ij}-1) + \sum_{ij}(1-\delta_{ij})Z_{ij}w_{ij}^c(S_{ij}w_{ij}-1). \end{aligned}$$

Except for the $w_{ij}^c$ factors, this expression is the same as $\tilde{Z}$ above. The full cohort contribution ($Z_{++}$) has cancelled here. Applying similar arguments to those in the Appendix,

$$\begin{aligned} \mathrm{Var}\{M_+(\hat{\mu}_Z - \hat{\mu}_Z^c)\} = \sum_j \mathrm{E}_\delta \Big\{ & w_j^c(1)w_j(1)D_j\mathrm{Var}(Z_{ij}|\delta_{ij}=1) \\ & + w_j^c(0)w_j(0)(M_j - D_j)\mathrm{Var}(Z_{ij}|\delta_{ij}=0)\Big\}, \end{aligned}$$

since again $w_{ij} - 1 = w_{ij}/w_{ij}^c$. The $\mathrm{Var}(Z_{ij}|\delta_{ij})$ terms can be estimated from the observed data as discussed above, except that here data from the full cohort can be used (since this comparison can only be made for $Z_{ij}$ observed on the full cohort).

## 3 Weighted partial likelihood estimators

The proportional hazards model assumes that the event hazard rate for subject $ij$ is

$$\lambda(t|x_{ij}) = \lambda_0(t)\exp(\beta' x_{ij}) \tag{4}$$

for an unspecified underlying hazard function $\lambda_0(t)$, where $\beta$ is the vector of unknown regression parameters. The cumulative underlying hazard is $\Lambda_0(t) = \int_0^t \lambda_0(u)\,du$. The centered event counting process is

$$M_{ij}(t,\beta) = N_{ij}(t) - \int_0^t Y_{ij}(u)\exp(\beta' x_{ij})\,d\Lambda_0(u),$$

where $N_{ij}(t) = I(T_{ij} \le t, \gamma_{ij} = 1)$ is the event counting process and $Y_{ij}(t) = I(T_{ij} \ge t)$ is the at risk counting process. Recall that the event time and status data being analyzed can be different from that used to define the sampling case status. Generally the $M_{ij}(t,\beta)$ are not martingales conditional on selection in the sample.

The log weighted (pseudo) partial likelihood is

$$L^w(\beta) = \sum_{ij} \gamma_{ij} S_{ij} w_{ij} \left[ \beta' x_{ij} - \log \left\{ \sum_{lk} S_{lk} w_{lk} Y_{lk}(T_{ij}) \exp(\beta' x_{lk}) \right\} \right],$$

the weighted score is

$$U^w(\beta) = \partial L^w(\beta)/\partial\beta = \sum_{ij} S_{ij} w_{ij} \int \{x_{ij} - \overline{x}^w(t,\beta)\} \, dN_{ij}(t),$$

where

$$\overline{x}^w(t,\beta) = \frac{\sum_{ij} S_{ij} w_{ij} Y_{ij}(t) x_{ij} \exp(\beta' x_{ij})}{\sum_{ij} S_{ij} w_{ij} Y_{ij}(t) \exp(\beta' x_{ij})},$$

and the weighted ‘information’ is

$$\begin{aligned} I^w(\beta) &= -\frac{\partial^2 L^w(\beta)}{\partial\beta\partial\beta'} \\ &= \int \left( \frac{\sum_{ij} S_{ij} w_{ij} Y_{ij}(t) x_{ij} x'_{ij} \exp(\beta' x_{ij})}{\sum_{ij} S_{ij} w_{ij} Y_{ij}(t) \exp(\beta' x_{ij})} \right. \\ &\qquad \left. - \overline{x}^w(t,\beta) \overline{x}^w(t,\beta)' \right) \sum_{ij} S_{ij} w_{ij} \, dN_{ij}(t). \end{aligned}$$

The weighted partial likelihood estimator $\hat{\beta}$ satisfies the score equation $U^w(\hat{\beta}) = 0$. Lin (2000) shows that in considerable generality, $\hat{\beta} - \beta$ is approximately normal with mean 0 and variance

$$\mathrm{E}\{I^w(\beta)\}^{-1} \mathrm{Var}\{U^w(\beta)\} \mathrm{E}\{I^w(\beta)\}^{-1}.$$

