University of Cincinnati

Date: 5/13/2011

I, Stephen J Bertke , hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Mathematical Sciences.

It is entitled:

A Simulation Study of the Cox Proportional Hazards Model and the Nested Case-Control Study Design

Student's name: **Stephen J Bertke**

This work and its defense approved by:

Committee chair: James Deddens, PhD

Committee member: Misty Hein, PhD

Committee member: Mary Schubauer-Berigan, PhD

Committee member: Paul Horn, PhD

Committee member: Siva Sivaganesan, PhD

Committee member: Seongho Song, PhD

1726

A Simulation Study of the Cox Proportional Hazards Model and the Nested Case-Control Study Design

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctorate of Philosophy

in the Department of Mathematical Sciences

of the College of Arts and Sciences

May 2011

by

Stephen Joseph Bertke

B.A. University of Cincinnati, June 2006

M.S. University of Cincinnati, June 2009

Committee Chair: J. A. Deddens, Ph.D.

Abstract

The Cox proportional hazards model is commonly used to analyze the exposure-response relationship in occupational cohort studies. This analysis involves identifying cases (those who experience the outcome of interest) and forming risk-sets for each case. The risk-set for a case is the set of cohort members whose failure times are at least as large as the case's failure time and are under observation immediately before the case's failure time. Thomas proposed the idea of randomly sampling controls from each risk-set to use for analysis, which results in a nested case-control study. It has been shown that the analysis using the full risk-sets and the analysis using the sampled risk-sets produce asymptotically unbiased results. Also, the asymptotic relative efficiency between analyzing the full risk-sets and using Thomas' estimator to analyze the sampled risk-sets (sampling m controls per case) is $\frac{m}{m+1}$ when there is no exposure-response relationship.

A simulation study investigated the non-asymptotic properties of the nested case-control study design and found that the relative efficiency decreased as the number of cases in the cohort decreased, the true exposure-response parameter increased, and the skewness of the exposure distribution of the risk-sets increased. There also appeared to be some bias in a nested case-control study and this bias tended to be away from the null, however, this was not a major issue. In fact, when 10 or more controls were matched with each case, the bias was never more than 10%.

A second simulation study compared the estimates obtained from a nested case-control analysis for a given cohort to the estimate obtained from analyzing the full cohort

with Cox proportional hazards regression. The nested case-control estimate generally overestimated the full cohort estimate and the size of this discrepancy varied from cohort to cohort. Also, the sample variance of the estimates from a nested case-control study for a given cohort decreased dramatically as the case: control ratio increased.

An alternative estimator for a nested case-control study was proposed by Chen and a set of simulations compared the performance of this estimator to that of the traditional Thomas estimator. Chen's estimator requires the user to define a function, $\varphi(t)$. The support of $\varphi(t)$, defines which controls It was shown that the performance of Chen's estimator is somewhat sensitive to the definition of $\varphi(t)$. In particular, if the support of $\varphi(t)$ was small, Chen's estimator performed poorly. However, for larger definitions of the support of $\varphi(t)$, Chen's estimator performed comparable, if not better than, Thomas' estimator in terms of the bias and relative efficiency.

Finally, a simulation study investigated the effect of classical measurement error on the Cox proportional hazards model. The simulations suggest that the introduction of measurement error may change the perceived shape of the exposure-response curve. In fact, the curve was more likely to level-off in the high exposure range which is commonly seen in occupational cohort studies and this effect became more severe as the magnitude of the error increased.

Acknowledgements

I would like to gratefully and sincerely thank my advisor, Dr. James A. Deddens, for his expertise, guidance, and patience. He has been abundantly helpful and extremely supportive in numerous ways through the years. I am also indebted to Dr. Mary Schubauer-Berigan and Dr. Misty Hein, who have acted as mentors during my doctoral research as well as during my professional development at the National Institute for Occupational Safety and Health. I am also grateful to Dr. Paul S. Horn, Dr. Siva Sivaganesan, and Dr. Seongho Song for their valuable input to this dissertation and holding me to a high research standard.

I would like to thank the Department of Mathematical Sciences for providing financial support as well as a superior, well-rounded educational experience. Also, I would like to thank the National Institute for Occupational Safety and Health, in particular my supervisors Cherie Estill and Elizabeth Whelan, for supporting me during my doctoral study. Most importantly, none of this would have been possible without the endless support from my parents Stephen and Patricia Bertke. Their encouragements are constantly inspiring me and pushing me to the best I can possibly be. It is to them that I dedicate this work.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Cohort Studies

A cohort study follows a defined group of individuals over time to study the effect of predictive factors on the occurrence of a particular outcome. In an occupational cohort study, the defined group of individuals is generally those who worked in a factory/plant dealing with a particular exposure of interest during a defined period of time. Further restrictions on the cohort definition may be made, such as selecting only those with a minimum duration of employment. The individuals are then considered at risk until the outcome of interest occurs (such as death from lung cancer) or until the observation is censored. Censoring may occur for various reasons, such as death from another disease, loss to follow-up, or survival until the end of study (Breslow and Day, 1987).

There are several methods used to analyze cohort studies, however, this paper will focus on the Cox proportional hazards model and the nested case-control study design.

1.2 Analysis of Cohort Study Data:

Cox Proportional Hazards:

Often, researchers are interested in evaluating the effect of various covariates on the survival time (i.e. the amount of time until an event occurs) of an individual. To describe the distribution of the survival time, *T*, the common functions are:

Cumulative Density Function: $F(t) = P[T \leq t]$

Survivor Function: $S(t) = P[T > t] = 1 - F(t)$

Hazard Function: $$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t} \quad (1.1)$$

If a specific distribution is assumed for *T*, standard maximum likelihood procedures may be used to estimate unknown parameters of the model. In particular, suppose there *n* observation in a cohort with *k* observed events (and therefore *n – k* censored events) and suppose the survival times are assumed to follow a distribution with pdf *f(t|θ)*, where θ is the unknown parameter to be estimated. Arranging the observations so that observations 1 to *k* are the observed events and observations *k* + 1 to *n* are the censored observations, the likelihood function will have the following setup:

$$L(\theta) = \prod_{i=1}^{k} f(t_i|\theta) \prod_{i=k+1}^{n} S(t_i|\theta) \quad (1.2)$$

(Lee and Wang, 2003).

Alternatively, the Cox proportional hazards model is frequently used in occupational cohort studies to evaluate the hazard associated with a given exposure. The Cox model is desirable because it does not require knowledge about the exact underlying distribution of the survival times. In fact, the only assumption that is made is that the hazard function of an individual with prognostic factors or covariates $\boldsymbol{X} = (x_1, x_2, \ldots)'$ can be expressed in the following manner:

$$h(t|\boldsymbol{X}) = h_0(t) g(\boldsymbol{X}) \quad (1.3)$$

where $h_0(t)$ is the baseline hazard function (i.e. the hazard function when $g(\boldsymbol{X}) = 1$) and $g(\boldsymbol{X})$ is a function only of the covariates $\boldsymbol{X}$. The function $g(\boldsymbol{X})$ may only implicitly be

a function of time, *t*, if the covariates of interest are functions of *t*. Often it is assumed that $g(\boldsymbol{X})$, which is the effect of the covariates on the hazard function, is log-linear and takes the following form:

$$g(\boldsymbol{X}) = e^{\boldsymbol{\beta}'\boldsymbol{X}} = e^{\boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \cdots} \tag{1.4}$$

where $\boldsymbol{\beta}$ represent the coefficients of the covariates and are the unknown parameters to be estimated. However, other forms of $g(\boldsymbol{X})$ may be used. This topic will be discussed further in Chapter 4.

Cox (1972) proposed the use of a partial likelihood function, which is free of $h_0(t)$, to estimate the unknown parameter $\boldsymbol{\beta}$. To construct the partial likelihood function, again suppose a given data set or cohort has *n* observations with *k* observed deaths, and let $t_{(1)}, t_{(2)}, \ldots t_{(k)}$ be the observed failure times. For each observed failure time, let $\boldsymbol{R}(t_{(i)})$ be the risk-set at time $t_{(i)}$, i.e. it is the set of individuals who are under observation at time $t_{(i)}$ and whose survival times are at least $t_{(i)}$. Let $\boldsymbol{X}_1(t), \boldsymbol{X}_2(t) \ldots \boldsymbol{X}_n(t)$ be the covariates of the *n* individuals evaluated at time *t*. Then, for a given risk-set $\boldsymbol{R}(t_{(i)})$, the probability that the death occurs to the observed individual is:

$$\frac{h(t_{(i)}|\boldsymbol{X}_{(i)}(t_{(i)}))}{\sum_{j\in R(t_{(i)})} h(t_{(i)}|\boldsymbol{X}_j(t_{(i)}))} = \frac{h_0(t_{(i)})e^{\boldsymbol{\beta}'\boldsymbol{X}_{(i)}(t_{(i)})}}{\sum_{j\in R(t_{(i)})} h_0(t_{(i)})e^{\boldsymbol{\beta}'\boldsymbol{X}_j(t_{(i)})}} = \frac{e^{\boldsymbol{\beta}'\boldsymbol{X}_{(i)}(t_{(i)})}}{\sum_{j\in R(t_{(i)})} e^{\boldsymbol{\beta}'\boldsymbol{X}_j(t_{(i)})}} \tag{1.5}$$

Each risk-set contributes a factor and therefore, the partial likelihood function becomes:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{e^{\boldsymbol{\beta}'\boldsymbol{X}_{(i)}(t_{(i)})}}{\sum_{j\in \boldsymbol{R}(t_{(i)})} e^{\boldsymbol{\beta}'\boldsymbol{X}_j(t_{(i)})}} \tag{1.6}$$

Since the above equation does not specify all factors of the model, it is not a true likelihood function.

However, even though the above equation is not a true likelihood, Cox (1974) outlined the theory that the standard asymptotic properties of likelihood functions hold. Namely, the estimator is a consistent estimator and is asymptotically normal with the inverse of the Fisher information matrix as the variance-covariance matrix.

Anderson and Gill (1982) reformulated the Cox-model using a counting process to rigorously prove the asymptotic properties of the partial maximum likelihood estimator. In particular, they defined $N_i(t)$ as the number of observed events for subject i up to time t. They also assign the following random intensity process for $N_i(t)$:

$$h_i(t) = Y_i(t)h_0(t)e^{\beta' X_i(t)} \tag{1.7}$$

again where $\boldsymbol{X}_i(t)$ is the covariate for subject i at time t and $h_0(t)$ is a baseline function. $Y_i(t)$ is an indicator process taking the values 1 and 0 representing when subject i is and is not under observation. Therefore, $N_i(t)$ only increases by one when $Y_i(t) = 1$.

Under this formulation, and with certain regularity conditions, Anderson and Gill were able to show that $\hat{\beta}$, the estimator from the partial likelihood function, is:

Consistent: $\hat{\beta} \xrightarrow{p} \beta$

Asymptotically Normal: $n^{\frac{1}{2}}(\hat{\beta} - \beta) \xrightarrow{D} N(\mathbf{0}, \Sigma^{-1})$ (1.8)

where Σ is an invertible, positive definite matrix such that:

$$-\left[\frac{\partial^2}{\partial\beta_i \partial\beta_j} \log L(\boldsymbol{\beta})\right]\Big|_{\beta=\hat{\beta}} \xrightarrow{p} \Sigma \tag{1.9}$$

Nested Case-control Analysis:

Performing Cox PH regression on a given cohort requires forming risk-sets for each observed failure time and evaluating the exposure history of every individual in the

risk-set at the failure time. Therefore, if an individual appears in multiple risk-sets, his/her exposure has to be re-evaluated at each failure time. This can be quite financially and computationally expensive to carry out for every member of each risk-set, especially if the cohort is very large.

To ease the burden of this cost, Thomas (Liddell et al, 1977) proposed the idea ofsampling controls from each of the risk-sets to use in analysis. Notice that the partial likelihood for Cox regression is identical to the conditional logistic likelihood used in matched case-control studies. In a matched case-control study, individuals with an outcome of interest (called cases) are randomly selected from an infinite population. Then one or more individuals without the outcome of interest (called controls) are randomly sampled and matched with each case based on matching variables (such as age or gender).

In a cohort study, for each case, one could match *m* controls by sampling from the case's risk-set. Here, the populations are the risk-sets and the matching variable is the failure time. Therefore, this design can be thought of as a matched case-control design nested within a cohort study. Randomly sampling *m* controls without replacement from each risk-set results in the following likelihood:

$$L_m(\boldsymbol{\beta}) = \prod_{i=1}^{k} \frac{e^{\boldsymbol{\beta}' \boldsymbol{X}_{(i)}(t_{(i)})}}{\sum_{j \in \widetilde{\boldsymbol{R}}(t_{(i)})} e^{\boldsymbol{\beta}' \boldsymbol{X}_{j}(t_{(i)})}} \qquad (1.10)$$

where $\widetilde{\boldsymbol{R}}(t_{(i)})$ consists of the case plus the *m* sampled controls from risk-set $\boldsymbol{R}(t_{(i)})$.

Goldstein and Langholz (1992) showed, using Anderson and Gill's counting process formulation of the Cox model, that $\hat{\beta}_T$, the estimator from the above likelihood, is:

Consistent: $\hat{\beta}_T \xrightarrow{P} \beta$

Asymptotically Normal: $n^{\frac{1}{2}}(\hat{\beta}_T - \beta) \xrightarrow{D} N(0, \Gamma^{-1})$ (1.11)

where Γ is an invertible, positive definite matrix such that:

$$-\left[\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log L_m(\boldsymbol{\beta})\right]|_{\beta=\hat{\beta}_T} \xrightarrow{P} \Gamma. \quad (1.12)$$

Furthermore, they gave a relationship between the variance parameters of the full cohort analysis using Cox regression and the nested case-control analysis when $\beta = 0$ (i.e. a null association). In particular, they showed that $\Gamma = \frac{m}{m+1}\Sigma$ where Σ is as in equation 1.8. Therefore, the asymptotic relative efficiency of 1:*m* nested case-control sampling relative to the full cohort with one predictor variable is $\frac{\Sigma^{-1}}{\Gamma^{-1}} = \frac{m}{m+1}$ when $\beta = 0$.

Application to Occupational Cohort Studies:

Generally, in occupational cohorts, the main covariate of interest is exposure to a particular physical or chemical agent and the outcome of interest is death or occurrence of a particular disease (such as lung cancer). As a result, age is usually used as the time scale (as opposed to calendar time or time on study) since age is one of the most important risk factors for most diseases (Breslow et al., 1983). Furthermore, in a simulation study, use of time-on-study as the time scale was associated with biased results whereas use of chronological age as the time scale was not (Thiébaut and Bénichou, 2004).

To calculate the exposure history for individuals in a cohort, detailed work history records are collected. From these work histories, information about employment begin and end dates for various job titles held are obtained. Then, for each job title, an exposure assessor assigns an exposure intensity to each job based on various information. For example, in the nuclear industry, radiation badge monitoring is practiced and information from these badges can be used to assign exposure intensities to different job titles or even to individual workers of the cohort.

Once exposure is summarized for each job title/worker the choice of exposure metric is also considered. The most common exposure metrics used are duration of exposure, cumulative exposure (i.e., the product of exposure duration and exposure intensity, summed over all jobs worked), peak exposure, and average exposure (i.e., cumulative exposure divided by duration of exposure). Depending on the outcome and underlying physiological process, different metrics might be more appropriate than others (Kriebel et al., 2007).

After the appropriate exposure metric is decided upon and the analysis is conducted, risk is generally summarized in terms of the hazard ratio. For example, suppose cumulative exposure is the exposure metric of interest and the following form of the hazard function is modeled assuming a log-linear relative risk model:

$$h(t|exp) = h_0(t)e^{\beta * exp} \tag{1.13}$$

Once β is estimated, e^{β} is reported and is interpreted as the hazard ratio for a 1 unit increase in exposure. A hazard ratio of 1 (i.e. $\beta = 0$) indicates no effect of exposure on the hazard function.

1.3 Issues with Nested Case-Control Studies:

Anderson and Gill (1982) proved all the asymptotic properties of the Cox proportional hazards model and Goldstein and Langholz (1992) proved all the asymptotic properties of nested case-control analysis using conditional logistic regression. However, the non-asymptotic properties of these analyses have not been formally investigated. A few issues with the use of the nested case-control design were highlighted in a study of workers at a naval shipyard.

