Communications for Statistical Applications and Methods
2013, Vol. 20, No. 6, 455–466

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.6.455

Analysis of Nested Case-Control Study Designs: Revisiting the Inverse Probability Weighting Method

Ryung S. Kim[1,a]

[a]Department of Epidemiology and Population Health, Albert Einstein College of Medicine

Abstract

In nested case-control studies, the most common way to make inference under a proportional hazards model is the conditional logistic approach of Thomas (1977). Inclusion probability methods are more efficient than the conditional logistic approach of Thomas; however, the epidemiology research community has not accepted the methods as a replacement of the Thomas' method. This paper promotes the inverse probability weighting method originally proposed by Samuelsen (1997) in combination with an approximate jackknife standard error that can be easily computed using existing software. Simulation studies demonstrate that this approach yields valid type 1 errors and greater powers than the conditional logistic approach in nested case-control designs across various sample sizes and magnitudes of the hazard ratios. A generalization of the method is also made to incorporate additional matching and the stratified Cox model. The proposed method is illustrated with data from a cohort of children with Wilm's tumor to study the association between histological signatures and relapses.

Keywords: Nested Case-Control, Inverse Probability Weighting, Approximate Jackknife Standard Error

1. Introduction

The nested case-control design, like the case-cohort design, is a schema in which a representative sample of a full cohort is used. It includes all cases and a pre-specified number of controls randomly chosen from the risk set of each failure time (Thomas, 1977). The design is also referred as incidence density sampling or risk set sampling. It is typically used to reduce the cost of exposure assessment in a prospective epidemiologic study since exposure and other covariate data are obtained from only a subset of the full cohort yet without much loss of efficiency. It is a representative sample of a full cohort and, unlike traditional case-control studies, one can estimate the hazard ratios of the full cohort and the population it represents.

A partial likelihood approach was put forth by Thomas (1977) to estimate the hazard ratios from nested case-control studies. Since his partial likelihood is mathematically equivalent to the conditional logistic likelihood used for matched case-control studies, one can maximize the proposed likelihood by 'tricking' statistical software written for conditional logistic regression or equivalently stratified Cox regression. This is accomplished by including multiple inputs for subjects who are selected multiple times, and converting all randomly selected failures to non-failures. It is worth noting that although the mathematical formula for the conditional logistic likelihood is being used, this is not a matched case-control design since failures can be selected as controls and a subject can be selected

This work was supported by National Institutes of Health Grants 1UL1RR025750-01, P30 CA01330-35, and the National Research Foundation of Korea Grant NRF-2012S1A3A2033416.

[1] Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, U.S.A. E-mail: ryung.kim@einstein.yu.edu

multiple times across different failure times. Consequently, linear coefficients should be interpreted as log hazard ratios and not as log odds-ratios (Langholz, 2010).

A few have proposed methods that are more efficient than that of Thomas. For example, Samuelsen (1997) proposed an analysis method in which the individual log-likelihood contributions are weighted by the inverse of the inclusion probabilities of ever being included in the nested case-control study. Chen (2001, 2004) proposed the use of the same form of likelihood but with refined weights by averaging the observed covariates from subjects with similar failure times to estimate the contribution from unselected controls. Both Samuelsen (1997) and Chen (2001) showed that their estimators were more efficient than Thomas'; however, their methods have not been accepted as replacements by the epidemiology research community for the Thomas' method. Nested case-control studies are still almost exclusively analyzed by means of the conditional logistic regression approach of Thomas. For instance, Kim (2013) recently reported that among sixteen nested case-control studies published in the American Journal of Epidemiology between 2009 and 2011, fourteen were analyzed by the conditional logistic regression approach of Thomas and the remaining two by the unconditional logistic regression approach; none used Samuelsen or Chen's method.

The lack of acceptance is in part due to the following reasons. First, there was a lack of explanation on how the magnitude of the hazard ratio and the sample size affect the relative performance of different methods. Second, common statistical software cannot compute the complex variance estimators which require a computational memory in the order of $O(n^2)$ to compute all pair-wise co-inclusion probabilities. Finally, the methods were not extended to incorporate additional matching factors or to perform the stratified Cox analysis.

The aim of this paper is to promote the use of the inverse probability weighting methods in nested case-control studies using Cox proportional hazards model through demonstration of their superiority over the conditional logistic approach. I also demonstrate that approximate jackknife standard errors can be used to make valid inferences about log hazard ratios. The standard error estimator was recently proposed for secondary outcome analyses in nested case-control studies (Kim, 2013) and is easily computable with existing software. As an illustration, we will use a cohort of children from the US with Wilm's tumor (Breslow and Chatterjee, 1999) to study the association of histology with relapses while incorporating matching variables and performing stratified analyses.