When all cases are sampled, various authors have given results for $\mathrm{Var}\{U^w(\beta)\}$, including Chen and Lo (1999), Borgan et al. (2000) and Samuelsen et al. (2007). An extension of these results to allow subsampling of the cases and the possibility of analyzing different endpoints is needed here. Let $\mu_x(t,\beta)$ be the limit in probability of $\overline{x}^w(t,\beta)$. A key part of Lin (2000) development is showing that the asymptotic properties of $M_+^{-1/2} U^w(\beta)$ are the same as those of $M_+^{-1/2} \sum_{ij} S_{ij} w_{ij} U_{ij}$, where

$$U_{ij} = \int \{x_{ij} - \mu_x(t,\beta)\} \, dM_{ij}(t,\beta). \tag{5}$$

Note that $U_{ij}$ is a function of only the data from subject $ij$, so the $U_{ij}$ are independent in the full cohort. Thus the problem of estimating the variance of $U^w(\beta)$ fits into the framework of Sect. 2, with the vectors $U_{ij}$ in the role of the $Z_{ij}$. A vector generalization of Proposition 1 leads to the formula

$$\begin{aligned}\mathrm{Var}\{U^w(\beta)\} \doteq \mathrm{Var}\{U_{++}(\beta)\} + \sum_j \mathrm{E}_\delta[\{w_j(1)-1\}D_j \mathrm{Var}(U_{ij}|\delta_{ij}=1)\\ +\{w_j(0)-1\}(M_j-D_j)\mathrm{Var}(U_{ij}|\delta_{ij}=0)]. \qquad (6)\end{aligned}$$

If the $\gamma_{ij}$ are identical to the $\delta_{ij}$ and all observed events in the cohort are sampled, so $d_j = D_j$, then this formula is equivalent to that given for $\mathrm{Var}(\hat\beta)$ on the top half of page 106 in Samuelsen et al. (2007).

The variance of $U_{++}(\beta)$, which is the score from the full cohort, can be estimated by $I^w(\hat\beta)$, since $M_+^{-1}I^w$ converges to the same limit as the average information from the full cohort. To estimate the conditional variances in (6), the unknown quantities in $U_{ij}$ need to be replaced by estimates. Let $\hat U_{ij}$ be $U_{ij}$ with $\bar x^w(t,\beta)$ substituted for $\mu_x(t,\beta)$, $\hat\beta$ substituted for $\beta$ and the Breslow-style estimator substituted for the cumulative hazard $\Lambda_0(t)$ (see Sect. 3.1 below for more details). These $\hat U_{ij}$ are simply the score residuals from the fit of the model, available from the output in many packages, so the following algorithm can be used to estimate $\mathrm{Var}(\hat\beta)$:

1. Fit the model using weighted partial likelihood, obtaining the estimates $\hat\beta$ and the inverse information $I^w(\hat\beta)^{-1}$.
2. Obtain the score residuals from the fit of the model.
3. Calculate the empirical variance-covariance matrices $\hat V^1_{Uj}$ and $\hat V^0_{Uj}$ of the score residuals for the cases and non-cases in stratum $j$, for each $j$. (The case status for this calculation is defined by the status in the sampling data.)
4. Estimate $\mathrm{Var}(\hat\beta)$ with

$$\begin{aligned}I^w(\hat\beta)^{-1} + I^w(\hat\beta)^{-1}\sum_j \Bigg(\frac{D_j-d_j}{d_j}D_j\hat V^1_{Uj}\\ +\frac{M_j-D_j-m_j+d_j}{m_j-d_j}(M_j-D_j)\hat V^0_{Uj}\Bigg) I^w(\hat\beta)^{-1}.\end{aligned}$$

This formula is equivalent in their setting to that defined by Samuelsen et al. (2007) using the DFBETA case influence residuals, which combine the score residuals and information factors of the second term above.

The ‘robust’ variance estimator (Barlow 1994; Therneau and Grambsch 2000, Sect. 7.3) estimates $\mathrm{Var}\{U^w(\beta)\}$ with the empirical variance-covariance matrix of the quantities $w_{ij}\hat U_{ij}$ (the weighted score residuals). While the robust variance estimator is valid under certain circumstances, it can overestimate the true variance from stratified case-cohort samples, especially for the effects of factors used in defining sampling strata (see Samuelsen et al. 2007).