In a study investigating the relationship between external ionizing radiation exposure and leukemia mortality among workers at the Portsmouth Naval Shipyard (PNS), the nested case-control model was used with 1:4 matching (Kubale et al, 2005). Risk was modeled with a log-linear model as well as a linear excess relative risk model. A significant hazard ratio of 1.08 per 10 mSv (95% confidence interval = 1.01 – 1.16) was reported for the log-linear model and $\hat{\beta}$ = 0.23 per 10 mSv (95% CI = 0.03 - 0.88) was reported for the linear model. These estimates were higher than those reported in other studies. In particular, in a similar study of A-bomb survivors, β was estimated to be 0.04 per 10 mSv of exposure when a linear hazard model was analyzed (Yiin et al, 2005).

To investigate this discrepancy further, the authors considered a subcohort of 13,468 radiation-monitored PNS workers, which included 34 leukemia cases (Kubale et al, 2006). Detailed exposure histories were collected for these individuals, which made it possible to analyze the full subcohort as well as perform a nested case-control analysis on this subcohort. To investigate the effect of the case: control ratio on bias and

variability of the risk estimates obtained from a nested case-control study, 4, 10, 15, 20 and 50 controls were randomly sampled from each risk-set (with age as the time scale) and matched with each case. For each case to control ratio, 250 control selections were conducted and analyzed using conditional logistic regression. The distribution of results for each case to control ratio was compared to the analysis of the full cohort.

Figure 1-1 gives a histogram of the exposure summary for the full subcohort of 13,468 individuals. The data was very right-skewed, and their distribution was best approximated as hybrid-lognormal as described by Kumazawa and Numakunai (1981). The hazard ratio for the total cohort analysis with a log-linear model was 1.05 per 10 mSv (95% CI = 1.01 – 1.09). Table 1-1 gives descriptive statistics of the nested case-control analyses and Figure 1-2 plots the 1,250 estimated hazard ratios (250 for each case-control matching ratio) versus the matching ratio. The authors noted that the nested case-control analysis had a tendency to over-estimate the full cohort analysis estimate however this bias decreased as the case/control matching ratio increased.

Intuitively, the estimates from a nested case-control analysis would be higher (on average) than the estimate of the full cohort analysis when the exposure variable is very right skewed, making it unlikely to randomly sample individuals with high exposure. This raises a couple of concerns. Namely, what are the non-asymptotic properties of a nested case-control analysis and how are the results affected by the number of cases in the cohort, the distribution of the exposure parameter and the strength of the exposure-response parameter? These questions will be investigated in Chapter 2.

There have been alternative methods of analysis for nested case-control studies proposed in the literature. One such estimator was proposed by Chen (2004), and this

estimator asymptotically outperforms the traditional estimator proposed by Thomas. How Chen's estimator performs non-asymptotically and how it compares to Thomas' estimator will be investigated in Chapter 3.

Finally, classical measurement error of the exposure variable is always a concern in cohort studies. Measurement error has been shown to cause attenuation of the exposure-response parameter estimate. However, measurement error may also effect the perceived shape of the exposure-response curve. In fact, in many occupational cohort studies, the exposure-response curve tends to level off for high exposures and sometimes begins to decrease. This issue will be considered in Chapter 4.

Figures and Tables

Figure 1-1: Histogram of exposure for the 13,468 PNS workers. (graph from Kubale et al, 2006)

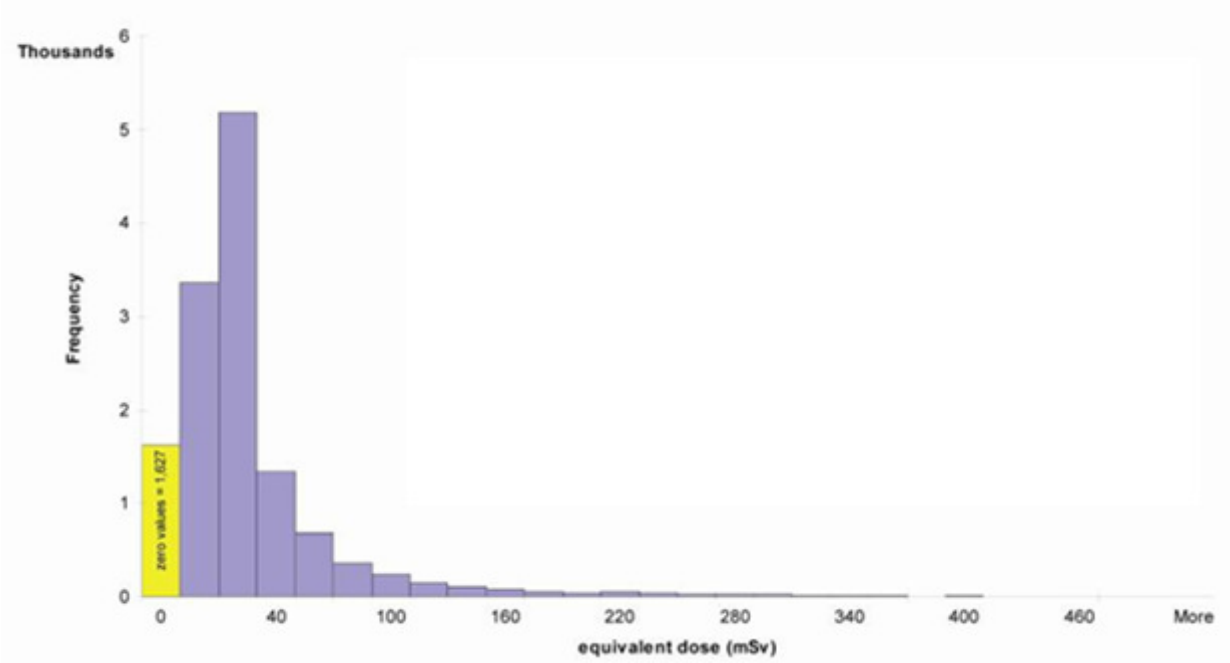
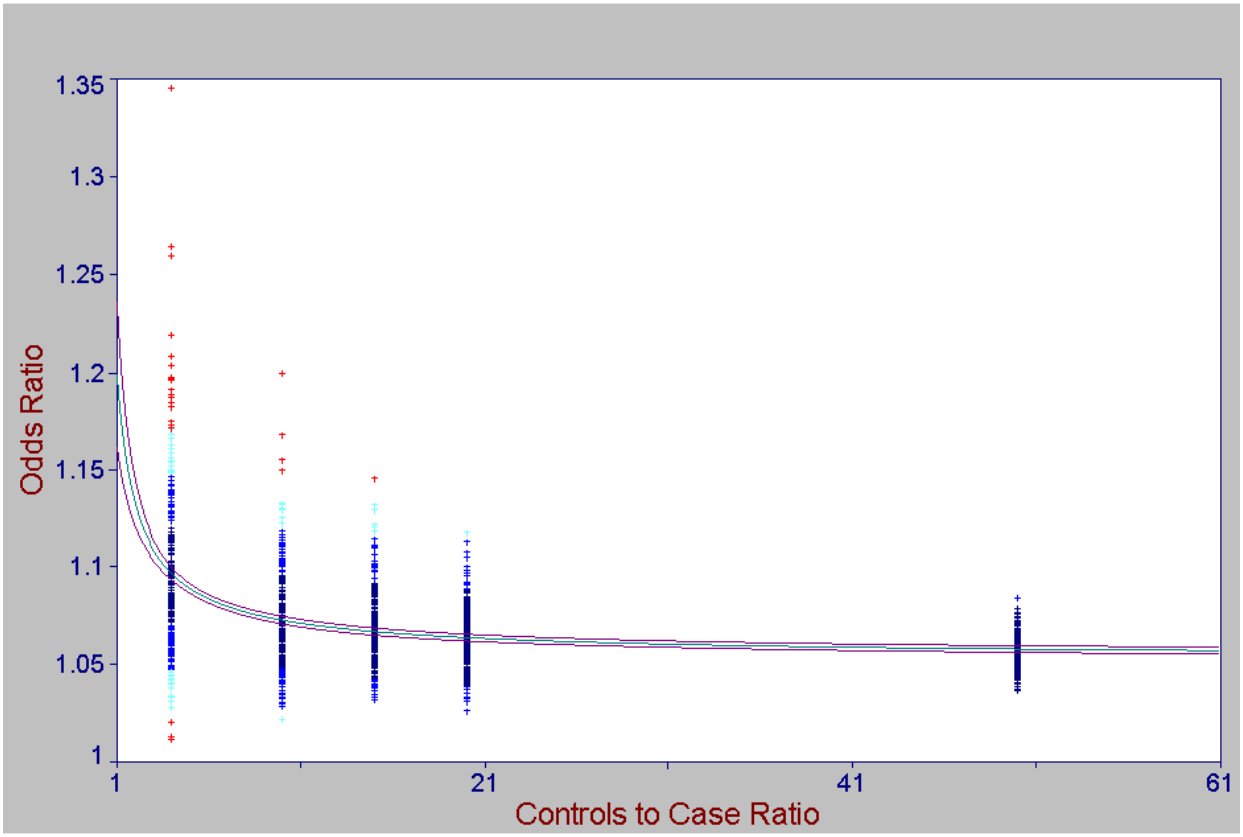


Table 1-1: Descriptive statistics from the multiple nested case-control analysis (n = 250) (table from Kubale et al, 2006)

Control to Case Ratio	Parameter Estimate Mean at 10 mSv	Odds Ratio Mean	Odds Ratio SD
4:1	0.091	1.10	0.048
10:1	0.069	1.07	0.026
15:1	0.066	1.07	0.020
20:1	0.061	1.06	0.016
50:1	0.055	1.06	0.008

Figure 1-2: Plot of hazard ratio estimate vs. case/control ratio. Items in black are within 1 standard deviation from the mean, items in blue are between 1 and 2, items in light blue are between 2 and 3 and items in red are over 3 standard deviations from the mean (graph from Kubale et al, 2006)



Chapter 2:

Relative Efficiency and Bias of Nested Case-Control Studies

Section 1

Objective:

The Cox proportional hazards model involves the analysis of the risk-sets of a given cohort. Often it is desirable to not analyze the full risk-sets but rather a subset of risk-set members for each case. This is because analysis of the full risk-sets would involve getting detailed exposure information for everyone in the cohort, which may be very costly. For example, one may randomly select five controls from each risk-set to match with each case. This study design is called the nested case-control.

Goldstein and Langholz proved that estimates from the nested case-control design are asymptotically unbiased and that when there is no exposure-response relationship (i.e. when the true exposure parameter is 0) the asymptotic relative efficiency from sampling *m* controls compared to analyzing the full cohort is $\frac{m}{m+1}$, regardless of the distribution of the exposure variable (1992). For instance, the asymptotic relative efficiency of sampling 1 control for each case would be $\frac{1}{2}$, which means the variance of the estimate from the sample risk-set analysis is twice as big as the variance of the estimate obtained if the full cohort was analyzed.

This formula gives the asymptotic relative efficiency, i.e. it is the limit of the relative efficiency as the sample size increases to infinity. Also, it only applies when there is no exposure-response relationship. There is very little in the literature that

addresses how the strength of the exposure-response relationship affects the relative efficiency (i.e. when $\beta \neq 0$).

Occupational cohorts were simulated to get a better understanding of the following questions:

1. How is the relative efficiency and bias affected by the number of cases in the cohort?
2. How is the relative efficiency and bias affected by the magnitude of the exposure-response relationship?
3. How is the relative efficiency and bias affected by the distribution of the exposure variable?

Method:

Simulations were conducted using SAS Software (version 9.1.3, SAS Institute Inc., Cary, NC). Cohorts were simulated based on a method developed by Richardson and Loomis (Richardson and Loomis, 2004) and further used in a simulation performed by Hein et al (2009). Thirty-six simulation scenarios were performed defined by the number of cases in the cohort (~30, ~100, or ~300 cases), the exposure-response relationship (hazard ratio per unit exposure = 1, 1.005, 1.01, or 1.015), and the distribution of the exposure intensity [Distribution 0: Normal(25, 64) - Truncated(0, 50), Distribution 1: Log - Normal(2.5, 0.5) - Truncated(0, 50), Distribution 2: Log - Normal(0.75, 1) – Truncated(0, 50)]. The distributions were chosen to study the effect of the distribution's skewness on the bias and relative efficiency. Distribution 0 is the least skewed (skewness of 0) and Distribution 2 is the most skewed. Figures 2-1 shows histograms of cumulative exposure for example cohort's risk-sets of each distribution.

For each scenario with ~30 cases, 10,000 cohorts were simulated. For the scenarios with ~100 cases, 3,000 cohorts were simulated and for the scenarios with ~300 cases, 1,000 cohorts were simulated. The number of cohorts simulated varied due to time (simulating and analyzing a cohort with ~30 cases didn't take very long compared to simulating a cohort with ~300 cases, therefore more cohorts could be simulated with ~30 cases). Also, presumably the results from the scenarios with more cases would be more stable and therefore fewer simulated cohorts would be required.

Each worker was randomly assigned values for age at first exposure (18 years plus a random exponential variable with mean 10), and maximum follow-up time (40 years minus a random exponential variable with mean 5). Each worker was assigned a maximum exposure duration of 15 years.

At each year of a workers maximum follow-up time, the workers current age and cumulative exposure (equal to the workers exposure intensity multiplied by exposure duration) was calculated. Also, at each year, a conditional probability of mortality from the outcome of interest (conditional on survival to that age), *h*, was assigned to each worker based on the workers age and cumulative exposure, *cumexp*, by the following formula:

$$h = e^{\alpha_0 + 1.5 * \ln\left(\frac{age}{55}\right) + \beta * cumexp} \quad (2.1)$$

where β is the exposure-response parameter (so that the hazard ratio per 1 unit increase in cumulative exposure is e^{β}). The variable α_0 is an intercept parameter which varied in each simulation scenario and was chosen to obtain the desired number of cases. It is not possible to completely control the number of cases in each cohort through this simulation method; rather the number of cases in each simulated cohort will

vary. The actual values for α_0 used as well as summary statistics of the resulting number of cases in each scenario can be found in Table 2-1 and Table 2-2, respectively. Additionally, at each year, a conditional probability of mortality from any another outcome (conditional on survival to that age), *c*, was assigned to each worker based only on the worker's age by the following formula:

$$c = e^{-5+\,5*\ln\left(\frac{age}{55}\right)}. \qquad (2.2)$$

Specific parameters for these conditional probabilities (hazard rates) were used by Richardson and Loomis (2004) as well as Hein et al (2009).

Two Bernoulli random variables were assigned to each worker at each year, one with probability *h* and one with probability *c*. A Bernoulli random variable of 1 represents a death in that year from the outcome of interest or from another outcome, respectively. A worker is followed up until his first death. A worker is considered censored if his first death is from another outcome or if he made it through all years of his maximum follow-up time with no deaths, in which case he is considered lost to follow-up. A worker is considered a case if his first death is from the outcome of interest. The final cohort consisted of 5,000 workers with variables indicating for each worker the age at first exposure, age at death/censor, age at last exposure (which is the minimum of age at first exposure plus 15 and age at death/censor), exposure intensity, and case-status.

Analysis:

Risk-sets were created for each cohort, with age as the time scale. For each case, 1, 5, 10, 15, and 20 controls were randomly sampled from the risk-sets. The full as well as the sampled risk-sets were analyzed using conditional logistic regression

(procedure PHREG in SAS) to obtain estimates of the exposure-response parameter using a log-linear model. For each scenario, 10,000, 3,000 and 1,000 estimates of the exposure-response parameter were obtained for the analysis of full risk-sets and for each of the sampled risk-sets from the cohorts with ~30, ~100, and ~300 cases, respectively. The sample variance of these estimates was obtained. The relative efficiency of 1:*m* sampling was estimated by dividing the sample variance obtained from the full risk-set analysis by the sample variance obtained from the *m*-sampled risk-set analysis. The bias was estimated by subtracting the true exposure-response parameter (i.e. the log of the true hazard ratio) from the mean of the estimated parameters and is reported as a percentage of the true parameter estimate.

Results and Discussion:

For each scenario, Tables 2-3 contains summary statistics of cumulative exposure for cases and controls of the risk-sets. As expected, the cases have higher cumulative exposures than controls when there is a positive exposure-response relationship. The tables also indicate that the distributions of cumulative exposure for the full risk-sets are similar for different values of the hazard ratio for Distribution 0 and 1 (as can be seen by comparing the summary statistics for the controls of each hazard ratio since these statistics will be similar to the summary statistics for the full risk-sets). However, for Distribution 2, the resulting distributions of cumulative exposure of the risk-sets for hazard ratio 1.015 are different compared to the other hazard ratios with Distribution 2. In particular, the max and variance of the distributions are smaller. Consequently, differences in the results from comparing the scenario with hazard ratio

1.015 with other hazard ratios for Distribution 2 may be due to either a change in the distribution or to the change in the true hazard ratio.