2. Methods

Consider a cohort of N subjects who are followed for the occurrence of a failure event. Let a_i denote the fixed time i^{th} subject ($i = 1, \ldots, N$) entered the study, T_i denote the time to the failure event, C_i denote the censoring time that is independent of T_i, and $Y_i = \min(T_i, C_i)$ denote the observed time. Assume that the hazard function $\lambda_i(t)$ of the failure time for the i^{th} subject follows the proportional hazards model

$$\lambda_i(t) = \lambda_0(t) e^{X_i(t)\beta}, \tag{2.1}$$

where $\lambda_0(t)$ is the baseline hazard function, β is the parameter vector of interest, and $X_i(Y_i)$ is a time-dependent covariate vector for the i^{th} subject. Then, in full cohort studies, inferences on β are typically made by maximizing the Cox partial likelihood (Cox, 1972):

$$L_C(\beta) = \prod_{i=1}^{N} \left[\frac{e^{X_i(Y_i)\beta}}{\sum_{j \in R_i} e^{X_j(Y_i)\beta}} \right]^{\delta_i}. \tag{2.2}$$

$\delta_i = 1$ if subject i failed during the study and 0 otherwise; $R_i = \{j : Y_j \geq Y_i > a_j\}$ is the set of subjects at risk in the underlying cohort at time Y_i. In a more general model that allows different baseline hazard functions across subgroups, the following partial likelihood is maximized where $S(i)$ is the index set of the subjects who have the same values of matching variables with subject i:

$$L_C^S(\beta) = \prod_{i=1}^{N} \left[\frac{e^{X_i(Y_i)\beta}}{\sum_{j \in R_i \cap S(i)} e^{X_j(Y_i)\beta}} \right]^{\delta_i}. \tag{2.3}$$

In nested case-control studies, m controls are sampled from $R_i \cap \{i\}^c$ without replacement at each Y_i where $\delta_i = 1$. That is, for each case, m controls are randomly selected from the subjects still at risk at the time of the failure of the case. Notice that the controls may include both failures and non-failures. Let S_i denote this set of m controls and $S = \{i : \delta_i = 1\} \bigcup (\bigcup_{i:\delta_i=1} S_i)$ denote all subjects included in the nested case-control study. Then $\tilde{R}_i = R_i \cap S$ is the set of all subjects in the nested case-control study who are at risk at time Y_i. Let n indicate the size S. Thomas (1977) proposed maximizing the following partial likelihood to make inferences on β's from nested case-control studies:

$$L_{Thomas}(\beta) = \prod_{i \in S} \left[\frac{e^{X_i(Y_i)\beta}}{\sum_{j \in \{i\} \cup S_i} e^{X_j(Y_i)\beta}} \right]^{\delta_i}. \tag{2.4}$$

The partial likelihood produces a model consistent estimator of log hazard ratios (Borgan and Langholz, 1993). For this paper, it is important to notice that the denominators in the Thomas' likelihood uses only $\{i\} \cap S_i$ and not all available subjects at risk, namely $\tilde{R}_i$. As we will see, such 'partial risk set' approach is inherently inefficient because it uses only a subset of available subjects at risk. Samuelsen (1997) proposed maximizing the following partial-likelihood:

$$L_{IPW}(\beta) = \prod_{i \in S} \left[\frac{e^{X_i(Y_i)\beta}}{\sum_{j \in \tilde{R}_i} w_j e^{X_j(Y_i)\beta}} \right]^{w_i \cdot \delta_i}, \tag{2.5}$$

where $w_i = 1/p_i$, and p_i is the probability of subject i ever being included in the nested case-control study. Samuelsen proved the consistency of the resulting estimator: roughly, this is because the partial likelihood (2.5) is a design consistent estimator of Cox's partial likelihood (2.2) which in turn yields an estimating equation for a model consistent estimator for β. The normality of the estimator was left as a conjecture due to the complexity of the sampling scheme. To allow baseline hazard functions to differ across subgroups, I propose the use of the following partial likelihood where $S(i)$ is the index set of the subjects that share a common baseline hazard function with subject i:

$$L_{IPW}^S(\beta) = \prod_{i \in S} \left[\frac{e^{X_i(Y_i)\beta}}{\sum_{j \in \tilde{R}_i \cap S(i)} w_j e^{X_j(Y_i)\beta}} \right]^{w_i \cdot \delta_i}. \tag{2.6}$$

Kim (2013) computed the inclusion probabilities in a nested case-control study that account for additional matching factors and ties in failure times. The probability for subject i was:

$$p_i = \begin{cases} 1, & \text{if } \delta_i = 1, \\ 1 - \prod_{j: a_i < Y_j < Y_i} \left(1 - \min\left(1, \frac{m b_{ji}}{k_{ji} - b_{ji}}\right)\right), & \text{if } \delta_i = 0, \end{cases} \tag{2.7}$$

where k_{ji} is the size of $R_j \cap H_i$ where H_i is the set of subjects in the full cohort with the same values of matching variables as subject i. In other words, k_{ji} is the number of subjects at risk at Y_j with the same values of the matching variables as subject i; b_{ji} is the number of tied subjects in H_i that failed exactly at Y_j. In the absence of additional matching variables or ties in failure times, the inclusion probability simplifies to that of Samuelsen (1997). The calculation of minimum is for the late failure times when all subjects in $R_j \cap H_i$ are sampled because $k_{ji} - b_{ji} < m\, b_{ji}$.