Score tests with estimated nuisance parameters will also be of interest in some settings. Write $\beta' = (\phi', \psi')$ and $U^w(\beta)' = \{U^w_\phi(\phi,\psi)', U^w_\psi(\phi,\psi)'\}$, and consider the hypothesis $H_0 : \phi = 0$. Let $\hat\psi_0$ be the weighted partial likelihood estimator under $H_0$, which is the solution to $U^w_\psi(0,\psi) = 0$. The score test statistic is $U^w_\phi(0,\hat\psi_0)'\widehat{\mathrm{Var}}\{U^w_\phi(0,\hat\psi_0)\}^{-1}U^w_\phi(0,\hat\psi_0)$. Using parameters as subscripts to denote the subvectors and submatrices corresponding to those parameters (as with $U^w_\phi$ and $U^w_\psi$), and using standard Taylor series methods, under $H_0$,

$$U^w_\phi(0, \hat{\psi}_0) \doteq U^w_\phi(0, \psi) - I^w_{\phi\psi}(0, \psi) I^w_{\psi\psi}(0, \psi)^{-1} U^w_\psi(0, \psi).$$

Thus an estimator of the null variance of $U^w_\phi(0, \hat{\psi}_0)$ is given by $B\hat{V}B'$, where $\hat{V}$ is the estimate of (6), $B = \{I, -I^w_{\phi\psi}(I^w_{\psi\psi})^{-1}\}$, and both $\hat{V}$ and $I^w$ are evaluated at $(\phi, \psi) = (0, \hat{\psi}_0)$.

### 3.1 Cumulative hazard and survivor function estimation

The cumulative hazard for a subject with covariates $x$, based on model (4) is $\Lambda(t|x) = \exp(\beta' x)\Lambda_0(t)$, and the event-free probability at $t$ is $S(t|x) = \exp\{-\exp(\beta' x)\Lambda_0(t)\}$. The generalization of the Breslow estimator of $\Lambda_0(t)$ to survey sampling is $\hat{\Lambda}(t, \hat{\beta}, 0)$, where

$$\begin{aligned}\hat{\Lambda}(t, \beta, x) &= \exp(\beta' x) \int_0^t \frac{\sum_{ij} S_{ij} w_{ij} dN_{ij}(u)}{\sum_{ij} S_{ij} w_{ij} Y_{ij}(u) \exp(\beta' x_{ij})} \\ &= \int_0^t \frac{\sum_{ij} S_{ij} w_{ij} dN_{ij}(u)}{\sum_{ij} S_{ij} w_{ij} Y_{ij}(u) \exp\{\beta'(x_{ij} - x)\}}.\end{aligned}$$

An estimator of $S(t|x)$ is then given by $\hat{S}(t|x) = \exp\{-\hat{\Lambda}(t, \hat{\beta}, x)\}$. Lin (2000) shows that, as for the other quantities considered here, $\mathrm{Var}\{\hat{\Lambda}(t, \hat{\beta}, x)\} \doteq V_c(t, x) + V_s(t, x)$, where $V_c(t, x)$ is the variance of the corresponding estimator computed from the full cohort and $V_s(t, x)$ is the extra variation due to the finite cohort subsampling (Lin only considers the case $x = 0$, but the extension to the version here follows trivially by centering the covariates at $x$). The quantity $V_c(t, x)$ can be estimated from the sample with

$$\hat{V}_c(t, x) = \int_0^t \frac{\sum_{ij} S_{ij} w_{ij}\, dN_{ij}(u)}{S_0(u, \hat{\beta}, x)^2} + R(t, \hat{\beta}, x)' I^w(\hat{\beta})^{-1} R(t, \hat{\beta}, x),$$

where $S_0(u, \beta, x) = \sum_{ij} S_{ij} w_{ij} Y_{ij}(u) \exp\{\beta'(x_{ij} - x)\}$ and

$$\begin{aligned}R(t, \beta, x) = -\int_0^t &\left\{ \sum_{ij} (x_{ij} - x) S_{ij} w_{ij} Y_{ij}(u) \exp\{\beta'(x_{ij} - x)\} \right\} \\ &\times \frac{\sum_{ij} S_{ij} w_{ij}\, dN_{ij}(u)}{S_0(u, \beta, x)^2}\end{aligned}$$

(this formula is essentially the same as that given on the upper part of page 43 of Lin 2000).