The parameter estimates from each scenario are summarized in Tables 2-4. Observations were deleted if the resulting standard error calculated by PHREG for the parameter estimate was greater than 1 because this was taken as an indication that the procedure had trouble converging. This was not an issue in most of the simulation scenarios. The most severe scenario was in the scenario using Distribution 1, ~30 cases per cohort and a true hazard ratio of 1.015. In this scenario, 10,000 cohorts were created. However, only 7,911 of the parameter estimates calculated with 1:1 matching had calculated standard errors less than 1 (some had standard errors greater than 5,000). Only these observations were used to calculate the mean and variance of the parameter estimates in Tables 2-4.

The procedure PHREG will not converge if, in every risk-set, the exposure variable for the case is higher (lower) than the exposure variable of every control in the risk-set. In fact, the maximum likelihood estimate for this situation will be infinity (-infinity). As a result, PHREG will report the last estimate from when the optimization algorithm stopped, which most likely will be a large estimate with a large standard error. As a result of deleting these observations, the mean and variance of the estimates listed in Tables 2-4 where not all observations were used will be biased. Presumably, the reported mean variance is lower since the very large estimates were removed.

It is interesting to note that this issue was most severe in the scenarios with ~30 cases and a true hazard ratio of 1.015. The issue became significantly less severe as the number of cases increased and/or as the true hazard ratio decreased. This is to be

expected because with a high true hazard ratio, the difference between the exposure variables for the controls and the exposure variable for the case will be larger in each risk-set. This is illustrated in Tables 2-3. Note that as the true hazard ratio increases, the mean of the cumulative exposure for the cases increases and the difference between the means for the cases and controls also increases. Therefore, the higher the true hazard ratio, the more likely it becomes to randomly select controls that have lower exposures than the case in every risk-set (especially when there are few cases and few controls being sampled).

From Tables 2-4, it appears that the estimated variance is an unbiased estimator of the true variance, even with few number of cases. It also appears that the variance is inversely proportional to the number of cases. Note that the variances of the estimates with ~30 cases are about 3-4 times larger than the corresponding variances with ~100, and they are about 10 times larger than the corresponding variances with ~300 cases.

Tables 2-5 lists the relative efficiency and percent bias for each scenario and Figures 2-2 and Figures 2-3 give graphical representations of this data. Note that the relative efficiency when the true hazard ratio is 1 is close to $\frac{m}{m+1}$ for 1:*m* matching, and it gets closer to this value as the number of cases increases. This supports what is reported in the literature (Goldstein and Langholz, 1992). However, when the true hazard ratio increases, the relative efficiency decreases substantially. The relative efficiencies do appear to increase as the number of cases increase, but it does not appear that they are approaching $\frac{m}{m+1}$. For example, even with ~300 cases per cohort, the relative efficiency of 1:5 matching with exposure intensity Distribution 1 and true hazard ratio 1.015 is approximately 18.17% which is considerably lower than 83.33%,

the estimate given by $\frac{m}{m+1}$.

The reason for this drop in relative efficiency may be similar to the reason that PHREG has trouble converging described above. With a high hazard ratio, the difference between the exposure variables for the controls and the case will be greater in each risk-set and as a result, it would be more likely to get very large exposure-response estimates, especially when only a few controls are sampled. This would result in a large variance for the exposure-response variable, and therefore a drop in relative efficiency.

It also appears that the bias of the estimate is affected by the strength of the exposure-response parameter. In particular, the bias increases as the exposure-response variable gets stronger and this bias tends to be away from the null. However, as the number of cases increases, the bias decreases significantly supporting the fact that the estimate is asymptotically unbiased.

Next, it seems that the relative efficiency and bias are dependent on the distribution of exposure-intensity (and consequently dependent on the distribution of cumulative exposure of a cohort). Cumulative exposure is exposure intensity times duration of exposure. In these simulations, everyone was assigned a duration of exposure of 15 years (unless their age at death or age at censor occurred earlier than their maximum 15 years of employment), therefore, the distribution of cumulative exposure is about 15 times the distribution of exposure intensity. Note that Distribution 0 has the smallest skewness value (close to 0) and Distribution 2 has the largest skewness value.

It seems that, in general, Distribution 0 yields the highest relative efficiency and

smallest bias for a fixed true hazard ratio and approximate number of cases. Distribution 2 yields the lowest relative efficiency and largest bias (away from the null).

Through these simulations, it seems apparent that $\frac{m}{m+1}$ gives a good estimate of the relative efficiency for 1:*m* matching in a nested case-control study when there is no exposure-response relationship. This is all the theory guarantees. However, as the true exposure-response becomes stronger (resulting in a higher hazard ratio) the relative efficiency begins to drop. Even with ~300 cases (which is large for a cohort of 5000 individuals) the relative efficiency still decreases notably as the hazard ratio increases. The relative efficiency seems to increase slightly as the number of cases increases and the efficiency also appears to be affected by the distribution of the exposure parameter distribution. In particular, as the skewness of the distribution of the exposure variable increases, the efficiency decreases. Also, in these simulations, there appears to be some bias away from the null in a nested case-control study when the exposure distribution is right skewed. This bias increases as the true exposure-response becomes stronger and as the skewness of the exposure distribution increases. However, the bias decreases significantly and is not a major issue as the number of cases increases and as the case-control ratio increases. In fact, when 10 or more controls were matched with each case, the bias was never more than 10%, even with ~30 cases per cohort. Furthermore, for all simulations with ~100 cases, when 5 or more controls were matched with each case, the bias was never more than 3%.

Also, it is important to note that the size of the variance was always much larger compared to the bias. For example, if we considered the mean square error (mse) of the estimates, the variance of the estimate made up over 90% of the mse in every

simulation scenario.

Section 2

Objective:

It is also of interest to investigate the results of a nested case-control study for a given a cohort. For example, how do the parameter estimates and the estimated standard errors from a 1:5 matched nested case-control design of a given cohort compare to those estimates from the full cohort. In practice, we are given a cohort of workers and perform 1:m matching expecting to get similar results as if we analyzed the full cohort.

In this section, within cohort bias is defined as the expected parameter estimate from 1:m matching given a cohort minus the parameter estimated from the given full cohort (as opposed to the theoretical parameter that was investigated in Section 1). The variance of the parameter estimate within a cohort was also considered. In particular, occupational cohorts were simulated to get a better understanding of the following questions:

1. How is the within cohort bias of 1:m matching affected by the exposure-response relationship of the full cohort, the distribution of the exposure variable, and the number of cases in the cohort?
2. How is the variance of the parameter estimate within a cohort affected by the exposure-response relationship of the full cohort, the distribution of the exposure variable, and the number of cases in the cohort?

Method and Analysis:

The same simulation scenarios from Section 1 were considered. However, for this section, 100 cohorts were simulated for each scenario. Risk-sets were created for each cohort with age as the time scale. Then for each case, 1, 5, 10, 15, and 20 controls were randomly sampled 500 times from the risk-sets. The full as well as the sampled risk-sets were analyzed using conditional logistic regression (procedure PHREG in SAS) to get estimates of the exposure-response parameter. Therefore, for each simulated cohort, one estimate of the exposure-response parameter was obtained from analyzing the full cohort and 500 estimates were obtained from each of the 1:1, 1:5, 1:10, 1:15 and 1:20 sampled risk-sets.

Bias for 1:m matching within a cohort was estimated by subtracting the exposure-response estimate of analyzing the full cohort from the average of the 500 exposure-response estimates obtained by analyzing the sampled risk-sets. Therefore, there were 100 estimates of the within cohort bias obtained for each simulation scenario.

Results and Discussions:

As discussed in Section 1, results were deleted if the estimated standard error was greater than 1, as this was a sign that the model had trouble converging. Again, this was not a problem for most of the simulation scenarios.

For each cohort, summary statistics for the 100 within cohort bias estimates from each simulation scenario are summarized in Tables 2-6. Again, the trends reflect the conclusions from Section 1. Namely increasing the parameter estimate of the full cohort, increasing the skewness of the exposure distribution, and decreasing the number of

cases causes the percent bias to increase. Also, this bias tends to be away from the null. It is also interesting to note that the mean of the within cohort percent bias values of Tables 2-6 are similar to the percent bias of Section 1 listed in Tables 2-5. However, the within cohort bias varies greatly between cohorts. For some cohorts, the within cohort bias was greater than 100%.

Looking at the bias within a cohort may give some insight as to why the bias tends to be away from the null, and how the distribution affects this bias. Figures 2-4 are graphs of the parameter estimates of an example cohort with ~30 cases, a true hazard ratio of 1.005 and exposure intensity Distribution 0 and 2. The 500 parameter estimates are plotted vs. the number of matched controls per case (in black) along with the average of the 500 estimates (in red). A line represents the estimate of the full cohort. Note that the estimates corresponding to Distribution 0 (which is approximately symmetric) seem to be symmetrically distributed around the full cohort parameter estimate, whereas the estimates corresponding to Distribution 2 seem to have a tendency to overestimate the full cohort estimate. This may be due to the fact that Distribution 2 is skewed, making it more likely to sample lower exposed controls for each case. This would thus make it more likely to give a larger parameter estimate.

However, it should be noted that in each simulation scenario, the 95% confidence interval from a nested case-control analysis contained the estimate from the full cohort analysis nearly 100% of the time. This indicates that even though the nested case-control parameter estimates are straying from the full cohort estimate, the estimated variances are also getting larger so that confidence intervals are large enough to capture the full cohort estimate.

It was noted at the end of Section 1 that the variance is the dominating factor of the mse. It is interesting to investigate the variance by separating it into two components; the within cohort variance and the between cohort variance. To do this, for each simulation, define X_{ci} as the i^{th} estimate from the c^{th} cohort where c = 1,…,100 and i = 1,…,I_c (generally I_c = 500, however some estimates were dropped if their corresponding estimated standard errors were greater than 1). Then let:

$$\overline{X_{c.}} = \frac{\sum_i X_{ci}}{I_c} \tag{2.3}$$

$$\overline{X_{..}} = \frac{\sum_c \overline{X_{c.}}}{100} \tag{2.4}$$

and define:

$$Within\ SS = \sum_c \sum_i (X_{ci} - \overline{X_{c.}})^2 \tag{2.5}$$

$$Between\ SS = \sum_c \sum_i (\overline{X_{c.}} - \overline{X_{..}})^2 \tag{2.6}$$

$$Total\ SS = \sum_c \sum_i (X_{ci} - \overline{X_{..}})^2 \tag{2.7}$$

Note that $Within\ SS + Between\ SS = Total\ SS$. Tables 2-7 summarize the $Within\ SS$, $Between\ SS$, and $Total\ SS$ for each simulation scenario. Note that the $Total\ SS$ is proportional to the sample variance of the parameter estimate from a nested case-control analysis with 1:*m* matching, which was estimated in Section 1. Also, when $\overline{X_{c.}}$ is close to the full cohort estimate (which is especially true when the case-control ratio is high or there are many cases in the cohort), the $Between\ SS$ is proportional to the sample variance of the parameter estimates from analyzing the full cohort. Therefore, in these situations, the percent contribution of the $Between\ SS$ to the $Total\ SS$ is very close to the relative efficiency calculated in Section 1.

Notice that generally with a small case to control ratio, the $Within\ SS$ makes up a large percentage of the $Total\ SS$ and the percentage increases substantially when the exposure-response increases. However, the $Within\ SS$ decreases substantially as the case: control ratio increases and it also appears to be inversely proportional to this ratio (i.e. the $Within\ SS$ is cut in half when the case: control ratio is doubled). In fact, it will eventually decrease to 0 when the ratio is large enough that the entire risk-sets are sampled. It therefore seems that there is much to be gained, in terms of the within cohort variance of the parameter estimate, by selecting a larger case-to control ratio.

Comparison to Gold Miners Data:

To illustrate the implications of these simulations, data collected for the analysis of silicosis among gold miners (Steenland, Brown 1995) will be considered. The study consisted of 3,330 gold miners who worked for at least 1 year between 1940 – 1965 and were exposed to high levels of silica. Workers were followed-up until 1990 and there were 170 cases of silicosis determined. For this analysis, silicosis has the advantage of being associated with silica exposure alone so that no confounders had to be evaluated. Also, in this study, it was determined that cumulative exposure (or log of cumulative exposure) was the best predictor of disease when compared to duration of exposure and average exposure. Therefore, cumulative exposure was used as the exposure metric.

Cumulative exposure data was available for the entire cohort so that the full risk-sets as well as sampled risk-sets could be analyzed. Table 2-8 gives descriptive statistics of cumulative exposure and log of cumulative exposure for the risk-sets of the

data and Figures 2-5 are a histogram of this data. In particular, the cumulative exposure of individuals in the risk-sets has a range of 275 – 225521 and a skewness value of 2.1. The log of cumulative exposure has a range of 5.62 – 12.33 and a skewness value of -0.07. To compare this data to the simulations, the exposure variable must be scaled so that the range of the exposure metric for the risk-sets matches that of the simulations. Table 2-9 gives descriptive statistics of each exposure metric after being appropriately scaled. Table 2-10 gives the results of Cox proportional hazards regression on the full cohort with age as the time scale and using the original data and the scaled data for each exposure metric as the independent variable.

The risk-sets were formed using age as the time scale. From each risk-set, 1, 5, 10, 15, and 20 controls were randomly selected 500 times. The sampled risk-sets were then analyzed using conditional logistic regression and thus 500 parameter estimates were obtained for each case to control ratio. Summary statistics of this analyses are summarized in Table 2-11. The results were then compared to the analysis of the full cohort and the results of this comparison are summarized in Table 2-12.

Again, bias was defined as the difference between the parameter estimate of the full cohort and the mean of the 500 estimates obtained from the sample risk-set analysis. Also, relative efficiency was estimated as the estimated variance from the full cohort analysis divided by the average of the 500 estimated variances from the sampled cohort analysis. Similar trends are seen in this cohort as were seen in the previous sections. Namely, when the exposure distribution is very skewed (as is the case when the exposure metric used is cumulative exposure) there tends to be bias away from the null and the estimated relative efficiency is quite small. When the analysis is performed

on the log of cumulative exposure (which is not very skewed), the bias is smaller, compared to the analysis with cumulative exposure as the exposure metric. In fact, the bias is towards the null when the analysis is performed on the log of cumulative exposure for this cohort. This demonstrates the effect the exposure distribution has on the nested case-control estimate and reinforces the trends seen in the simulations. Namely, when the exposure distribution is skewed, the within cohort bias tends to be larger and it is more likely that the bias is away from the null.

Figures 2-6 plot the 500 parameter estimates vs. the sampling ratio with a red dot indicating the average of the 500 estimates and a dotted line indicating the estimate from the full cohort analysis.

Notice that when cumulative is used as the exposure metric, nested case-control analysis tends to overestimate the exposure-response parameter and this bias is greater than that of the nested case-control analysis when log of cumulative exposure is used as the exposure metric. Furthermore, the efficiency is smaller for the cumulative exposure analysis as compared to the log of cumulative exposure analysis. This supports the fact that when the exposure distribution is skewed, nested case-control analysis is more likely to give a larger estimate than the estimate from the full cohort analysis as was seen in the simulations.

Chapter 2 Conclusions:

Nested case-control studies are used to analyze cohorts in hopes of saving time and money. In practice, a ratio of 4 – 6 controls are selected and generally the justification for this ratio is that this analysis yields an asymptotically unbiased result and

the relative efficiency is about 80%, under the null hypothesis, when compared to the analysis performed on the full cohort. However, these are asymptotic properties, and we rarely conduct a study with the expectation of having no exposure-response relationship. Therefore, the non-asymptotic properties of a nested case-control study should be investigated.

These simulations have shown that the relative efficiency decreases and bias increases as:

- the exposure-response relationship becomes stronger
- the skewness of the exposure distribution increases
- the number of cases decreases

for nested case-control studies. Furthermore, the variance of the estimator is the dominating term of the mean square error when compared to the bias.

In Section 1, when the theoretical bias was estimated by comparing estimates to the true exposure-response relationship, the bias appeared to not be a major issue. It only seemed to be somewhat problematic in the simulations with ~30 cases per cohort, a skew distribution and a strong exposure-response relationship. However, when 10 controls were matched with each case, the bias was never more than 10%.

In practice, nested case-control studies are carried out in the hope of getting an estimate that is similar to the estimate from analyzing the full cohort. When the bias of a nested case-control study was considered with respect to the full cohort estimate, as was investigated in Section 2, the issue was a bit more severe. The within cohort bias varied quite a bit between cohorts and it was possible to still have a large within cohort bias, especially with a strong exposure-response relationship and a skewed exposure

distribution. However, as the bias increased, so did the variance estimate, which resulted in wide confidence intervals that almost always contained the full cohort parameter.