Notice that we used two different subgroup notations $S(i)$ in (2.3) and $H(i)$ in (2.7) because the stratifications at design and analysis stages need not be the same. While the disproportionate representation caused by matching is accounted by the inclusion probabilities (2.7), the stratified analysis is used strictly to allow varying baseline hazard functions. One may choose less number of different baseline hazard functions than the number of matching subgroups in order to increase statistical power. The decision is strictly a modeling issue that does not need to be determined in the study design stage.

3. Standard Error Estimation

Samuelsen (1997) and Chen (2001) both derived asymptotic variances for β, but these cannot be computed using commonly available statistical software and require computational memory in the order of $O(n^2)$ to compute all pair-wise co-inclusion probabilities. Kim (2013) recently proposed an approximate jackknife variance estimator for secondary outcome analysis that can be easily computed with existing software:

$$\mathrm{Cov}\left(\hat{\beta}\right) = I_w^{-1} S_w^T \Lambda_w^2 S_w I_w^{-1}|_{\beta=\hat{\beta}}. \tag{3.1}$$

Λ_w is the diagonal matrix of weights, $S_w = (S_{ij}^w(\hat{\beta}))$ where $S_{ij}^w(\beta)$ is the typical score residual for Cox's partial likelihood if w_i were the frequency weight, and I_w is the negative Jacobian. We will use the estimator for the primary outcome analysis. The approximate jackknife standard error has been used for full cohort analysis by Reid and Crepeau (1985), and Lin and Wei (1989), and for case-cohort analysis by Barlow (1994). Kim (2013) provided simple syntax of codes in R and SAS software.

4. Simulation Studies

Exponential failure times T with the rate of $\exp(\beta_1 X_1 + \beta_2 X_2)$ were generated for full cohorts of size $N = 500$, 1,000, or 2,000. X_1, the main exposure variable of interest, was assumed to be distributed as standard normal variable, and X_2 was specified as a Bernoulli variable with a probability of success set at $(1 + \exp(X_1))^{-1}$. The distribution of the covariates was set up so that a mild multicollinearity existed. The true log hazard ratio of interest, β_1, assumed the values of $0, 0.1, 0.2, \ldots, 1.0$; therefore, the hazard ratios ranged between 1 and 3.86 by an increase equivalent to the inter-quartile range of the standard normal variable. The log hazard ratio for X_2 was set as $\beta_2 = 0.5$. Censoring times were uniformly distributed between 0 and a. The upper limit of the censoring, a, was chosen so that the proportion of failure was 15% in the full cohort. For each subject, either the failure or censoring time was observed, whichever occurred earlier. The log hazard ratios and their standard errors in the full cohort were estimated under the Cox proportional hazards model (Cox, 1972).

I then selected nested case-control samples with varying numbers of controls, $m = 1, 2$, or 5, at each failure time. For each simulated nested case-control data set, the log hazard ratios and their standard errors are estimated by three methods: a naïve Cox method (Cox, 1972) treating the nested case-control data as if they were a full cohort, the conditional logistic approach of Thomas (1977),