The term $V_s(t, x)$ can be obtained using the general approach from Sect. 2 (and the Appendix). From Eq. 2.7 of Lin (2000) and the subsequent material there, the appropriate $Z_{ij}$ terms here, based on an expansion of $\hat{\Lambda}(t, \hat{\beta}, x) - \Lambda(t|x)$ can be estimated from the sample with

$$\begin{aligned}\hat{Z}_{ij} &= \int_0^t \frac{dN_{ij}(u)}{S_0(u, \hat{\beta}, x)} - \int_0^t \frac{Y_{ij}(u)\exp\{\hat{\beta}'x_{ij}\}}{S_0(u, \hat{\beta}, x)} d\hat{\Lambda}(t, \hat{\beta}, 0) + R(t, \hat{\beta}, x)' I^w(\hat{\beta})^{-1}\hat{U}_{ij} \\ &= \int_0^t \frac{dN_{ij}(u)}{S_0(u, \hat{\beta}, x)} - \int_0^t \frac{Y_{ij}(u)\exp\{\hat{\beta}'(x_{ij} - x)\}}{S_0(u, \hat{\beta}, x)} d\hat{\Lambda}(t, \hat{\beta}, x) \\ &\quad + R(t, \hat{\beta}, x)' I^w(\hat{\beta})^{-1}\hat{U}_{ij},\end{aligned}$$

where $U_{ij}$ is defined in (5) and $\hat{U}_{ij}$ is the score residual estimate of $U_{ij}$ defined above. An estimate $\hat{V}_s(t, x)$ of $V_s(x, t)$ can be computed from the $\hat{Z}_{ij}$ as in (3).

The specialization of $\hat{\Lambda}(t, \hat{\beta}, x)$ to the one-sample case, which can be defined by $\hat{\Lambda}^w(t) = \hat{\Lambda}(t, 0, 0)$, is also of interest. In most applications, such estimates would be computed in subsets of the data defined by baseline characteristics. The calculations for a subset follow the same as for the full sample here, with the sums restricted to the subset of interest.

The general results for $\hat{\Lambda}(t, \hat{\beta}, x)$ still apply in this simplified setting, and $\hat{Z}_{ij}$ simplifies to

$$\hat{Z}_{ij} = \int_0^t \frac{dN_{ij}(u) - Y_{ij}(u)d\hat{\Lambda}^w(u)}{\sum_{lk} S_{lk} w_{lk} Y_{lk}(u)}, \tag{7}$$

and $\hat{V}_c(t, 0)$ can be written

$$\sum_{ij} S_{ij} w_{ij} \int_0^t \frac{dN_{ij}(u)}{\{\sum_{lk} S_{lk} w_{lk} Y_{lk}(u)\}^2}. \tag{8}$$

A simple estimate of $S(t)$ is then given by $\tilde{S}^w(t) = \exp\{-\hat{\Lambda}(t)\}$. By the delta method, an estimator of $\text{Var}\{\tilde{S}(t)\}$ is given by $\tilde{S}^w(t)^2 \hat{\text{Var}}\{\hat{\Lambda}^w(t)\}$. A product limit version can also be given. Specifically, by analogy with the unweighted case, let $t_1, \ldots, t_K$ be the unique event times, and define

$$\begin{aligned}\hat{S}^w(t) &= \prod_{t_k \le t}\{1 - \hat{\Lambda}^w(t_k) + \hat{\Lambda}^w(t_k-)\} \\ &= \prod_{t_k \le t}\left(1 - \frac{\sum_{ij} S_{ij} w_{ij} I(T_{ij} = t_k, \delta_{ij} = 1)}{\sum_{ij} S_{ij} w_{ij} Y_{ij}(t_k)}\right).\end{aligned}$$

The asymptotic variances of $\hat{S}^w(t)$ and $\tilde{S}^w(t)$ will be the same.