In Section 1, the variance of the parameter estimate for a nested case-control study was shown to dominate the mean square error. In Section 2, it was shown that most of the variance occurred within a cohort with a small case-control ratio. However, the within cohort variance of the parameter estimates appears to be inversely proportional to the case: control ratio and therefore, there could be significant gains by selecting a larger ratio. This is especially true when there are a few number of cases, since the variance is the largest in this situation and the impact on the cost of the study would be smaller. For example, performing a 1:5 matched nested case-control study for a cohort with 300 cases would require as much work as performing a 1:50 matched nested case-control study for a cohort with 30 cases.

Section 1 Tables and Graphs

Table 2-1: Values of α_0 used in each simulation scenario

	Distribution 0[a]			Distribution 1[b]			Distribution 2[c]		
	~30 cases	~100 cases	~300 cases	~30 cases	~100 cases	~300 cases	~30 cases	~100 cases	~300 cases
True Hazard Ratio	α_0	α_0	α_0	α_0	α_0	α_0	α_0	α_0	α_0
1	-8.20	-7.10	-6.00	-8.20	-7.10	-6.00	-8.20	-7.10	-6.00
1.005	-10.10	-9.00	-7.85	-9.50	-8.20	-7.05	-8.70	-7.40	-6.60
1.01	-12.30	-11.25	-10.10	-10.90	-9.70	-8.45	-9.20	-7.90	-7.80
1.015	-14.90	-13.80	-12.55	-13.20	-11.70	-10.05	-10.10	-8.50	-9.10

[a] – Distribution 0: normal (25, 64) - truncated (0, 50)
[b] – Distribution 1: log-normal (2.5, 0.5) - truncated (0, 50)
[c] – Distribution 2: log-normal (0.75, 1) – truncated (0, 50)

Table 2-2: Summary statistics of the number of cases for each simulation scenario
a)

	Distribution 0											
	~30 cases				~100 cases				~300 cases			
True Hazard Ratio	mean	median	min	max	mean	median	min	max	mean	median	min	max
1	34.73	35	13	58	103.17	103	67	138	304.08	304	251	360
1.005	33.60	33	12	58	99.83	100	69	135	305.41	305	244	354
1.01	36.51	36	18	60	102.17	102	67	136	300.22	300	256	362
1.015	36.76	37	18	62	103.55	103	71	138	303.17	303	253	360

b)

	Distribution 1											
	~30 cases				~100 cases				~300 cases			
True Hazard Ratio	mean	median	min	max	mean	median	min	max	mean	median	min	max
1	34.70	35	15	61	103.57	103	73	138	304.35	304	243	363
1.005	27.70	28	9	50	100.52	100	65	138	304.73	305	251	356
1.01	34.78	35	14	58	105.41	105	72	148	307.12	308	255	359
1.015	33.52	33	10	57	102.20	102	67	141	307.90	309	250	365

c)

	Distribution 2											
	~30 cases				~100 cases				~300 cases			
True Hazard Ratio	mean	median	min	max	mean	median	min	max	mean	median	min	max
1	34.68	35	16	59	103.53	104	70	138	303.39	303	246	363
1.005	28.20	28	12	50	101.92	102	70	148	296.32	296	253	357
1.01	31.53	31	14	55	101.48	101	68	136	295.65	295	251	349
1.015	34.02	34	14	58	105.94	106	77	141	307.49	307	249	377

Table 2-3: Averaged summary statistics of the cumulative exposure for cases and controls of the full risk-sets in each simulation scenario. For each cohort simulated, the risk-sets were formed and the mean, variance, skewness, min and max of the cumulative exposures across all risk-sets were calculated. Then these statistics were averaged across all cohorts and are summarized below.

a)

Distribution 0

		~30 cases					~ 100 cases					~ 300 cases				
		Cumulative Exposure					Cumulative Exposure					Cumulative Exposure				
True Hazard Ratio	Group	Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max
1	case	325.081	20592.19	-0.10389	44.835	606.674	324.702	20559.87	-0.08661	23.236	655.976	323.726	20590.51	-0.08553	14.315	692.522
	cont	323.912	20114.56	-0.0579	0.967	743.025	323.465	20218.86	-0.05947	0.66	743.181	322.338	20334.91	-0.05576	0.515	742.736
1.005	case	420.635	17554.89	-0.26963	121.711	669.954	420.075	17469.63	-0.2944	65.2	705.352	417.398	17550.58	-0.30936	31.215	727.841
	cont	335.713	18863.08	-0.09052	1.26	743.042	334.994	18870.57	-0.09198	0.848	743.004	332.971	18825.87	-0.09139	0.655	742.824
1.01	case	499.092	14169.74	-0.31913	220.821	713.522	496.172	13992.8	-0.35243	157.731	731.688	488.593	13638.49	-0.3673	98.688	740.068
	cont	340.914	18232.66	-0.10326	1.433	742.95	339.643	18176.76	-0.1081	1.073	742.944	336.392	17944.18	-0.11876	0.837	742.954
1.015	case	558.663	11013.94	-0.43632	309.967	731.288	551.871	10759.55	-0.46241	250.941	739.853	534.388	10039.18	-0.44613	190.601	742.411
	cont	342.634	17972.82	-0.11202	1.532	742.911	340.972	17841.92	-0.12491	1.152	742.697	336.008	17416.64	-0.1526	0.969	742.547

b)

Distribution 1

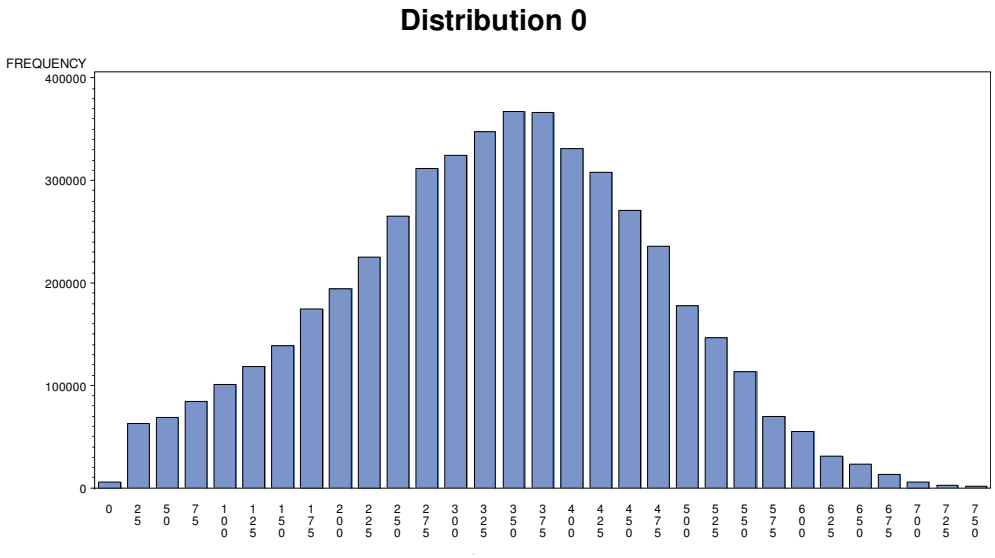
True Hazard Ratio	Group	~30 cases: Cumulative Exposure					~ 100 cases: Cumulative Exposure					~ 300 cases: Cumulative Exposure				
		Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max
1	case	178.127	11746.55	1.01864	23.183	483.99	178.163	11780.39	1.15077	11.914	574.606	177.592	11774.27	1.19354	7.472	644.814
	cont	177.404	11574.94	1.23746	2.39	740.38	177.235	11598.19	1.23458	2.139	740.166	176.551	11601.85	1.23361	2.039	740.88
1.005	case	263.038	23976.9	0.85517	47.622	634.868	261.051	23563.95	0.89836	21.532	704.581	255.366	22326.59	0.91801	11.499	728.371
	cont	182.042	11479.13	1.24004	2.552	740.239	181.325	11383.63	1.23045	2.231	740.1	179.467	11087.07	1.21065	2.086	739.522
1.01	case	408.371	34728.35	0.0481	84.367	726.094	388.514	32132.97	0.12724	47.949	737.02	349.49	26423.56	0.27261	23.911	736.481
	cont	185.138	11253.99	1.2241	2.576	739.753	183.365	10886.1	1.17281	2.327	739.178	179.039	10065.31	1.07253	2.19	736.65
1.015	case	537.765	23688.83	-0.78623	171.304	737.846	482.093	22740.9	-0.49277	100.268	737.109	403.74	19026.45	-0.18024	51.686	712.977
	cont	185.824	11033.34	1.18461	2.622	738.691	182.121	10404.67	1.07983	2.377	736.722	175.856	9237.21	0.92274	2.241	712.859

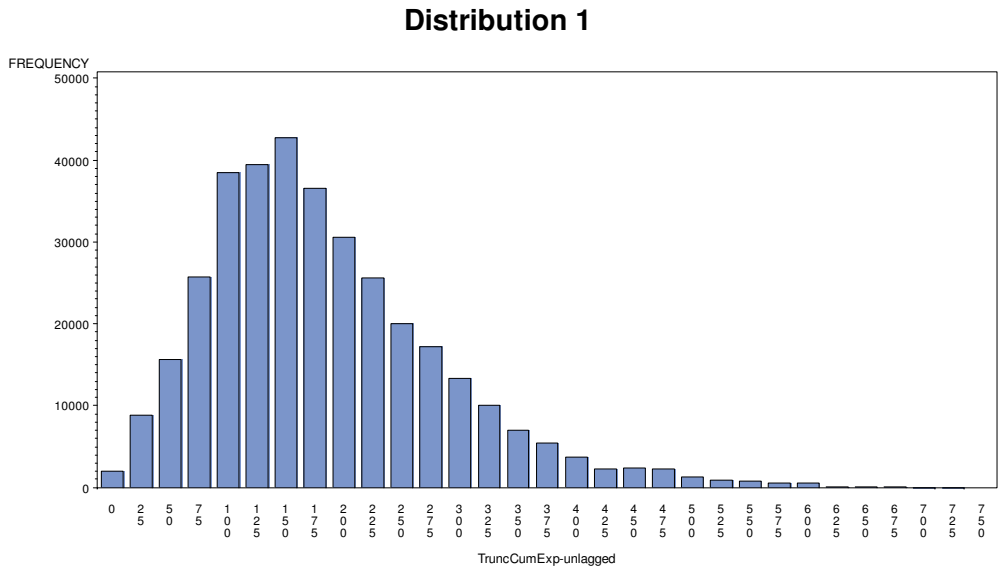
c)

Distribution 2

		~30 cases					~ 100 cases					~ 300 cases				
		Cumulative Exposure					Cumulative Exposure					Cumulative Exposure				
True Hazard Ratio	Group	Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max	Mean	Variance	Skew	Min	Max
1	case	44.648	3329.91	2.37747	2.34064	252.885	44.415	3265.82	3.07615	1.18338	355.083	44.568	3294.77	3.50839	0.6443	463.604
	cont	44.554	3301.63	3.86759	0.08303	704.384	44.51	3296.41	3.86505	0.0673	704.758	44.337	3282.89	3.87194	0.06103	705.347
1.005	case	79.936	14198.81	2.48816	3.27776	478.333	78.25	13427.19	2.91972	1.38916	630.207	73.469	11413.22	3.03429	0.75332	678.114
	cont	44.926	3315.11	3.83868	0.08886	704.138	44.643	3239.02	3.7976	0.06863	704.346	44.011	3077.04	3.70717	0.06098	698.692
1.01	case	207.442	49165.55	1.01213	4.54812	691.181	157.908	32697.3	1.34277	1.90514	688.298	117.156	19132.32	1.68722	0.92986	651.426
	cont	45.213	3175.34	3.60282	0.09056	696.879	44.221	2894.2	3.36787	0.07104	687.452	42.784	2525.98	3.09729	0.06241	651.354
1.015	case	329.562	43954.6	-0.14686	9.19886	669.689	218.439	29901.04	0.4295	2.78664	600.67	144.119	16918.88	0.89394	1.1687	523.298
	cont	44.514	2885.79	3.25436	0.08945	654.948	43.171	2519.9	2.95098	0.07269	597.407	41.242	2096.55	2.66393	0.06273	522.927

Figures 2-1: Histograms of cumulative exposure for an example cohort's risk-sets of each distribution





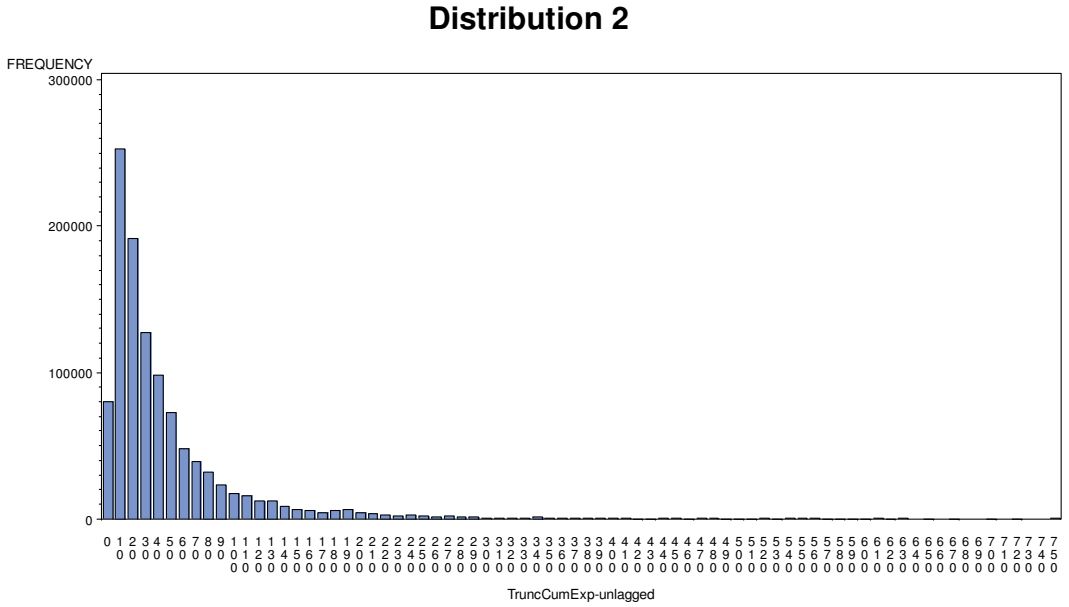


Table 2-4: Summary statistics of the exposure-response parameter estimates for each scenario

a)