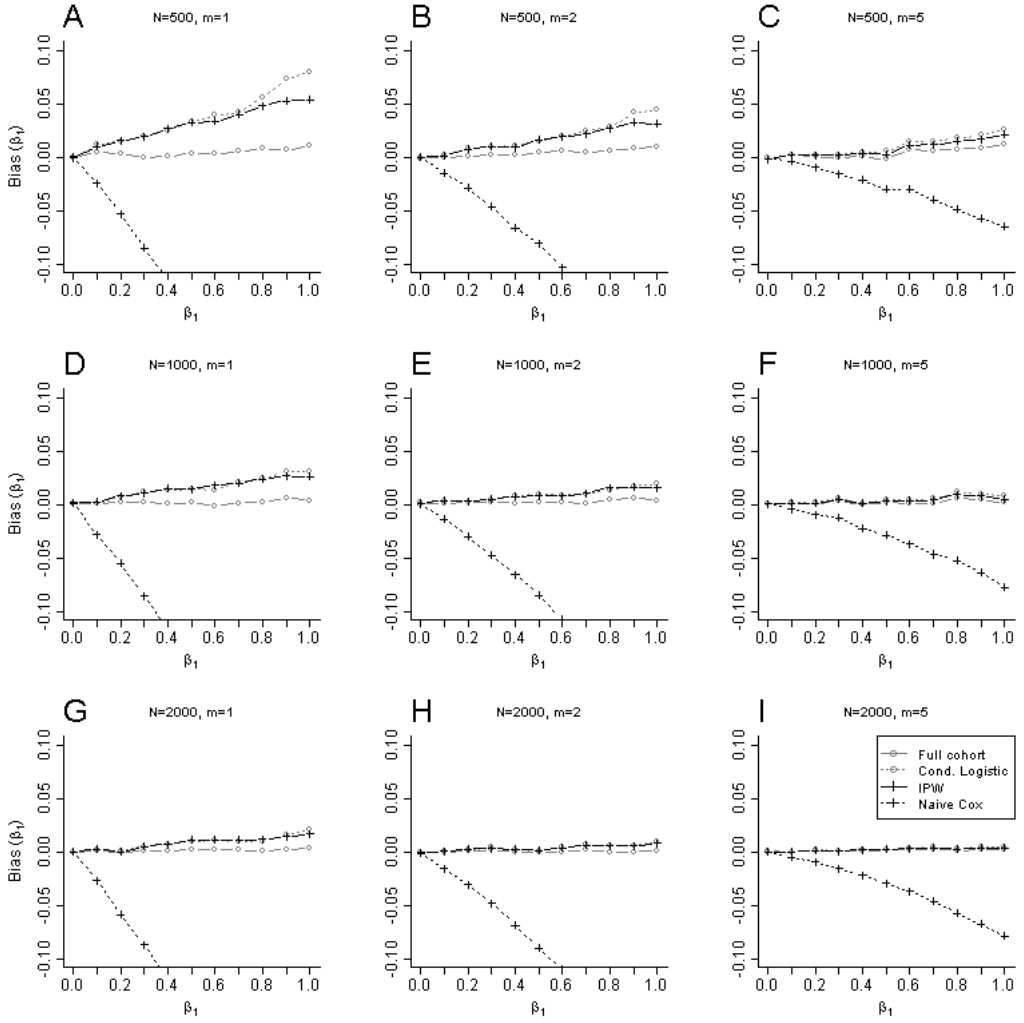

Figure 1: *The Empirical Biases in the Simulation Study (The empirical biases of the full cohort estimator and the three estimators based on the nested case-control samples - a naïve Cox estimator, the conditional logistic estimator, and the inverse probability weighted (IPW) estimator.)*

and the inverse probability method proposed by Samuelsen (1997). When I used the method by Samuelsen, both the approximate jackknife standard error proposed by Kim (2013) and the original standard error proposed by Samuelsen (1997) were calculated. For simplicity, additional matching factors were not used in the simulation studies. This overall process including the generation of the full cohort, the nested case-control sample, and the recording of estimates was repeated 5,000 times. The standard error proposed by Samuelsen (1997) was computed with only 500 simulated data sets due to prohibitive computing time.

Figure 1 shows the empirical biases of the full cohort estimator and of the three aforementioned estimators based on the nested case-control samples. As expected, the naïve estimates were severely biased since they did not adjust for the disproportionate representation in the nested case-control

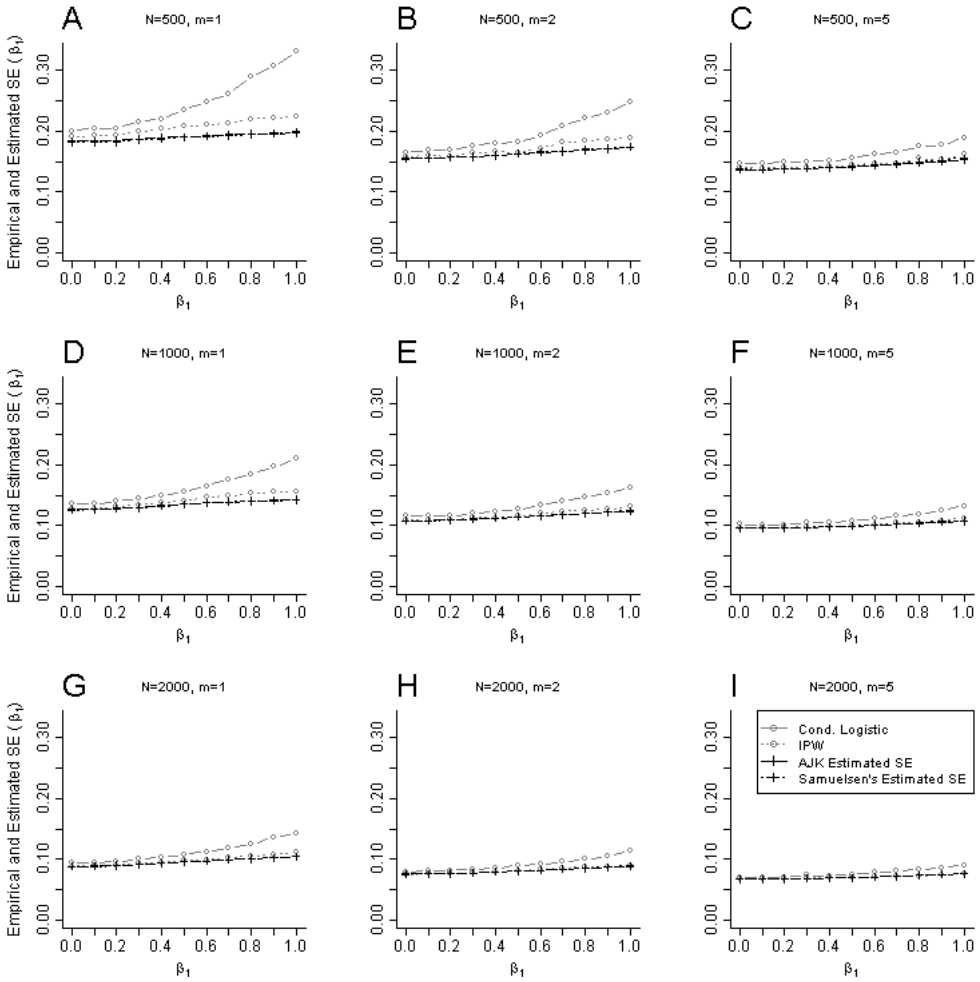

Figure 2: *The Empirical and the Estimated Standard errors in the Simulation Study (The empirical standard errors of the conditional logistic estimator and the inverse probability weighted estimator are shown. In addition, the average standard error estimates by the approximate jackknife (AJK) method and by Samuelsen (1997)'s method are shown.)*

designs. The inverse probability weighting method estimator by Samuelsen showed noticeably less bias than the conditional logistic estimator of Thomas when $N = 500$ and the true hazard ratio was large. Most importantly, both the inverse probability weighting method and the conditional logistic approach yielded estimates with low empirical biases ($< 5\%$ from the true of log hazard ratio except when $N = 500$ and $m = 1$).