### 3.2 A simulation study of the cumulative hazard estimator

Here performance of the variance estimator $\hat{V}$ for $\hat{\Lambda}^w(t)$, based on (7) and (8) is considered for the one-sample estimator computed for a single stratum. Full cohort event times are generated from an exponential distribution with 5-year event-free probabilities $S(5)$ of either 0.90 or 0.95. Censoring times are drawn from the uniform (3,8) distribution. The cumulative hazard $\hat{\Lambda}^w(t)$ is computed at $t = 3, 5, 7$. Through $t = 3$ there is complete follow-up and at $t = 7$ the risk sets are becoming small, so the times span a considerable range of conditions. Cohort sizes in the stratum of $M_j = 1000$, 2952 are considered. In all cases $m_j - d_j = 150$ and $d_j = \min\{50, D_j\}$. Cohorts with $D_j < 50$ only occur for the combination $M_j = 1000$ and $S(5) = 0.95$, where the expected number of events in the cohort is 54.76 and approximately 24% of the generated cohorts should contain $<$ 50 events. For each configuration, 10,000 samples are generated. In each case, for each time point, the average of the $\hat{\Lambda}^w(t)$, the empirical variance of the $\hat{\Lambda}^w(t)$ (which estimates the true variance), the average of the variance estimators computed from each sample (E($\hat{V}$)), and the true coverage probabilities of nominal 95% confidence intervals are computed. Confidence intervals are computed from the normal approximation on both the $\Lambda(t)$ scale and the $\log\{\Lambda(t)\}$ scale, with the variance estimates for the latter case obtained using the delta method. For each case, the proportion of the samples where the true value was smaller than the lower confidence limit ($<$LCL) and the proportion where true value was larger than the upper confidence limit ($>$UCL) are given (both should be 2.5%). The results, given in Table 1, show that the bias in the estimator and in the variance estimator are minimal, but the performance of the confidence intervals is mixed. The scenarios and time points have been chosen to have low event probabilities, where cohort sampling methods may be most useful, but where convergence to normality may be slow. There is generally some skewness in the distribution of $\hat{\Lambda}^w(t)$ (that is sometimes overcorrected

**Table 1** Simulation results for estimating the cumulative hazard (entries for Var$\{\hat{\Lambda}^w(t)\}$ and E($\hat{V}$) are multiplied by 10,000)

| $M_j$ | $S(5)$ | $t$ | $\Lambda(t)$ | E$\{\hat{\Lambda}^w(t)\}$ | Var$\{\hat{\Lambda}^w(t)\}$ | E($\hat{V}$) | CI on $\Lambda(t)$ | | CI on $\log\{\Lambda(t)\}$ | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | <LCL (%) | >UCL (%) | <LCL (%) | >UCL (%) |
| 1000 | 0.90 | 3 | 0.0632 | 0.0632 | 1.04 | 1.01 | 1.7 | 4.3 | 3.3 | 1.9 |
| | | 5 | 0.1054 | 0.1052 | 1.50 | 1.54 | 1.7 | 3.3 | 2.6 | 2.0 |
| | | 7 | 0.1475 | 0.1474 | 3.31 | 3.45 | 0.6 | 4.8 | 1.6 | 3.2 |
| 2952 | 0.90 | 3 | 0.0632 | 0.0633 | 0.78 | 0.78 | 2.4 | 3.1 | 3.8 | 1.4 |
| | | 5 | 0.1054 | 0.1053 | 0.89 | 0.88 | 2.3 | 2.8 | 3.3 | 1.8 |
| | | 7 | 0.1475 | 0.1478 | 2.11 | 2.13 | 0.6 | 5.7 | 1.4 | 4.6 |
| 1000 | 0.95 | 3 | 0.0308 | 0.0307 | 0.33 | 0.33 | 1.1 | 4.9 | 3.1 | 2.0 |
| | | 5 | 0.0513 | 0.0512 | 0.62 | 0.61 | 1.5 | 4.6 | 3.0 | 2.2 |
| | | 7 | 0.0718 | 0.0720 | 1.29 | 1.31 | 0.8 | 4.7 | 2.2 | 2.8 |
| 2952 | 0.95 | 3 | 0.0308 | 0.0307 | 0.21 | 0.22 | 1.9 | 3.5 | 3.2 | 1.5 |
| | | 5 | 0.0513 | 0.0513 | 0.29 | 0.29 | 1.8 | 3.2 | 2.9 | 2.0 |
| | | 7 | 0.0718 | 0.0722 | 0.67 | 0.67 | 0.9 | 4.8 | 1.9 | 3.3 |

by use of the log transformation) and sometimes fairly high correlation between the estimate and its estimated standard error (that can be in the opposite direction on the log scale), which jointly lead to the observed problems with the confidence interval coverage. Both factors vary with the follow-up time point and the scenario, leading to the complex pattern in Table 1, which makes it difficult to predict performance.