		Distribution 0											
		~30 cases				~100 cases				~300 cases			
True Hazard Ratio	Match	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]
$1 = e^{0}$	1:1	10000	-2.01E-05	4.66E-06	4.23E-06	3000	1.50E-05	1.28E-06	1.24E-06	1000	-1.91E-05	4.07E-07	4.05E-07
	1:5	10000	8.52E-06	2.30E-06	2.23E-06	3000	-2.84E-06	7.26E-07	7.17E-07	1000	-2.16E-06	2.44E-07	2.40E-07
	1:10	10000	-6.35E-06	2.03E-06	2.01E-06	3000	5.31E-06	6.72E-07	6.54E-07	1000	-5.46E-06	2.30E-07	2.20E-07
	1:15	10000	1.83E-06	1.95E-06	1.94E-06	3000	-6.28E-06	6.56E-07	6.33E-07	1000	-5.18E-06	2.19E-07	2.13E-07
	1:20	10000	-3.83E-06	1.92E-06	1.91E-06	3000	2.85E-06	6.53E-07	6.23E-07	1000	-5.37E-06	2.17E-07	2.09E-07
	full	10000	-1.75E-06	1.80E-06	1.81E-06	3000	-1.72E-06	6.14E-07	5.93E-07	1000	-5.74E-06	2.10E-07	1.99E-07
$1.005 = e^{0.00499}$	1:1	10000	5.57E-03	3.37E-05	1.02E-04	3000	5.20E-03	1.98E-06	1.91E-06	1000	5.06E-03	6.16E-07	5.86E-07
	1:5	10000	5.11E-03	2.92E-06	2.82E-06	3000	5.03E-03	9.20E-07	8.93E-07	1000	5.01E-03	3.02E-07	2.88E-07
	1:10	10000	5.07E-03	2.46E-06	2.40E-06	3000	5.03E-03	7.81E-07	7.74E-07	1000	5.00E-03	2.56E-07	2.50E-07
	1:15	10000	5.04E-03	2.31E-06	2.25E-06	3000	5.01E-03	7.26E-07	7.31E-07	1000	4.99E-03	2.42E-07	2.37E-07
	1:20	10000	5.03E-03	2.22E-06	2.18E-06	3000	5.01E-03	7.37E-07	7.12E-07	1000	5.00E-03	2.29E-07	2.31E-07
	full	10000	5.00E-03	2.00E-06	1.98E-06	3000	5.00E-03	6.63E-07	6.51E-07	1000	4.99E-03	2.16E-07	2.12E-07
$1.01 = e^{0.00995}$	1:1	9993	1.14E-02	3.21E-05	3.93E-05	3000	1.04E-02	4.55E-06	4.35E-06	1000	1.01E-02	1.28E-06	1.33E-06
	1:5	10000	1.03E-02	4.94E-06	4.57E-06	3000	1.01E-02	1.55E-06	1.49E-06	1000	9.97E-03	5.43E-07	4.95E-07
	1:10	10000	1.02E-02	3.49E-06	3.36E-06	3000	1.00E-02	1.17E-06	1.13E-06	1000	9.97E-03	4.13E-07	3.85E-07
	1:15	10000	1.01E-02	3.03E-06	2.96E-06	3000	1.00E-02	1.03E-06	1.01E-06	1000	9.96E-03	3.48E-07	3.45E-07
	1:20	10000	1.01E-02	2.81E-06	2.77E-06	3000	1.00E-02	9.68E-07	9.52E-07	1000	9.96E-03	3.31E-07	3.25E-07
	full	10000	1.00E-02	2.12E-06	2.12E-06	3000	9.97E-03	7.49E-07	7.47E-07	1000	9.95E-03	2.64E-07	2.59E-07
$1.015 = e^{0.01489}$	1:1	9782	1.88E-02	2.24E-04	4.39E-04	3000	1.58E-02	1.72E-05	1.33E-05	1000	1.52E-02	3.45E-06	3.28E-06
	1:5	9999	1.57E-02	1.26E-05	1.09E-05	3000	1.51E-02	3.28E-06	3.03E-06	1000	1.49E-02	1.03E-06	9.86E-07
	1:10	10000	1.53E-02	7.20E-06	6.62E-06	3000	1.50E-02	2.13E-06	2.07E-06	1000	1.49E-02	6.75E-07	6.97E-07
	1:15	10000	1.52E-02	5.74E-06	5.30E-06	3000	1.50E-02	1.73E-06	1.73E-06	1000	1.49E-02	5.67E-07	5.94E-07
	1:20	10000	1.52E-02	5.06E-06	4.70E-06	3000	1.50E-02	1.55E-06	1.56E-06	1000	1.49E-02	5.20E-07	5.41E-07
	full	10000	1.50E-02	2.74E-06	2.66E-06	3000	1.49E-02	8.95E-07	9.44E-07	1000	1.49E-02	3.19E-07	3.46E-07

[a] N is the number of parameter estimates with corresponding standard error calculated by Proc Phreg less than 1

[b] The sample variance of the parameter estimates

[c] Average of the variance estimates from each analysis

b)

		Distribution 1											
		~30 cases				~100 cases				~300 cases			
True Hazard Ratio	Match	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]
$1 = e^{0}$	1:1	10000	-2.23E-06	8.07E-06	7.17E-06	3000	3.00E-05	2.02E-06	1.99E-06	1000	3.38E-05	6.47E-07	6.43E-07
	1:5	10000	-1.45E-04	3.80E-06	3.65E-06	3000	-1.46E-05	1.13E-06	1.14E-06	1000	1.59E-05	3.76E-07	3.78E-07
	1:10	10000	-1.72E-04	3.38E-06	3.29E-06	3000	-2.30E-05	1.01E-06	1.03E-06	1000	-5.34E-06	3.43E-07	3.46E-07
	1:15	10000	-1.83E-04	3.24E-06	3.17E-06	3000	-2.84E-05	9.92E-07	1.00E-06	1000	2.22E-06	3.15E-07	3.35E-07
	1:20	10000	-1.75E-04	3.18E-06	3.11E-06	3000	-3.44E-05	9.60E-07	9.81E-07	1000	9.29E-06	3.16E-07	3.30E-07
	full	10000	-1.89E-04	2.98E-06	2.94E-06	3000	-3.61E-05	9.05E-07	9.34E-07	1000	3.64E-06	2.98E-07	3.14E-07
$1.005 = e^{0.00499}$	1:1	9999	5.80E-03	1.60E-05	1.58E-05	3000	5.20E-03	2.26E-06	2.13E-06	1000	5.05E-03	6.91E-07	6.72E-07
	1:5	10000	5.13E-03	3.49E-06	3.30E-06	3000	5.03E-03	8.32E-07	8.24E-07	1000	4.99E-03	3.01E-07	2.75E-07
	1:10	10000	5.05E-03	2.70E-06	2.57E-06	3000	5.01E-03	6.63E-07	6.54E-07	1000	4.99E-03	2.36E-07	2.20E-07
	1:15	10000	4.99E-03	2.46E-06	2.30E-06	3000	5.00E-03	5.98E-07	5.93E-07	1000	4.97E-03	2.19E-07	2.01E-07
	1:20	10000	4.97E-03	2.27E-06	2.17E-06	3000	4.98E-03	5.65E-07	5.61E-07	1000	4.98E-03	2.03E-07	1.91E-07
	full	10000	4.89E-03	1.82E-06	1.76E-06	3000	4.97E-03	4.69E-07	4.64E-07	1000	4.98E-03	1.74E-07	1.59E-07
$1.01 = e^{0.00995}$	1:1	9952	1.23E-02	1.29E-04	3.23E-04	3000	1.04E-02	4.63E-06	4.61E-06	1000	1.01E-02	1.42E-06	1.36E-06
	1:5	10000	1.04E-02	4.88E-06	4.55E-06	3000	1.01E-02	1.25E-06	1.27E-06	1000	9.99E-03	4.50E-07	4.35E-07
	1:10	10000	1.02E-02	2.94E-06	2.80E-06	3000	1.00E-02	8.40E-07	8.48E-07	1000	9.98E-03	3.18E-07	3.06E-07
	1:15	10000	1.01E-02	2.35E-06	2.24E-06	3000	1.00E-02	7.13E-07	7.01E-07	1000	9.94E-03	2.55E-07	2.56E-07
	1:20	10000	1.01E-02	1.98E-06	1.94E-06	3000	1.00E-02	6.17E-07	6.22E-07	1000	9.96E-03	2.38E-07	2.31E-07
	full	10000	9.96E-03	9.19E-07	9.21E-07	3000	9.97E-03	3.15E-07	3.18E-07	1000	9.95E-03	1.32E-07	1.32E-07
$1.015 = e^{0.01489}$	1:1	7911	2.04E-02	4.84E-04	1.64E-03	2992	1.69E-02	4.38E-05	4.52E-05	1000	1.52E-02	3.53E-06	3.51E-06
	1:5	9897	1.74E-02	1.19E-04	2.71E-04	3000	1.53E-02	4.29E-06	3.88E-06	1000	1.49E-02	8.97E-07	9.55E-07
	1:10	9987	1.62E-02	5.98E-05	8.20E-05	3000	1.51E-02	2.41E-06	2.30E-06	1000	1.49E-02	6.14E-07	6.31E-07
	1:15	9996	1.58E-02	1.46E-05	1.24E-05	3000	1.51E-02	1.88E-06	1.79E-06	1000	1.49E-02	5.21E-07	5.13E-07
	1:20	9996	1.56E-02	1.05E-05	9.51E-06	3000	1.50E-02	1.52E-06	1.50E-06	1000	1.49E-02	4.39E-07	4.50E-07
	full	10000	1.50E-02	1.50E-06	1.47E-06	3000	1.49E-02	4.46E-07	4.82E-07	1000	1.49E-02	1.63E-07	1.91E-07

[a] N is the number of parameter estimates with corresponding standard error calculated by Proc Phreg less than 1

[b] The sample variance of the parameter estimates

[c] Average of the variance estimates from each analysis

c)

		Distribution 2											
		~30 cases				~100 cases				~300 cases			
True Hazard Ratio	Match	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]	N[a]	Mean	Variance[b]	Mean of Estimated Variance[c]
$1 = e^{0}$	1:1	10000	-1.71E-04	4.56E-05	3.60E-05	3000	-1.61E-04	8.91E-06	7.74E-06	1000	3.66E-05	2.67E-06	2.25E-06
	1:5	10000	-8.56E-04	1.82E-05	1.61E-05	3000	-3.40E-04	4.59E-06	4.20E-06	1000	-5.85E-05	1.34E-06	1.29E-06
	1:10	10000	-9.51E-04	1.59E-05	1.43E-05	3000	-3.85E-04	4.01E-06	3.79E-06	1000	-8.98E-05	1.20E-06	1.17E-06
	1:15	10000	-9.90E-04	1.49E-05	1.37E-05	3000	-4.01E-04	3.83E-06	3.66E-06	1000	-7.93E-05	1.15E-06	1.14E-06
	1:20	10000	-1.02E-03	1.47E-05	1.34E-05	3000	-4.06E-04	3.77E-06	3.59E-06	1000	-9.87E-05	1.16E-06	1.12E-06
	full	10000	-1.08E-03	1.35E-05	1.25E-05	3000	-4.13E-04	3.57E-06	3.40E-06	1000	-1.03E-04	1.09E-06	1.06E-06
$1.005 = e^{0.00499}$	1:1	10000	7.01E-03	7.07E-05	5.13E-05	3000	5.51E-03	5.97E-06	5.89E-06	1000	5.14E-03	2.14E-06	1.89E-06
	1:5	10000	5.14E-03	1.08E-05	9.20E-06	3000	5.09E-03	2.02E-06	1.94E-06	1000	4.98E-03	6.70E-07	6.90E-07
	1:10	10000	4.90E-03	7.92E-06	6.79E-06	3000	5.05E-03	1.42E-06	1.45E-06	1000	4.97E-03	5.35E-07	5.25E-07
	1:15	10000	4.79E-03	6.94E-06	5.94E-06	3000	5.00E-03	1.30E-06	1.27E-06	1000	4.98E-03	4.64E-07	4.65E-07
	1:20	10000	4.72E-03	6.49E-06	5.50E-06	3000	4.98E-03	1.22E-06	1.17E-06	1000	4.96E-03	4.45E-07	4.31E-07
	full	10000	4.44E-03	4.81E-06	4.07E-06	3000	4.91E-03	8.59E-07	8.40E-07	1000	4.95E-03	3.28E-07	3.19E-07
$1.01 = e^{0.00995}$	1:1	9999	1.36E-02	1.16E-04	1.01E-04	3000	1.08E-02	8.97E-06	8.85E-06	1000	1.02E-02	2.93E-06	2.74E-06
	1:5	10000	1.07E-02	9.79E-06	8.96E-06	3000	1.02E-02	2.30E-06	2.26E-06	1000	9.98E-03	7.64E-07	8.39E-07
	1:10	10000	1.04E-02	5.54E-06	5.13E-06	3000	1.01E-02	1.43E-06	1.45E-06	1000	1.00E-02	5.83E-07	5.80E-07
	1:15	10000	1.03E-02	4.02E-06	3.86E-06	3000	1.01E-02	1.14E-06	1.17E-06	1000	1.00E-02	4.79E-07	4.80E-07
	1:20	10000	1.02E-02	3.20E-06	3.20E-06	3000	1.00E-02	9.87E-07	1.01E-06	1000	9.98E-03	3.95E-07	4.25E-07
	full	10000	9.96E-03	8.00E-07	7.91E-07	3000	9.97E-03	3.14E-07	3.39E-07	1000	9.97E-03	1.78E-07	1.91E-07
$1.015 = e^{0.01489}$	1:1	9791	2.51E-02	1.08E-03	1.82E-03	3000	1.61E-02	1.77E-05	1.67E-05	1000	1.52E-02	4.38E-06	4.19E-06
	1:5	10000	1.67E-02	4.41E-05	4.43E-05	3000	1.52E-02	3.95E-06	3.72E-06	1000	1.50E-02	1.20E-06	1.21E-06
	1:10	9998	1.59E-02	1.53E-05	1.42E-05	3000	1.51E-02	2.35E-06	2.27E-06	1000	1.50E-02	8.30E-07	8.05E-07
	1:15	10000	1.56E-02	8.95E-06	8.25E-06	3000	1.50E-02	1.74E-06	1.76E-06	1000	1.49E-02	6.21E-07	6.51E-07
	1:20	10000	1.54E-02	6.91E-06	6.42E-06	3000	1.50E-02	1.48E-06	1.50E-06	1000	1.49E-02	5.31E-07	5.68E-07
	full	10000	1.50E-02	8.68E-07	8.96E-07	3000	1.49E-02	3.14E-07	3.73E-07	1000	1.49E-02	1.82E-07	2.12E-07

[a] N is the number of parameter estimates with corresponding standard error calculated by Proc Phreg less than 1

[b] The sample variance of the parameter estimates

[c] Average of the variance estimates from each analysis

Table 2-5: Relative Efficiency and Bias of the exposure-response parameter estimate for each scenario.

a)

		Distribution 0								
		~30 cases			~ 100 cases			~ 300 cases		
True Hazard Ratio	Match	Percent Bias[a]	Relative Efficiency[b]	95% CI Captures True Value	Percent Bias[a]	Relative Efficiency[b]	95% CI Captures True Value	Percent Bias[a]	Relative Efficiency[b]	95% CI Captures True Value
1	1:1		38.68%	96.41%		48.16%	95.77%		51.60%	95.00%
	1:5		78.34%	95.36%		84.57%	95.23%		86.07%	94.90%
	1:10		88.68%	95.37%		91.37%	94.80%		91.30%	95.30%
	1:15		92.60%	95.33%		93.60%	94.57%		95.89%	95.80%
	1:20		94.00%	95.26%		94.03%	94.77%		96.77%	94.60%
	full		100.00%	95.16%		100.00%	94.74%		100.00%	95.41%
1.005	1:1	11.68%	5.94%	96.57%	4.26%	33.42%	95.87%	1.45%	35.06%	94.60%
	1:5	2.46%	68.41%	95.54%	0.85%	72.07%	94.93%	0.45%	71.52%	95.50%
	1:10	1.65%	81.43%	95.18%	0.85%	84.89%	95.03%	0.25%	84.38%	95.20%
	1:15	1.05%	86.72%	94.99%	0.45%	91.32%	95.13%	0.05%	89.26%	95.40%
	1:20	0.85%	90.21%	95.13%	0.45%	89.96%	94.83%	0.25%	94.32%	94.70%
	full	0.25%	100.00%	94.96%	0.25%	100.00%	94.77%	0.05%	100.00%	95.31%
1.01	1:1	14.47%	6.60%	96.55%	4.02%	16.45%	95.90%	1.20%	20.56%	94.00%
	1:5	3.51%	42.91%	95.36%	1.40%	48.39%	95.63%	0.20%	48.62%	95.50%
	1:10	2.11%	60.78%	95.60%	0.80%	64.02%	95.07%	0.20%	63.92%	94.70%
	1:15	1.71%	69.99%	95.49%	0.50%	73.07%	95.40%	0.10%	75.86%	95.20%
	1:20	1.60%	75.55%	95.47%	0.70%	77.38%	94.87%	0.10%	79.76%	94.60%
	full	0.70%	100.00%	95.07%	0.20%	100.00%	95.04%	0.00%	100.00%	95.31%
1.015	1:1	26.07%	1.22%	95.60%	6.32%	5.19%	96.07%	2.02%	9.26%	95.40%
	1:5	5.32%	21.64%	96.15%	1.35%	27.33%	94.87%	0.28%	31.00%	95.10%
	1:10	2.90%	37.97%	95.67%	1.02%	42.06%	94.90%	0.14%	47.26%	96.20%
	1:15	2.16%	47.61%	95.31%	0.61%	51.82%	95.03%	0.14%	56.26%	95.30%
	1:20	2.02%	54.08%	95.17%	0.75%	57.78%	95.67%	0.28%	61.35%	95.70%
	full	0.48%	100.00%	94.79%	0.21%	100.00%	95.77%	0.14%	100.00%	95.60%

[a] Percent bias between the true exposure-response relationship (log of the true hazard ratio) and the mean of the exposure-response parameter estimates

[b] Relative efficiency is estimated by 100% times dividing the sample variance of the exposure-response parameter estimates for the full cohort by the sample variance of the exposure-response parameter estimates for the sampled cohort

b)