Figure 2 shows that the empirical standard error of the inverse probability weighted method estimator by Samuelsen is less than that of the conditional logistic estimator by Thomas especially when the hazard ratio is large. In addition, it shows the average approximate jackknife standard errors which accurately estimate the empirical standard errors of the inverse probability weighted estimator and were remarkably close to the standard errors of Samuelsen (1997). The standard errors were

Table 1: The empirical type 1 errors in the simulation study

Methods	N = 500			N = 1,000			N = 2,000		
	m = 1	m = 2	m = 5	m = 1	m = 2	m = 5	m = 1	m = 2	m = 5
Thomas(1977)	0.043	0.046	0.052	0.043	0.055	0.051	0.047	0.047	0.051
IPW & Samuelsen (1997)	0.052	0.058	0.046	0.044	0.048	0.056	0.046	0.030	0.046
IPW & Approx Jackknife	0.052	0.055	0.055	0.048	0.054	0.054	0.054	0.048	0.050

Controlling the nominal type 1 error rate at 0.05, and setting $\beta_1 = 0$, the empirical type 1 error rates to test the null hypothesis $H_0 : \beta_1 = 0$ were measured by three methods: the conditional logistic approach of Thomas (1977), and the inverse probability weighting (IPW) method by Samuelsen (1997) in combination with either the standard error he originally proposed or the approximate jackknife standard error.

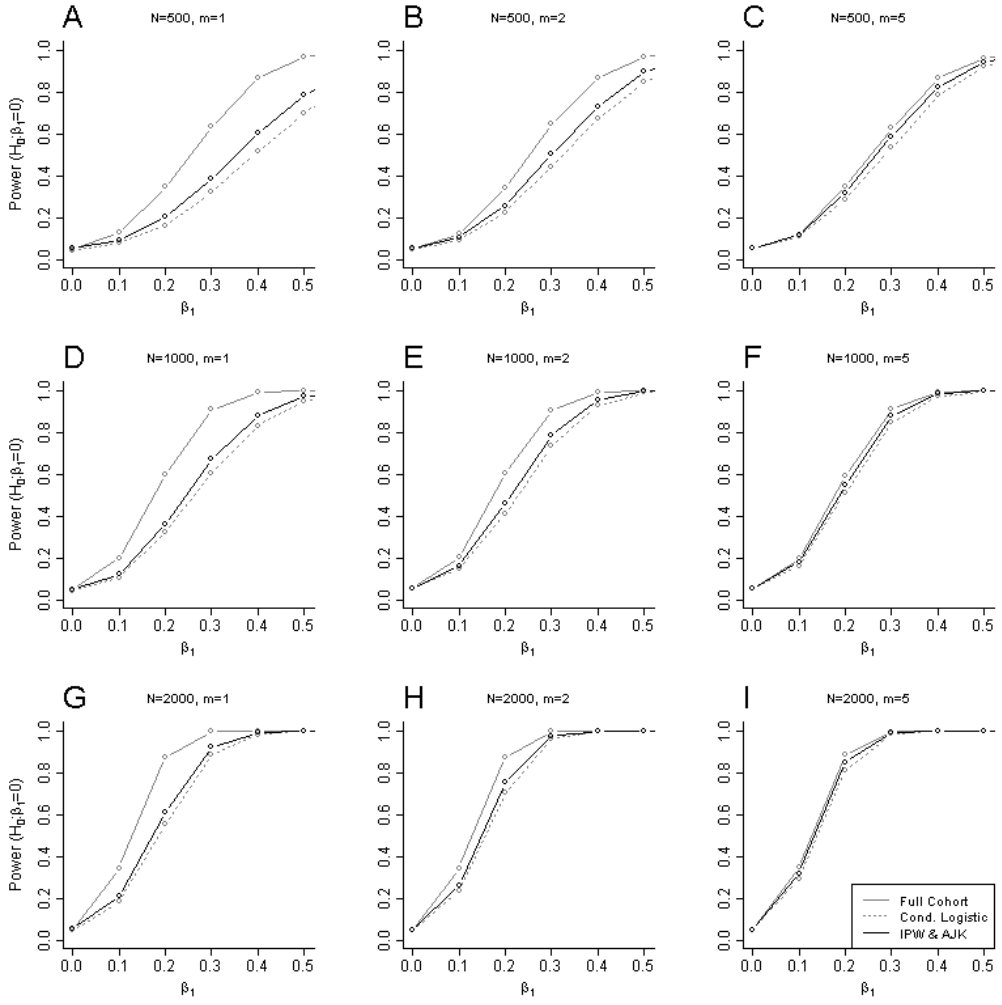

Figure 3: *The Empirical Power in the Simulation Study (Controlling the nominal type 1 error rate at 0.05, the empirical power by three methods to test the null hypothesis $H_0 : \beta_1 = 0$ is measured: Cox analysis in the full cohort, the conditional logistic approach, and the inverse probability weighting method in combination with an approximate jackknife (AJK) standard error).*