## 4 Cross-validation and bootstrap resampling

Cross-validation provides a general approach to validation of results when the amount of available data is limited. Cross-validation can be used to estimate the accuracy of the procedure used to develop a classifier or predictor in the context of the particular problem (see e.g. Dupuy and Simon 2007; Rademacher et al. 2002), but not the accuracy of the particular model or predictor. Because of the amount of data needed to give valid estimates of accuracy for the right-censored endpoints of interest here, $K$-fold cross-validation, where the data are randomly divided into $K$ groups, is considered. The model or predictor is developed on the data with one of the groups omitted, and then the omitted group is used to evaluate how well the model classifies subjects or predicts outcomes. This process is repeated with each group omitted in turn, and the average prediction accuracy over the omitted sets gives an estimate of the accuracy when the procedure is applied to the full data set (the data sets used to fit the model in the cross-validation procedure contain $(K-1)/K$ of the full sample, and so should be reasonably representative if $K$ is not too small).

The dependence in the sample is a problem, since the cross-validation sets need to be independent. Since subjects in the full cohort are independent, cross-validation can conceptually be applied at the level of the full cohort and then the sampling of cases and non-cases can be done within the cross-validation subsets. This can be implemented by dividing the actual sample into $K$ groups and then recomputing the weights based on a corresponding division of the portion of the cohort that is not in the sample, as follows. The only information needed on the full cohort are the values of the $M_j$ and the $D_j$. First the $m_j$ subjects in the sample in stratum $j$ can be randomly divided into $K$ subsets of size $m_{kj}$, $k=1,\ldots,K$. Then the $M_j - m_j$ subjects in stratum $j$ who are not in the sample can also be randomly divided into $K$ subsets of size $M^*_{kj}$, $k=1,\ldots,K$. The $m_{kj}$ and $M^*_{kj}$ can be fixed and each can be set to be as close to equal (in $k$) as possible. While the number of subjects in the groups can be fixed, in order to obtain independent subsets, the number of cases and non-cases cannot be. Let $d_{kj}$ be the number of cases in the sample and $D^*_{kj}$ the number of cases in the cohort not in the sample assigned to the $k$th cross-validation set in stratum $j$. For the subjects not in the sample, all that matters in the random division into $K$ subsets is determining the $D^*_{kj}$, and under random division into subsets, these have a multivariate hypergeometric distribution, and so can be generated without access to individual patient data for the subjects not in the sample. For analyses of the $k$th cross-validation training set, the sampling weights are then set to $\sum_{l\neq k}(D^*_{lj}+d_{lj})/\sum_{l\neq k} d_{lj}$ for cases and $\sum_{l\neq k}(M^*_{lj}+m_{lj}-D^*_{lj}-d_{lj})/\sum_{l\neq k}(m_{lj}-d_{lj})$ for non-cases, and for analyses of the $k$th validation set, the sampling weights are set to $(D^*_{kj}+d_{kj})/d_{kj}$ for cases and to $(M^*_{kj}+m_{kj}-D^*_{kj}-d_{kj})/(m_{kj}-d_{kj})$ for non-cases. In this way, each omitted validation set in the $K$-fold cross validation is independent of the complementary training set, since each is constructed from independent subsets of the full cohort.

It is possible to use nested application of this process, applying cross validation to each training set in the top level cross validation procedure, say for the purpose of optimizing tuning parameters in a classification algorithm on each training set.

The dependence in the sample also creates a challenge for bootstrap sampling, since the bootstrap resamples should reproduce the dependence structure in the sample. In general, bootstrapping dependent data is a complex problem, but here there is again a simple solution based on generating bootstrap samples of the full cohort and repeating the process of selecting cases and non-cases for each bootstrap data set. As with the cross-validation procedure, it is only necessary to have information on the $M_j$ and $D_j$ from the full cohort to do this. Since the original sampling was performed separately within strata, the bootstrap resampling can also be done within strata. Within each stratum, the population distribution of the data can be estimated with the weighted empirical distribution of the subjects in the stratum in the sample, where the weights are the $w_{ij}$. Sampling the required number of subjects from the empirical distribution can be done by weighted sampling with replacement from the subjects in the stratum in the sample. Other than the fact that subjects with and without observed events have different sampling weights, the generation of the full cohort ignores event status (that is, it does not attempt to replicate the number of events in the original cohort). This process can be used to generate a full $M_+$-subject study cohort. Cases and non-cases can then be sampled at random within strata from this cohort to give a new sample, with the same number of cases and non-cases within strata. Sampling weights can then be computed from the bootstrap cohort resample and the selected subset as in the original data set. This process is equivalent to just resampling subjects in the sample, with a further random adjustment to the weights based on the distribution of the potential cases and non-cases within strata in the full cohort bootstrap resamples.