		Distribution 1								
		~30 cases			~ 100 cases			~ 300 cases		
True Hazard Ratio	Match	Percent Bias	Relative Efficiency	95% CI Captures True Value	Percent Bias	Relative Efficiency	95% CI Captures True Value	Percent Bias	Relative Efficiency	95% CI Captures True Value
1	1:1		36.94%	97.09%		44.71%	95.83%		46.06%	95.40%
	1:5		78.49%	95.61%		80.16%	95.50%		79.26%	95.60%
	1:10		88.33%	95.27%		89.43%	95.63%		86.88%	94.40%
	1:15		92.12%	95.34%		91.23%	95.43%		94.60%	96.10%
	1:20		93.86%	95.20%		94.27%	95.47%		94.30%	95.50%
	full		100.00%	95.09%		100.00%	95.84%		100.00%	95.70%
1.005	1:1	16.29%	11.34%	96.64%	4.26%	20.80%	95.13%	1.25%	25.18%	94.50%
	1:5	2.86%	52.14%	95.91%	0.85%	56.37%	95.23%	0.05%	57.81%	93.70%
	1:10	1.25%	67.40%	95.84%	0.45%	70.74%	95.57%	0.05%	73.73%	95.10%
	1:15	0.05%	73.95%	95.26%	0.25%	78.43%	95.60%	-0.35%	79.45%	93.60%
	1:20	-0.35%	80.04%	95.72%	-0.15%	83.01%	95.27%	-0.15%	85.71%	94.40%
	full	-1.96%	100.00%	95.16%	-0.35%	100.00%	94.84%	-0.15%	100.00%	93.81%
1.01	1:1	23.81%	0.71%	96.24%	4.72%	6.80%	96.57%	1.50%	9.27%	94.90%
	1:5	4.12%	18.82%	96.28%	1.30%	25.18%	96.10%	0.40%	29.33%	94.90%
	1:10	2.51%	31.27%	95.82%	0.60%	37.50%	95.27%	0.30%	41.51%	94.80%
	1:15	1.91%	39.06%	95.88%	0.60%	44.18%	95.10%	-0.10%	51.76%	95.80%
	1:20	1.50%	46.41%	95.53%	0.60%	51.05%	95.47%	0.10%	55.46%	95.40%
	full	0.10%	100.00%	95.41%	0.20%	100.00%	95.14%	0.00%	100.00%	94.91%
1.015	1:1	36.95%	0.31%	94.57%	13.44%	1.02%	96.52%	2.23%	4.61%	95.50%
	1:5	16.80%	1.26%	96.27%	2.76%	10.40%	95.47%	0.35%	18.17%	95.70%
	1:10	9.01%	2.51%	96.41%	1.69%	18.51%	95.17%	0.21%	26.55%	96.60%
	1:15	6.12%	10.29%	96.57%	1.42%	23.69%	94.90%	0.28%	31.29%	94.50%
	1:20	4.71%	14.28%	96.08%	1.02%	29.40%	95.43%	0.28%	37.13%	96.40%
	full	1.02%	100.00%	95.56%	0.28%	100.00%	95.54%	-0.13%	100.00%	96.60%

[a] Bias is the true exposure-response relationship (log of the true hazard ratio) minus the mean of the exposure-response parameter estimates

[b] Relative efficiency is the variance of the exposure-response parameter estimates for the full cohort divided by the variance of the exposure-response parameter estimates for the sampled cohort times 100%

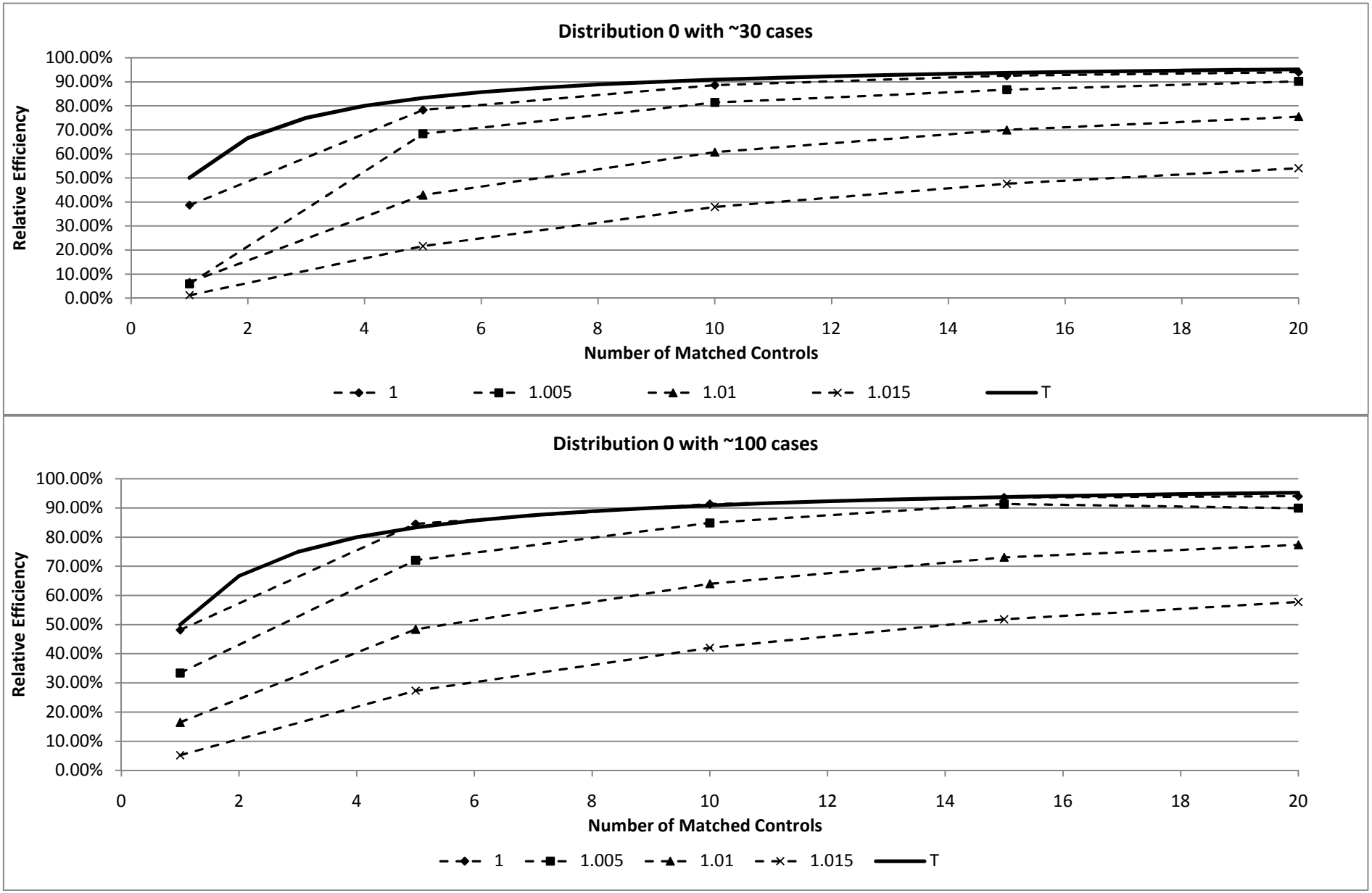
c)

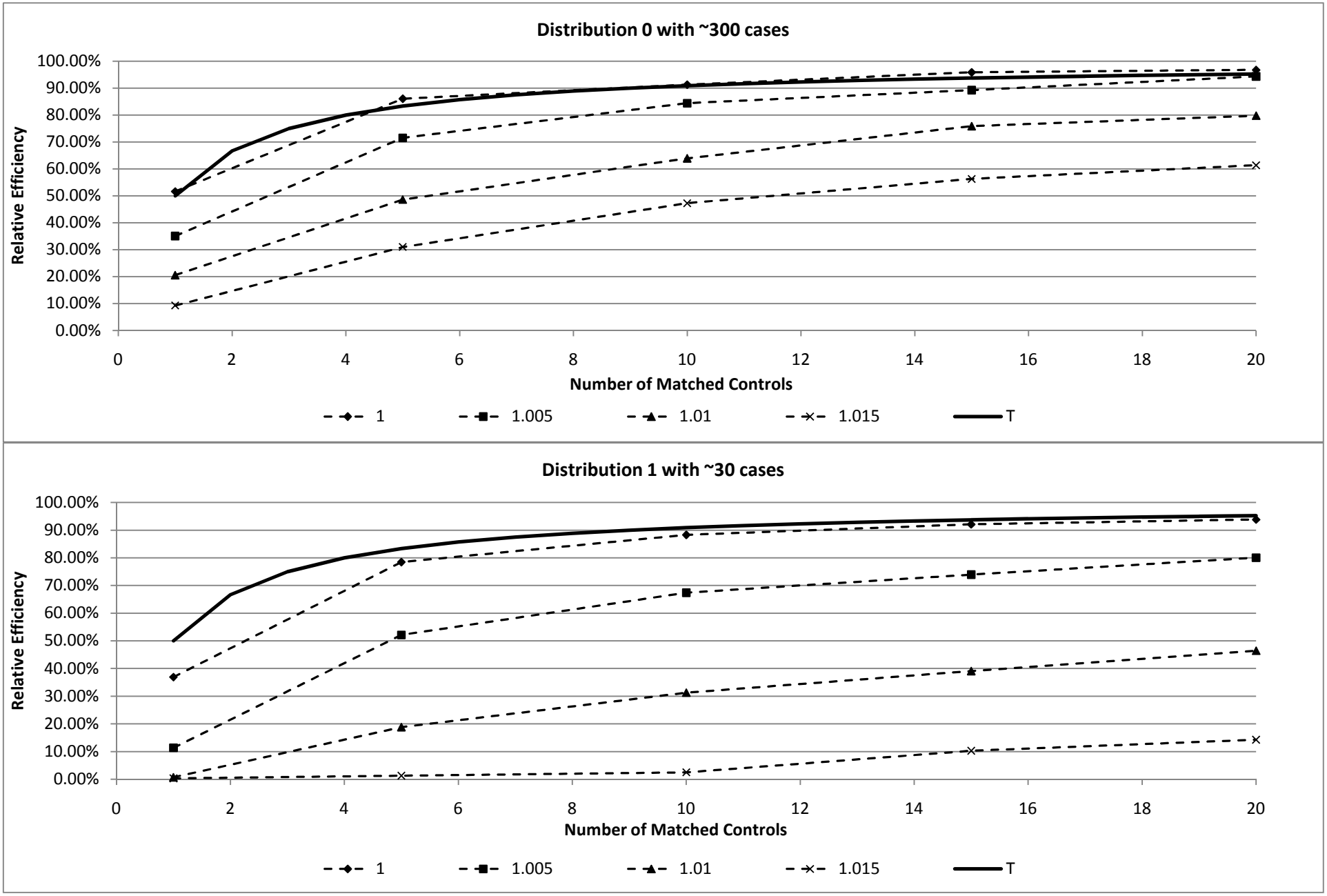
		Distribution 2								
		~30 cases			~ 100 cases			~ 300 cases		
True Hazard Ratio	Match	Percent Bias	Relative Efficiency	95% CI Captures True Value	Percent Bias	Relative Efficiency	95% CI Captures True Value	Percent Bias	Relative Efficiency	95% CI Captures True Value
1	1:1		29.60%	98.35%		40.06%	96.37%		40.79%	94.30%
	1:5		74.35%	96.20%		77.85%	95.53%		80.95%	95.30%
	1:10		85.11%	95.82%		88.94%	95.00%		90.67%	95.40%
	1:15		90.36%	95.71%		93.21%	95.57%		95.02%	95.60%
	1:20		91.95%	95.50%		94.69%	94.97%		93.63%	95.30%
	full		100.00%	95.05%		100.00%	95.10%		100.00%	95.01%
1.005	1:1	40.55%	6.81%	96.86%	10.48%	14.38%	96.37%	3.06%	15.36%	94.10%
	1:5	3.06%	44.70%	97.31%	2.05%	42.59%	95.13%	-0.15%	48.96%	95.10%
	1:10	-1.76%	60.73%	97.28%	1.25%	60.37%	95.67%	-0.35%	61.31%	95.00%
	1:15	-3.96%	69.32%	97.34%	0.25%	66.33%	96.07%	-0.15%	70.69%	95.30%
	1:20	-5.36%	74.10%	97.05%	-0.15%	70.41%	95.47%	-0.55%	73.71%	95.70%
	full	-10.98%	100.00%	95.96%	-1.55%	100.00%	95.10%	-0.75%	100.00%	95.01%
1.01	1:1	36.28%	0.69%	95.45%	8.04%	3.50%	95.67%	2.71%	6.09%	95.20%
	1:5	7.84%	8.17%	95.71%	2.01%	13.67%	95.33%	0.30%	23.30%	96.00%
	1:10	4.82%	14.43%	95.21%	1.00%	21.90%	95.73%	0.50%	30.53%	96.10%
	1:15	3.41%	19.89%	95.55%	1.10%	27.45%	95.17%	0.50%	37.16%	95.20%
	1:20	2.81%	24.97%	95.94%	0.80%	31.81%	95.53%	0.30%	45.06%	95.70%
	full	0.10%	100.00%	95.32%	0.20%	100.00%	96.07%	0.20%	100.00%	95.80%
1.015	1:1	68.25%	0.08%	95.52%	7.80%	1.77%	96.13%	2.23%	4.15%	96.40%
	1:5	12.43%	1.97%	96.07%	2.23%	7.94%	95.07%	0.55%	15.20%	95.30%
	1:10	6.79%	5.68%	95.83%	1.29%	13.36%	94.90%	0.61%	21.93%	95.00%
	1:15	4.64%	9.70%	95.77%	0.82%	18.01%	95.60%	0.35%	29.31%	94.70%
	1:20	3.64%	12.56%	95.91%	0.88%	21.27%	95.17%	0.28%	34.27%	95.70%
	full	0.68%	100.00%	96.02%	0.14%	100.00%	96.87%	0.01%	100.00%	96.80%

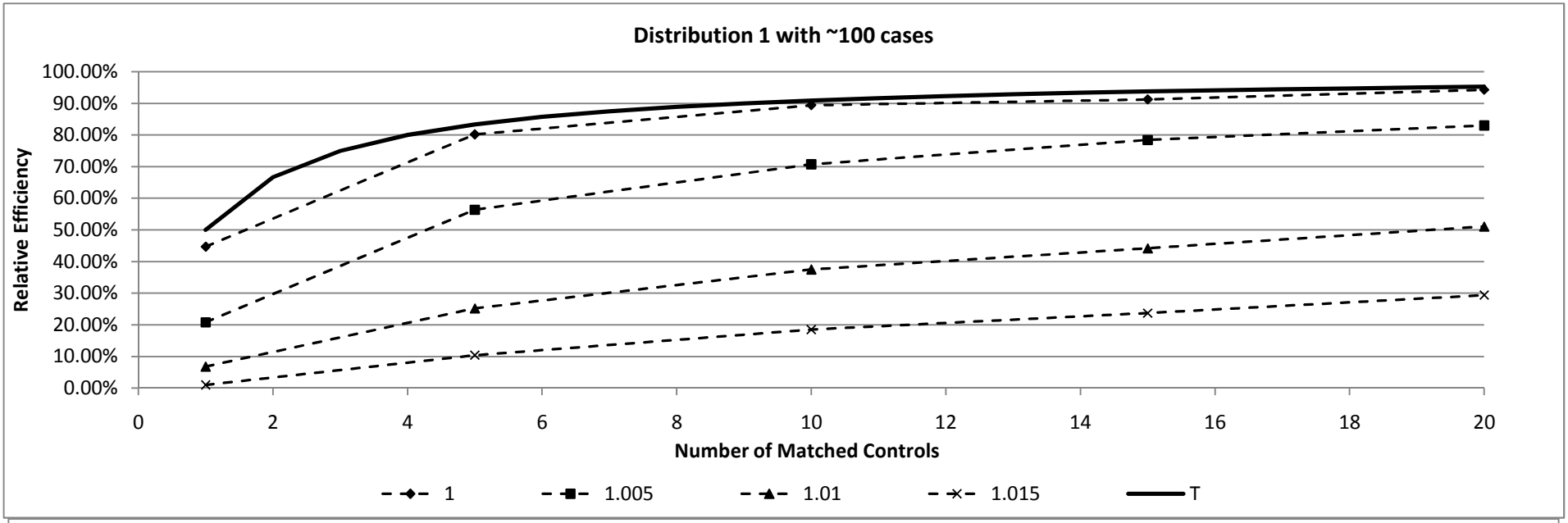
[a] Bias is the true exposure-response relationship (log of the true hazard ratio) minus the mean of the exposure-response parameter estimates