Table 2: The numbers of children in matching subgroups from the Wilm's Tumor study

		Tumor stage			
		I	II	III	IV
Age	Infant	215	272	291	167
	1 to 2	331	332	304	171
	3 to 4	640	358	290	108
	5 to 15	386	90	59	14

mildly under-estimated when the sample size was small ($N = 500$ and $m = 1$).

Controlling the nominal type 1 error rate at 0.05, and setting $\beta_1 = 0$, all methods showed valid empirical type 1 error rates to test the null hypothesis, $H_0 : \beta_1 = 0$ (Table 1); however, the empirical power of the inclusion probability weighting methods to test the null hypothesis, $H_0 : \beta_1 = 0$, was consistently higher than that of the conditional logistic approach (Figure 3). The difference was noticeable when the sample size was moderate. For example, when $N = 500$, $m = 1$, and $\beta = 0.5$, the empirical power of the inverse probability weighting method in combination with an approximate jackknife standard error was 0.79 and while that for the conditional logistic approach of Thomas was 0.70. When the sample size was sufficient, the differences between two methods were negligible.

5. Example

I analyzed the full cohort data of the relapse or censored times of 4,028 children with Wilms' tumor from the third and fourth clinical trials of the National Wilms Tumor Study Group (Breslow and Chatterjee, 1999; D'Angio *et al.*, 1989; Green *et al.*, 1998) to further compare the performances of the inverse probability weighting method to the conditional logistic approach of Thomas in the presence of additional matching. Among the 4,028 children in the full cohort, 571 had relapsed. The main predictor is a binary 'histology outcome of tumors (x_1)' from a central-lab which is either 'favorable' or 'unfavorable'. Patients with tumors composed of one of the rare cell types known collectively as 'unfavorable histology' are more likely to relapse and die than patients with tumors of 'favorable histology' (Beckwith and Palmer, 1978; Breslow and Chatterjee, 1999). Two additional variables considered important to control in the analyses were the tumor stage (x_2) and the age at baseline (x_3). I considered two analytic models: an additive proportional hazards model and a stratified proportional hazards model. For the additive model, I included all three predictors as additive covariates. The estimate of the log hazard ratio β_1 was 1.593 (SE = 0.089). For the stratified model, I used x_1 as the only covariate, and x_2, x_3 as the stratification factors allowing different shapes of baseline hazards. The estimate of the log hazard ratio β_1^S was 1.499 (SE = 0.092). Interactions were not considered for the sake of simplicity.

A total of 5,000 nested case-control samples were repeatedly selected from the full cohort with varying numbers of controls ($m = 1, 2, 5$) matched on tumor stage (x_2) and age (x_3) following the typical practices of controlling for confounding factors with matching. Table 2 shows the number of subjects across different matching subgroups. I used the aforementioned inclusion probabilities (2.7) to account for matching. First, for the additive model, I compared for each sample the full cohort estimate of β_1 with the estimates from naïve Cox regression with additive covariates x_1, x_2, and x_3, Thomas' conditional logistic method with the only covariate x_1, and the inverse probability weighting method based on the partial likelihood (2.5) with additive covariates x_1, x_2, and x_3. Similarly, for the stratified model, I compared the full cohort estimate of β_1^S with the estimates from naïve unweighted stratified Cox regression with covariate x_1 and stratification factors x_2 and x_3, Thomas' conditional logistic method with the only covariate x_1, and the inverse probability weighting method based on the

Table 3: The empirical biases in the Wilm's Tumor study

β_1^S	1	2	5	β_1	1	2	5
Full (stratified)		0		Full (additive)		0	
Naïve Cox (stratified)	−0.448	−0.272	−0.103	Naïve Cox (additive)	−0.480	−0.296	−0.118
Conditional Logistic	0.033	0.030	0.025	Conditional Logistic	−0.061	−0.064	−0.069
IPW & Approx Jackknife (stratified)	0.049	0.019	0.004	IPW & Approx Jackknife (stratified)	0.023	0.008	0.001

The empirical bias averaged over 5,000 iterations is shown in each cell. The bias is defined by the difference between the estimates from the nested case-control sample and the full cohort estimate. For the stratified model, I used histology (x_1) as the only covariate, and tumor stage (x_2), and the age at baseline (x_3) as the stratification factors to allow different baseline hazard functions. The full cohort estimate of the log hazard ratio β_1^S was 1.499 (SE = 0.092). For the additive model, all three predictors were included as the additive covariates. The full cohort estimate of the log hazard ratio β_1 was 1.593 (SE = 0.089).