## 5 Discussion

This paper gives details for implementation of weighted analyses for event-stratified subsampling for biomarker studies in clinical trials and epidemiologic cohort studies. Such designs potentially have wide application in studies with low event rates. They allow straightforward analysis of multiple endpoints (e.g. survival, disease-free survival and recurrence-free interval), and also allow incorporating additional follow-up after the sample has been selected.

While stratified cohort sampling designs have advantages in flexibility over nested case–control sampling (where the controls for each case are sampled from the risk set at the case's event time), these advantages could be offest by the possible effects of analytic batch, long-term storage, and freeze-thaw cycles on the biomarker evaluations. In the E2197 genomic project, RNA was extracted from tumor blocks created at the time of diagnosis (shortly before study entry), and the RNA extraction and genomic evaluation were performed on the entire sample within a short period of time, so these factors should be minimal. However, for settings where these factors can be substantial, Rundle et al. (2005) argue that the nested case–control design is preferred, since it allows the tissue sample for each case to be compared to a matched set of control samples.

The purpose of stratified cohort sampling designs is to improve efficiency over a simple random sample from the cohort of study subjects. Further work is needed to guide the choice of strata and the ratio of non-cases to cases in such designs. For example, if the purpose is to estimate the marginal relationship between factor $X$ and risk of recurrence, and factor $X$ is thought to be correlated with factor $Y$, which has been measured on the full cohort, then better understanding is needed of how to best use the information on $Y$ in selecting the sample. Currently, it appears simulation methods are required to investigate power of alternative sampling designs.

The weighted analysis methods used here are not efficient, but efficient estimators in related designs tend to be complex; see e.g. Nan (2004) and Scheike and Martinussen (2004). The local averaging method of Chen (2001) also has better efficiency than the weighted methods discussed here for case-cohort sampling. This method was extended to stratified sampling by Samuelsen et al. (2007), who showed that the method corresponds to sampling (or post stratification) within intervals defined by length of follow-up, in addition to case/non-case status. However, as discussed in more detail in Breslow and Wellner (2007) and elsewhere, the convenience and flexibility of the weighted estimation methodology makes it a reasonable option except when the efficiency loss may be substantial.

**Acknowledgements** Discussions with collaborators at Genomic Health Inc. and sanofi aventis on the E2197 analysis led to numerous refinements in the paper. Tianxi Cai suggested the concept for cross-validation based on dividing the full cohort. Partial funding was provided by sanofi aventis and a grant from the National Cancer Institute.

## Appendix

Here additional details of the derivation of (2) are given. Using standard results on Horvitz-Thompson estimators (see e.g. Overton and Stehman 1995; Lin 2000), and since sampling is conducted independently within stratum by event status groups, it follows that asymptotically in the super-population model,

$$
\begin{aligned}
\mathrm{Var}(\tilde{Z}) &= \sum_{j=1}^{J} \mathrm{E}\left\{\sum_{i}\sum_{l} \mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})Z_{ij}Z_{lj}\right\} \\
&= \sum_{j=1}^{J} \mathrm{E}_{\delta}\left\{\sum_{i}\sum_{l} \mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})\mathrm{E}(Z_{ij}Z_{lj}|\boldsymbol{\delta})\right\} \\
&= \sum_{j=1}^{J} \mathrm{E}_{\delta}\left\{\sum_{i} \mathrm{Var}(S_{ij}w_{ij}-1|\boldsymbol{\delta})\mathrm{Var}(Z_{ij}|\boldsymbol{\delta})\right.
\end{aligned}
$$

$$+\sum_i \sum_l \mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})\mathrm{E}(Z_{ij}|\boldsymbol{\delta})\mathrm{E}(Z_{lj}|\boldsymbol{\delta})\Bigg\}$$

using the independence of the $Z_{ij}$. Now $\mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta}) = 0$ when $\delta_{ij} \neq \delta_{lj}$, and $\mathrm{E}(Z_{ij}|\boldsymbol{\delta})$ depends on $i$ only through the value of $\delta_{ij}$, so