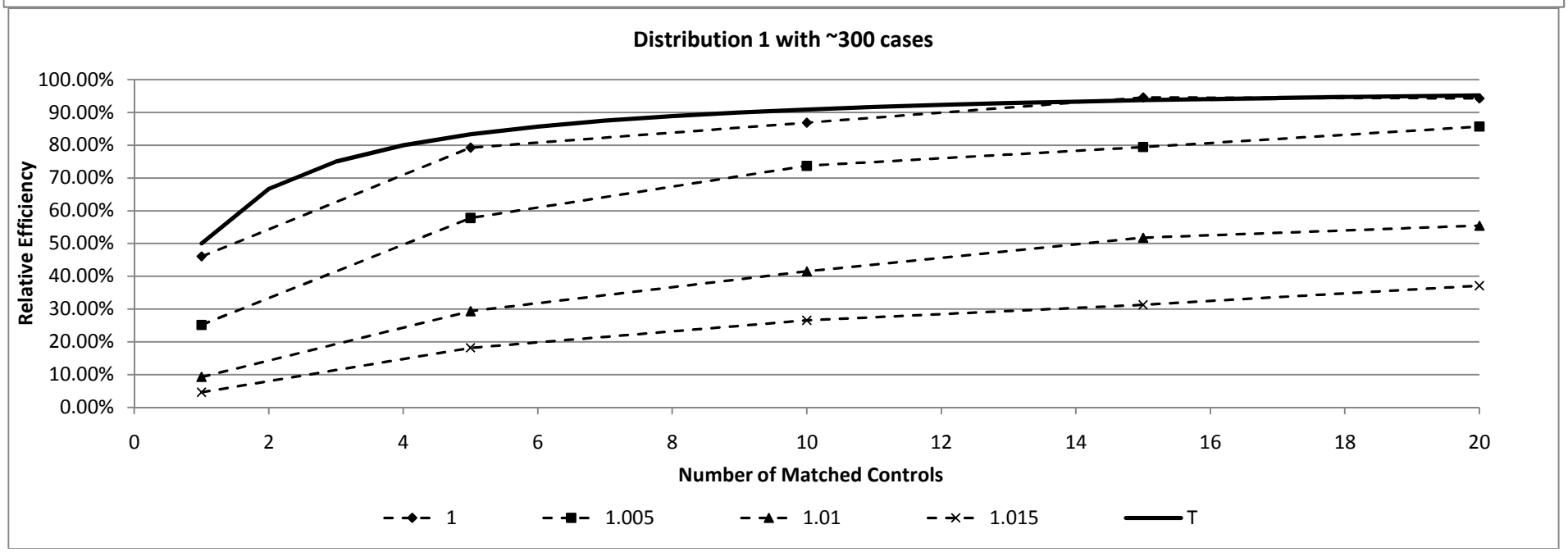
[b] Relative efficiency is the variance of the exposure-response parameter estimates for the full cohort divided by the variance of the exposure-response parameter estimates for the sampled cohort times 100%

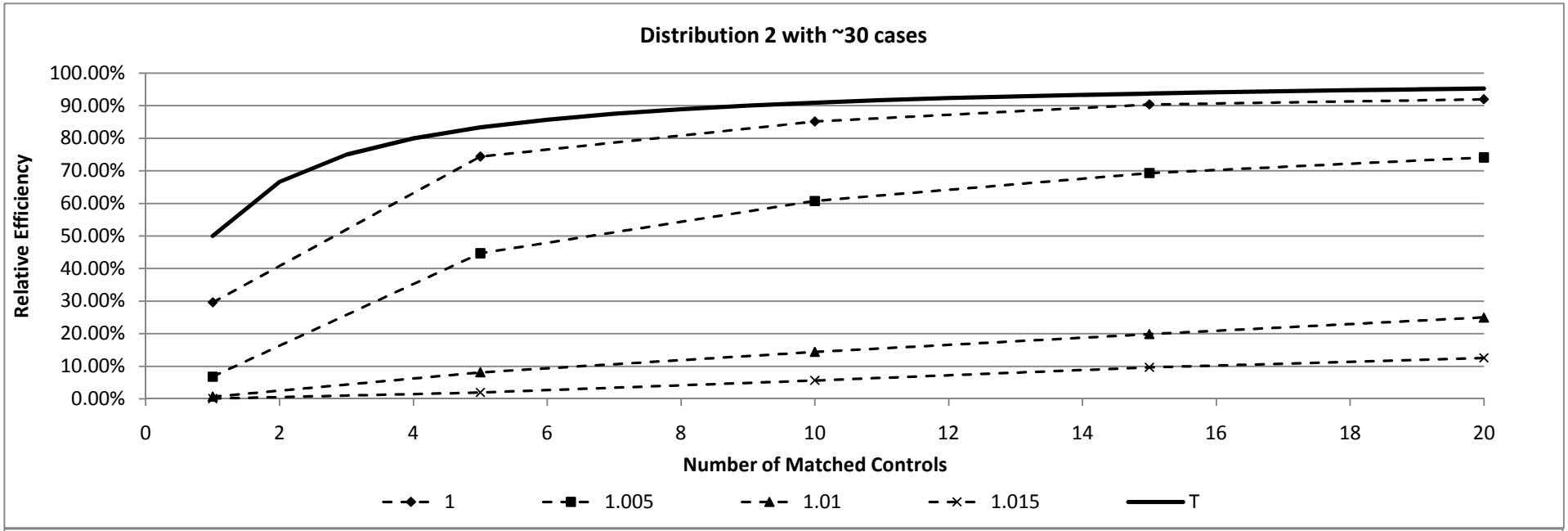
Figures 2-2: relative efficiency vs. number of matched controls by true hazard ratio. The line T represents the graph of the equation $\frac{m}{m+1}$ which is the theoretical relative efficiency when $\beta = 0$

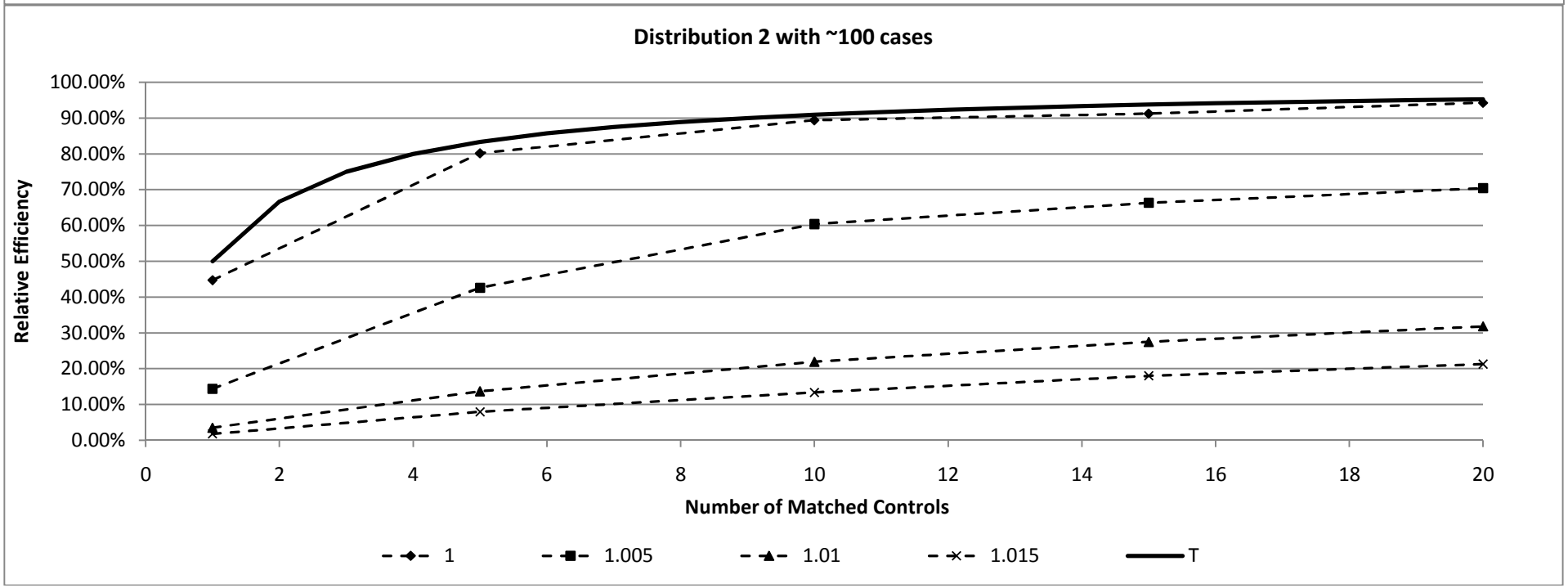


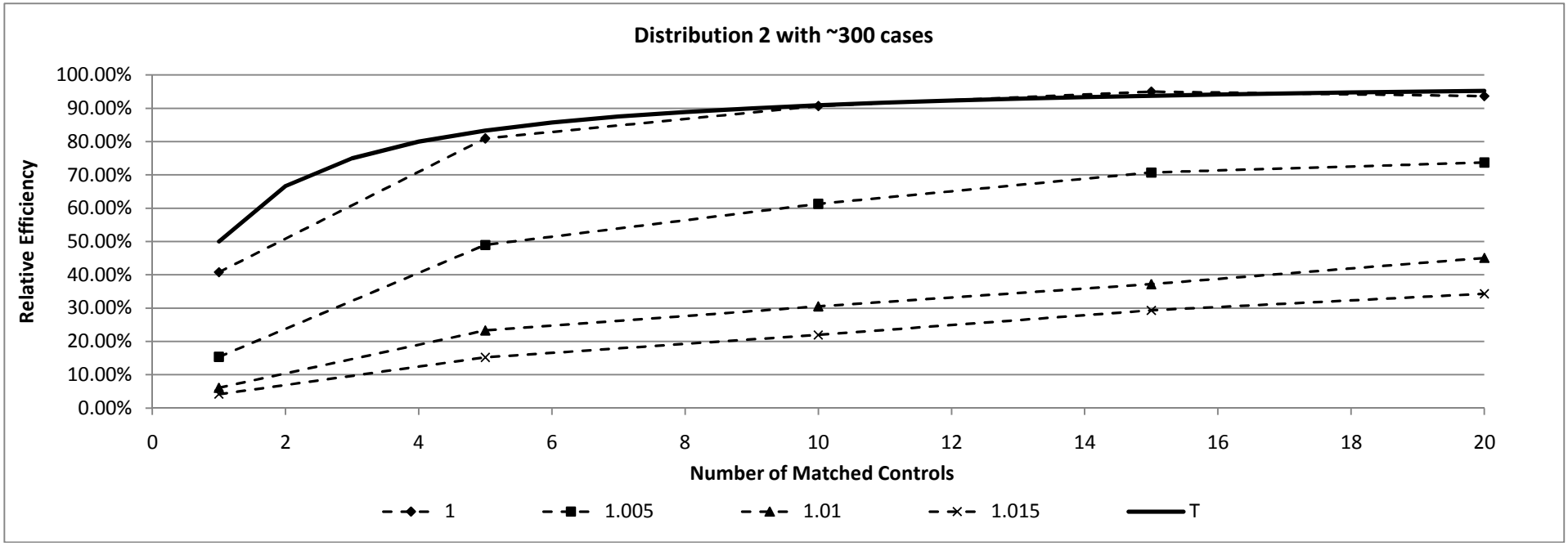

Distribution 0 with ~300 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T
Distribution 1 with ~30 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T