Table 4: The empirical and the estimated standard errors in the Wilm's Tumor study

β_1^S	1	2	5	β_1	1	2	5
Conditional Logistic (Thomas)	0.186 (0.185)	0.144 (0.137)	0.116 (0.096)	Conditional Logistic (Thomas)	0.184 (0.185)	0.142 (0.137)	0.113 (0.096)
IPW & Approx Jackknife (stratified)	0.157 (0.151)	0.124 (0.123)	0.102 (0.102)	IPW & Approx Jackknife (stratified)	0.102 (0.102)	0.121 (0.121)	0.099 (0.100)

In each cell, the empirical standard error is first shown followed by the estimated standard error averaged over 5,000 simulated data sets in parentheses. To capture both infinite and finite sampling variation, the empirical variance is defined as the sum of the empirical variance due to the simulation and the estimated variances of the full cohort estimates. The approximate jackknife standard errors were computed for the inverse probability weighting methods.

stratified partial likelihood (2.6) with covariate x_1 and stratification factors x_2 and x_3.

The empirical bias for the log hazard ratio was calculated as the average difference between the full cohort estimate and the nested case-control estimates. Both the inverse probability weighting method and the conditional logistic approach of Thomas yielded low empirical biases (< 5% from the full cohort estimate; Table 3). This is consistent with the results from the earlier simulations studies. Still, the inclusion probability weighting method showed the lowest bias. In comparison, the average biases of the naïve unweighted Cox estimators were from 9 to100 times greater than those of the inverse probability weighting methods.

Table 4 shows the empirical standard errors. The empirical variance due to both infinite and finite sampling was computed by adding the estimated variance for the full cohort estimates to the empirical variance resulting from the simulation. The inverse probability weighting method again resulted in less empirical standard error than the conditional logistic approach of Thomas in all settings. In addition, the approximate jackknife standard errors showed very good approximation of the empirical standard errors of the inverse probability weighting method estimators.

6. Discussion

The inclusion probability methods are more efficient than the conditional logistic approach of Thomas; however, the epidemiology research community has not accepted the methods as replacements of the Thomas' method. To promote the use of the inverse probability weighted estimators, I demonstrated that the estimators have greater power than the conditional logistic approach consistently across various sample sizes and magnitudes of the hazard ratios. I then demonstrated that the easily computable approximate jackknife standard errors originally proposed for secondary outcome analysis in nested case-control studies (2013) showed good approximation of the empirical standard errors in the pri-

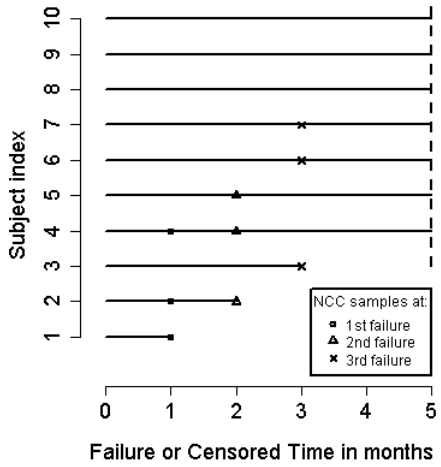

Figure 4: *An example of a nested case-control study design (This is an example of a nested case-control design ($m = 2$) from a small cohort of ten subjects. For example, three subjects ($i = 1, 2, 3$) failed with no ties before the remaining seven subjects ($i = 4, \ldots, 10$) were censored. The subjects 2 and 4 were selected as the controls from the risk set at 1 month. Overall all three failures ($i = 1, 2, 3$) and four non-failures ($i = 4, 5, 6, 7$) were included in the sample.)*

mary outcome analysis. And they were remarkably close to the standard errors by Samuelsen (1997). Finally, I made a modest generalization to the partial likelihood and inclusion probabilities in order to incorporate additional matching and the stratified Cox analysis.

Consider a small cohort of ten subjects and one covariate, x, of interest. Three subjects ($i = 1, 2, 3$) failed with no ties before the remaining seven subjects ($i = 4, \ldots, 10$) were censored (Figure 4). In a nested case-control design with two controls for each case (*i.e.*, $m = 2$), the inclusion probabilities of the subjects were $p_1 = p_2 = p_3 = 1$, $p_4 = \cdots = p_{10} = 1 - (1 - 2/9)(1 - 2/8)(1 - 2/7) = 0.58$. Overall all three failures ($i = 1, 2, 3$) and four non-failures ($i = 4, 5, 6, 7$) were included in the sample. Consider the contribution of the subjects at 1 month. The figure demonstrates the nested case-control sample with subjects 2 and 4 selected as controls from the risk set at 1 month. The denominator of Cox's partial likelihood (2.2) is

$$\sum_{i=1}^{10} e^{\beta x_i} = e^{\beta x_1} + \sum_{i=2}^{10} e^{\beta x_i}. \tag{6.1}$$

To approximate this, the partial likelihood of Samuelsen uses

$$e^{\beta x_1} + \left(\sum_{i=2}^{3} e^{\beta x_i} + \frac{1}{0.58} \sum_{i=4}^{7} e^{\beta x_i} \right). \tag{6.2}$$

The conditional logistic partial likelihood of Thomas uses

$$\frac{9}{2} e^{\beta x_1} + \frac{9}{2} \left(e^{\beta x_2} + e^{\beta x_4} \right). \tag{6.3}$$

Notice that the constant 9/2 does not change the resulting estimator. The second terms in (6.2) and (6.3) are both consistent estimators of the second term in (6.1) with respect to the finite sampling

distribution of nested case-control samples. This is why the partial likelihoods of both methods are design consistent estimators of Cox's partial likelihood. The standard error of (6.3) suffers because the covariate contribution is estimated from only 3 subjects while all 7 available subjects are used in (6.2).