$$\begin{aligned}
&\sum_i \sum_l \mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})\mathrm{E}(Z_{ij}|\boldsymbol{\delta})\mathrm{E}(Z_{lj}|\boldsymbol{\delta})\\
&= \mathrm{E}(Z_{ij}|\boldsymbol{\delta}, \delta_{ij}=1)^2 \sum_i \sum_l \delta_{ij}\delta_{lj}\mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})\\
&\quad + \mathrm{E}(Z_{ij}|\boldsymbol{\delta}, \delta_{ij}=0)^2 \sum_i \sum_l (1-\delta_{ij})(1-\delta_{lj})\mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta})\\
&= 0,
\end{aligned}$$

since, for example,

$$\sum_{il} \delta_{ij}\delta_{lj}\mathrm{Cov}(S_{ij}w_{ij}-1, S_{lj}w_{lj}-1|\boldsymbol{\delta}) = \mathrm{Var}\left\{\sum_i \delta_{ij}(S_{ij}w_{ij}-1)\middle|\boldsymbol{\delta}\right\}$$

and $\sum_i \delta_{ij}(S_{ij}w_{ij}-1) = \sum_i \delta_{ij}\{S_{ij}w_j(1)-1\} = d_j(D_j/d_j) - D_j = 0$. Also, $\mathrm{Var}(S_{ij}w_{ij}-1|\boldsymbol{\delta}) = w_{ij}^2(1/w_{ij})(1-1/w_{ij}) = w_{ij}-1$, so

$$\mathrm{Var}(\tilde{Z}) = \sum_{j=1}^{J} \mathrm{E}\left\{\sum_i (w_{ij}-1)\mathrm{Var}(Z_{ij}|\boldsymbol{\delta})\right\}.$$

Formula (2) follows by splitting this expression into sums over subjects with $\delta_{ij} = 1$ and $\delta_{ij} = 0$.

## References

Barlow WE (1994) Robust variance estimation for the case-cohort design. Biometrics 50:1064–1072

Borgan O, Langholz B, Samuelsen SO, Goldstein L, Pagoda J (2000) Exposure stratified case-cohort designs. Lifetime Data Anal 6:39–58

Breslow NE, Wellner JA (2007) Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. Scand J Stat 34:88–102

Chen K, Lo S-H (1999) Case-cohort and case–control analysis with Cox's model. Biometrika 86:755–764

Chen K (2001) Generalized case-cohort sampling. J R Stat Soc B 63:791–809

Dupuy A, Simon RM (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst 99:147–157

Goldstein LJ, Gray R, Badve S, Childs BH, Yoshizawa C, Rowley S, Shak S, Baehner FL, Ravdin PM, Davidson NE, Sledge GW, Perez E, Shulman LN, Martino S, Sparano JA (2008) Prognostic utility of the 21-gene assay in hormone receptor-positive operable breast cancer compared with classical clinicopathologic features. J Clin Oncol 26(25). doi:10.1200/JCO.2007.14.4501

Lin DY (2000) On fitting Cox's proportional hazards models to survey data. Biometrika 87:37–47

Nan B (2004) Efficient estimation for case-cohort studies. Can J Stat 32:403–419

Overton WS, Stehman SV (1995) The Horvitz-Thompson theorem as a unifying perspective for probability sampling: with examples from natural resource sampling. Am Stat 49:261–268

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

Prentice RL, Breslow NE (1978) Retrospective studies and failure time models. Biometrika 65:153–158

Rademacher MD, McShane LM, Simon R (2002) A paradigm for class prediction using gene expression profiles. J Comput Biol 9:505–511

Rodrigues L, Kirkwood BR (1990) Case–control designs in the study of common diseases: updates on the demise of the rare disease assumption and the choice of sampling scheme for controls. Int J Epidemiol 19:205–213

Rothman KJ, Greenland S (1998) Modern Epidemiology, 2nd edn. Lippincott Williams & Wilkins, Philadelphia

Rundle AG, Vineis P, Ahsan H (2005) Design options for molecular epidemiology research within cohort studies. Cancer Epidemiol Biomarkers Prev 14:1899–1907

Samuelsen SO, Anestad H, Skrondal A (2007) Stratified case-cohort analysis of general cohort sampling designs. Scand J Stat 34:103–119

Scheike TH, Martinussen T (2004) Maximum likelihood estimation for Cox's regression model under case-cohort sampling. Scand J Stat 31:283–293

Therneau TM, Grambsch PM (2000) Modeling survival data. Springer, New York