We may view a nested case-control design as an unequal probability sampling from a full cohort. This explains why the same form of partial-likelihood (2.5) has been used in complex survey designs (Binder, 1992; Lin, 2000). However there are differences between a nested case-control study and a complex survey. In the complex survey, a sample is drawn from each disjoint stratum. On the contrary, in nested case-control studies, m subjects are repeatedly selected from overlapping risk sets. Further, unlike in typical survey settings, nested case-control studies almost always need to take both finite and infinite sampling variations into consideration because full cohorts are often only moderate in sizes and they themselves are considered a sample from target populations.

The statistical power of the inverse probability weighting methods decreases as more strata with varying baseline hazards functions are allowed in the analysis stage. If matching is fully exhaustive and the number of failures equals the number of subgroups in the full cohort for matching, and if one decides to assume different hazard functions for all matching subgroups, the method should result in similar power as that of the conditional logistic approach. However, matching factors in the design stage and stratification factors in the analysis stage need not be the same since over-stratification in the analysis stage can unnecessarily decrease power. In addition, if the over-matching causes the inclusion probability of some subjects to be zero, the study by definition does not represent the full-cohort any more.

From the simulation study, we observed a moderate under-estimation of the standard errors of the inverse probability weighted estimators by both the approximate jackknife method and Samuelsen's method when the sample size was small ($N = 500$, $m = 1$; Figure 2). However, the type 1 errors were still acceptable (< 0.06; Table 1). Further research to improve the small sample property will be helpful.

Local averaging of covariates to adjust for weights will likely further improve the performance (Chen, 2001). The performance would depend on the correlation between failure times and covariates as well as the parameters for local averaging.

The inverse probability weighting methods break the matching. Any matching variable that could affect the hazard ratio of interest should be controlled for by including them in the regression model. Finally, notice that the inverse probability weighting methods used in this paper requires retrospective access to the censoring times of all non-failures in the full cohort while Thomas (1977) and Chen (2004) do not.

Acknowledgements

The author thanks Professors Mimi Kim and Xiaonan Xue for their encouragement and comments on this work.

References

Barlow, W. E. (1994). Robust variance estimation for the case-cohort design, *Biometrics*, **50**, 1064–1072.

Beckwith, J. B. and Palmer, N. F. (1978). Histopathology and prognosis of Wilms tumor, *Cancer*, **41**, 1937–1948.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data, *Biometrika*, **79**, 139–147.

Borgan, O. and Langholz, B. (1993). Nonparametric estimation of relative mortality from nested case-control studies, *Biometrics*, **49**.

Breslow, N. and Chatterjee, N. (1999). Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis, *Applied Statistics*, **48**.

Chen, K. N. (2001). Generalized case-cohort sampling, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**.

Chen, K. N. (2004). Statistical estimation in the proportional hazards model with risk set sampling, *Annals of Statistics*, **32**, 1513–1532.

Cox, D. R. (1972). Regression models and life tables, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **34**, 187–220.

D'angio, G. J., Breslow, N., Beckwith, B., Evans, A., Baum, E., Delorimier, A., Fernbach, D., Hrabovsky, E., Jones, B., Kelalis, P., Othersen, B., Tefft, M. and Thomas, P. R. M. (1989). Treatment of Wilms' tumor, *Cancer*, **64**, 349–360.

Green, D. M., Breslow, N. E., Beckwith, J. B., Finklestein, J. Z., Grundy, P. G., Thomas, P. R. M., Kim, T., Shochat, S., Haase, G. M., Ritchey, M. L., Kelalis, P. P. and D'angio, G. J. (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.

Kim, R. S. (2013). *Analysis of Secondary Outcomes in Nested Case-Control Study Designs*, Technical Reports. Division of Biostatistics, Albert Einstein College of Medicine.

Langholz, B. (2010). Case-control studies = Odds ratios: Blame the retrospective model, *Epidemiology*, **21**, 10–12.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data, *Biometrika*, **87**, 37–47.

Lin, D. Y. and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model, *Journal of the American Statistical Association*, **84**.

Reid, N. and Crepeau, H. (1985). Influence function for proportional hazards regression, *Biometrika*, **72**, 1–9.

Samuelsen, S. (1997). A pseudo-likelihood approach to analysis of nested case-control studies, *Biometrika*, **84**, 379–394.

Thomas, D. (1977). Addendum to 'Methods of cohort analysis: Appraisal by application to asbestos mining' by Liddell FDK, McDonald JC, Thomas DC, *Journal of the Royal Statistical Society*, **A140**, 469–491.

Received July 26, 2013; Revised October 1, 2013; Accepted October 24, 2013