Distribution 1 with ~100 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T

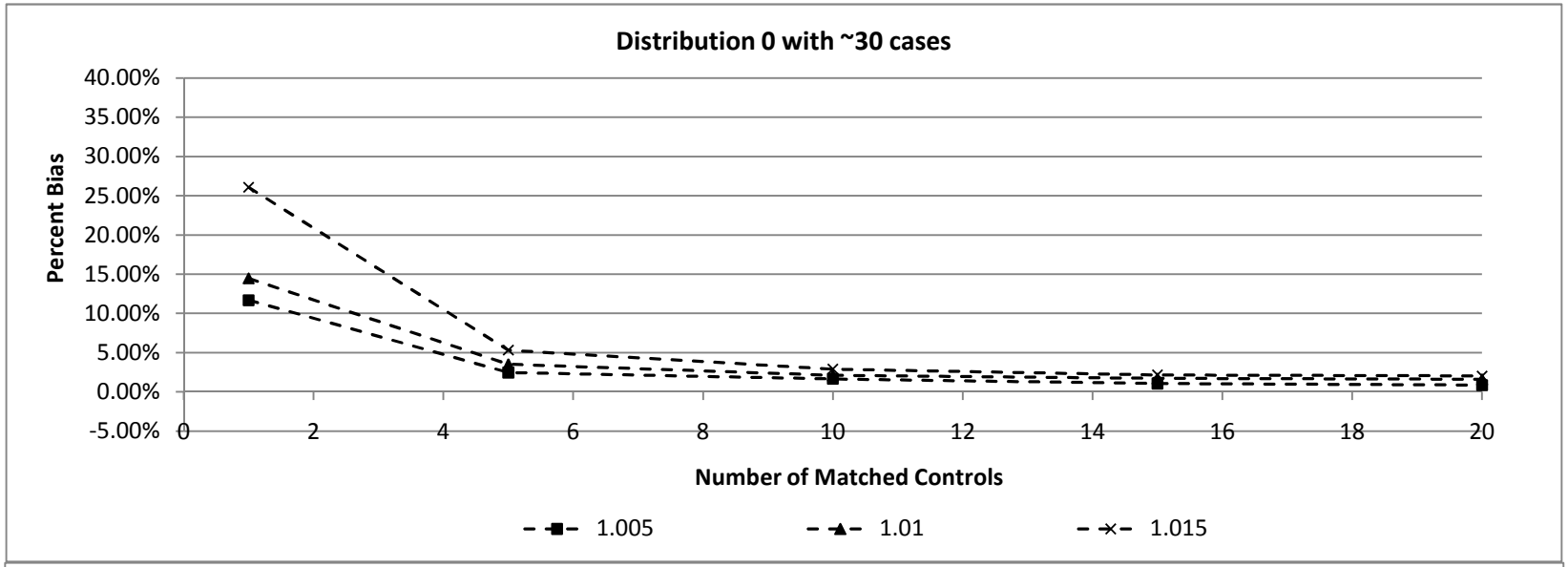

Distribution 1 with ~300 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T

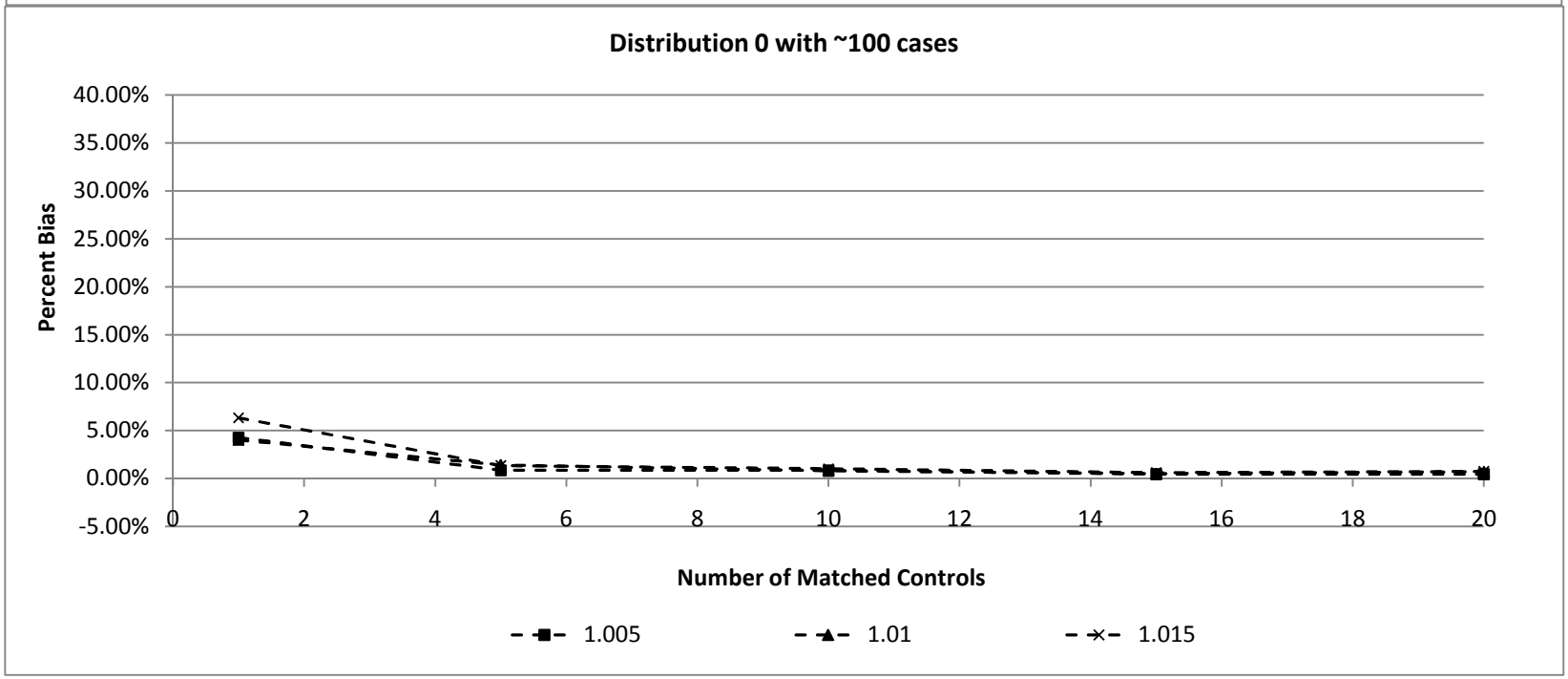

Distribution 2 with ~30 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T

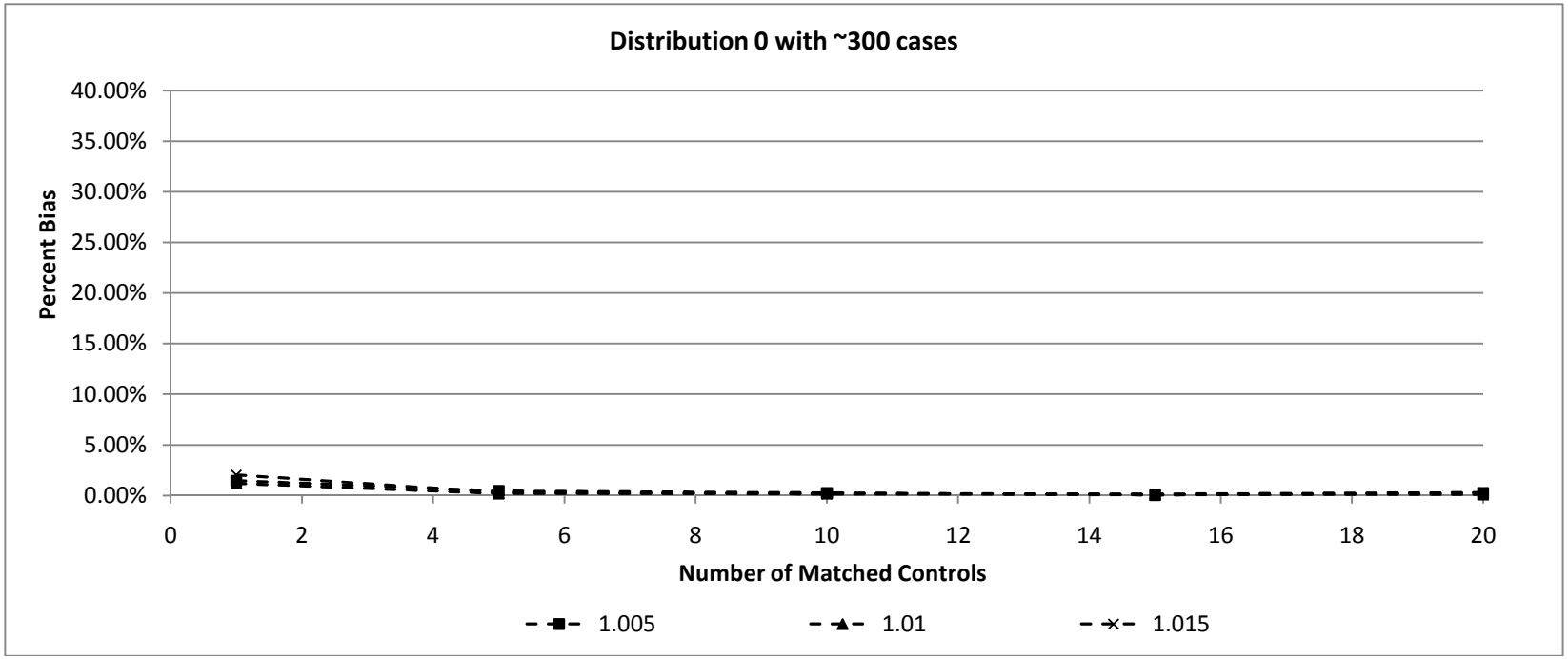

Distribution 2 with ~100 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T

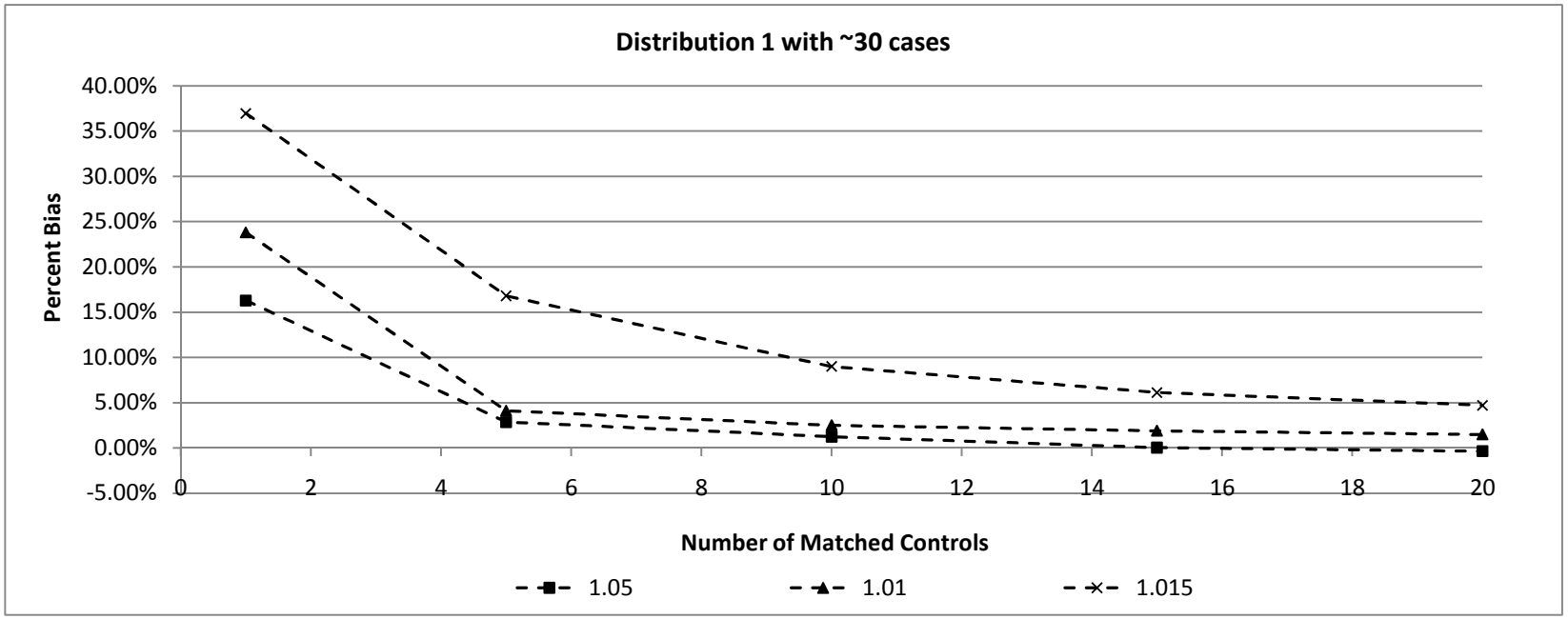

Distribution 2 with ~300 cases
Relative Efficiency
100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1
1.005
1.01
1.015
T

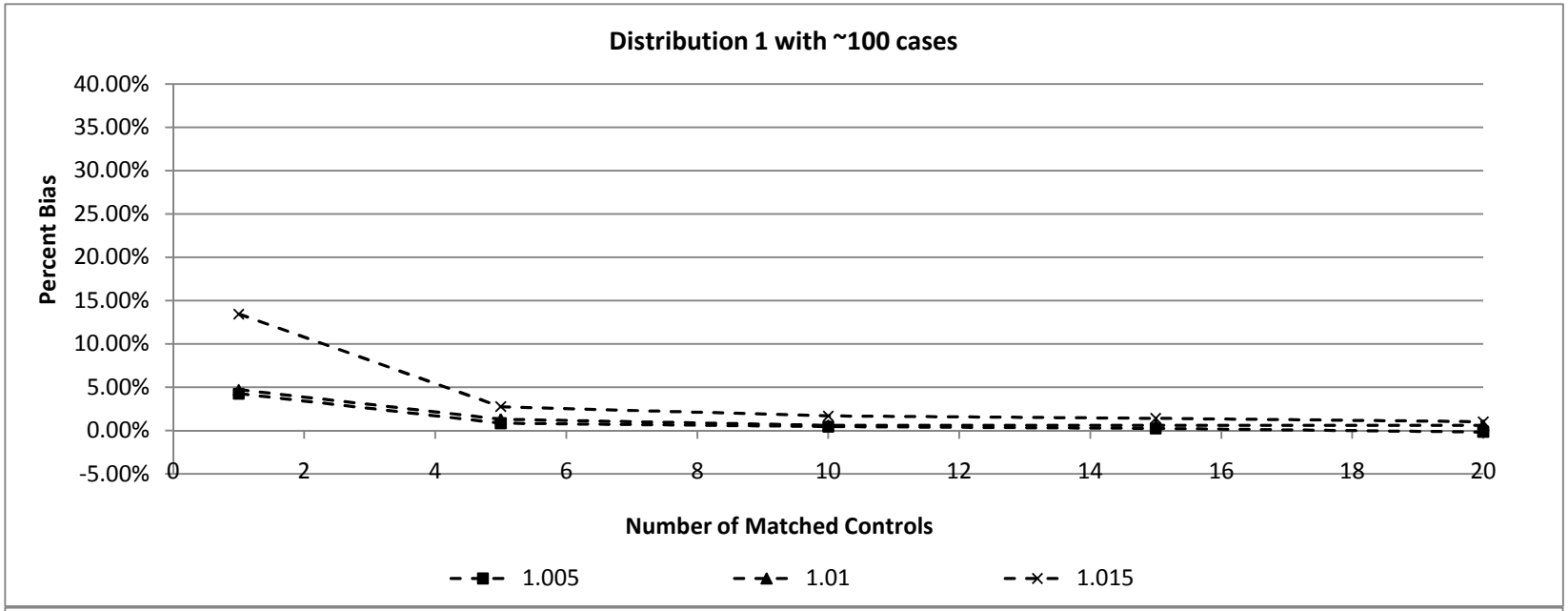
Figures 2-3: Percent Bias vs. Number of matched controls by true hazard ratio

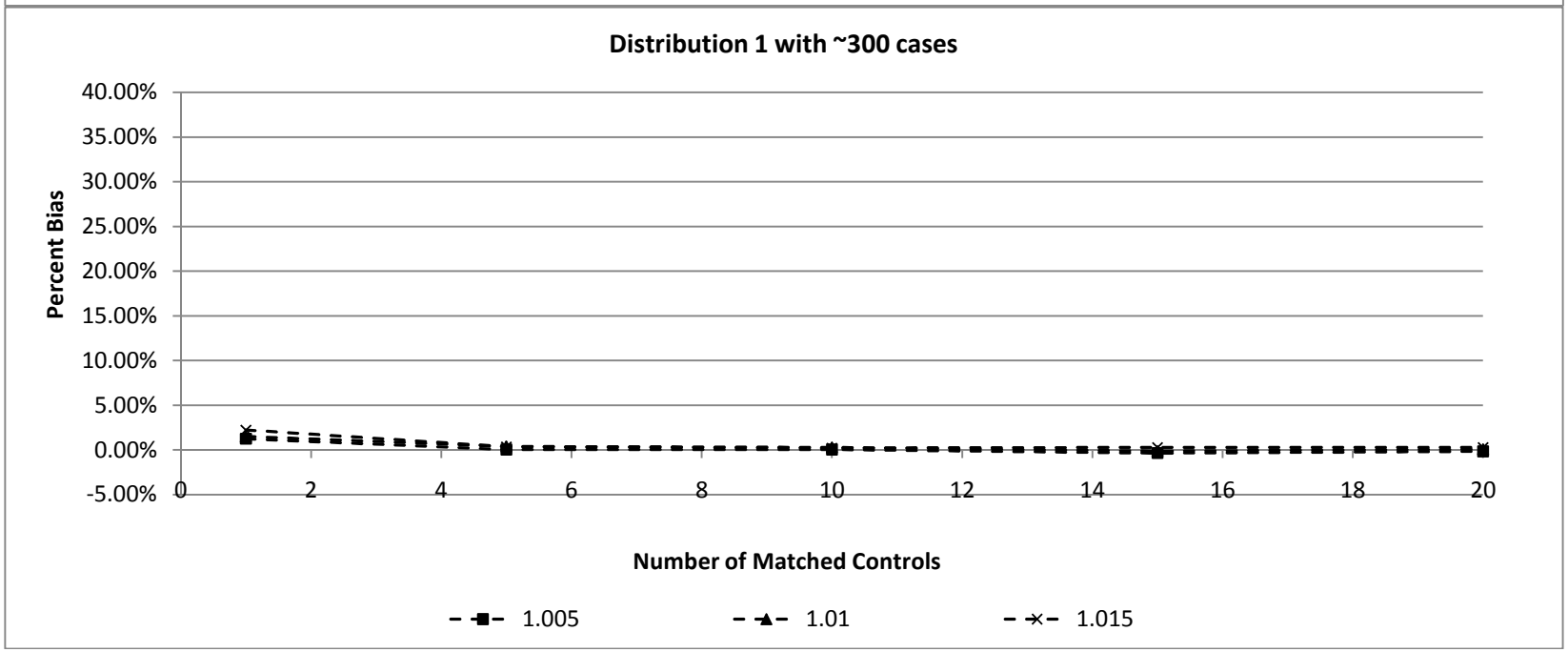


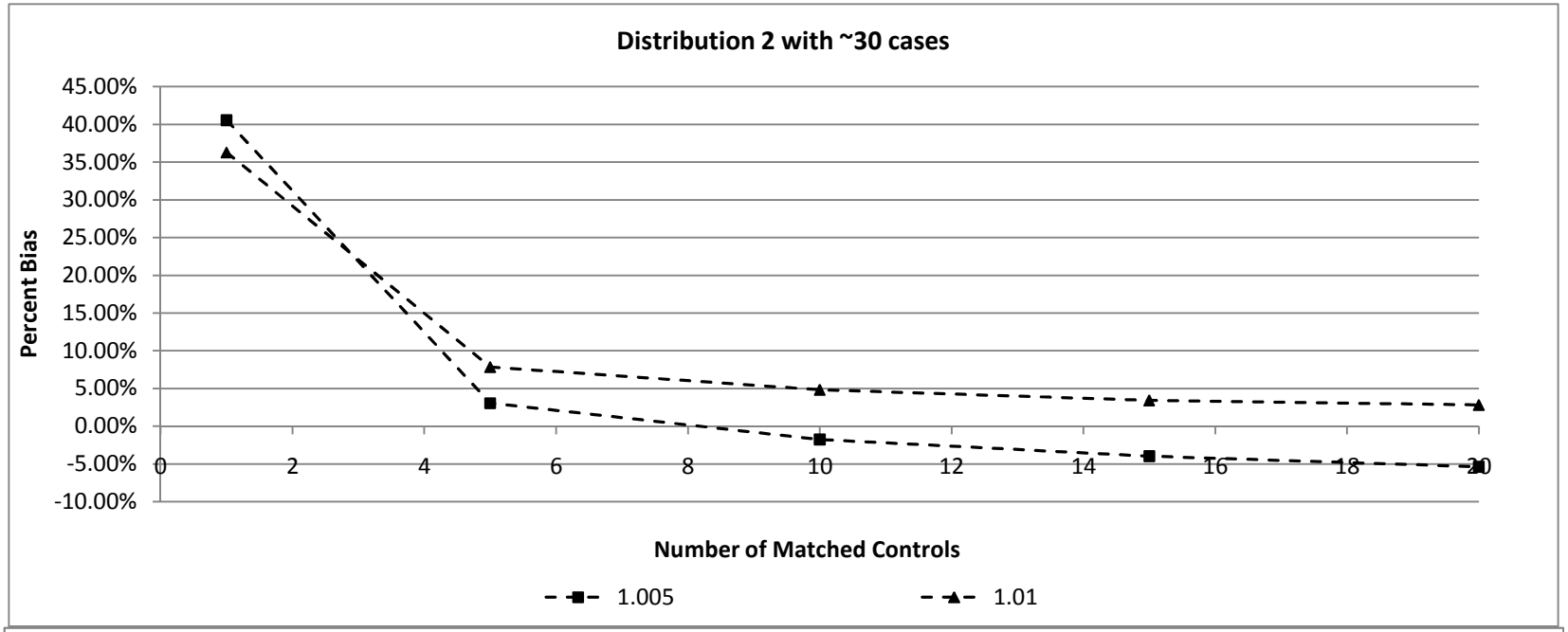


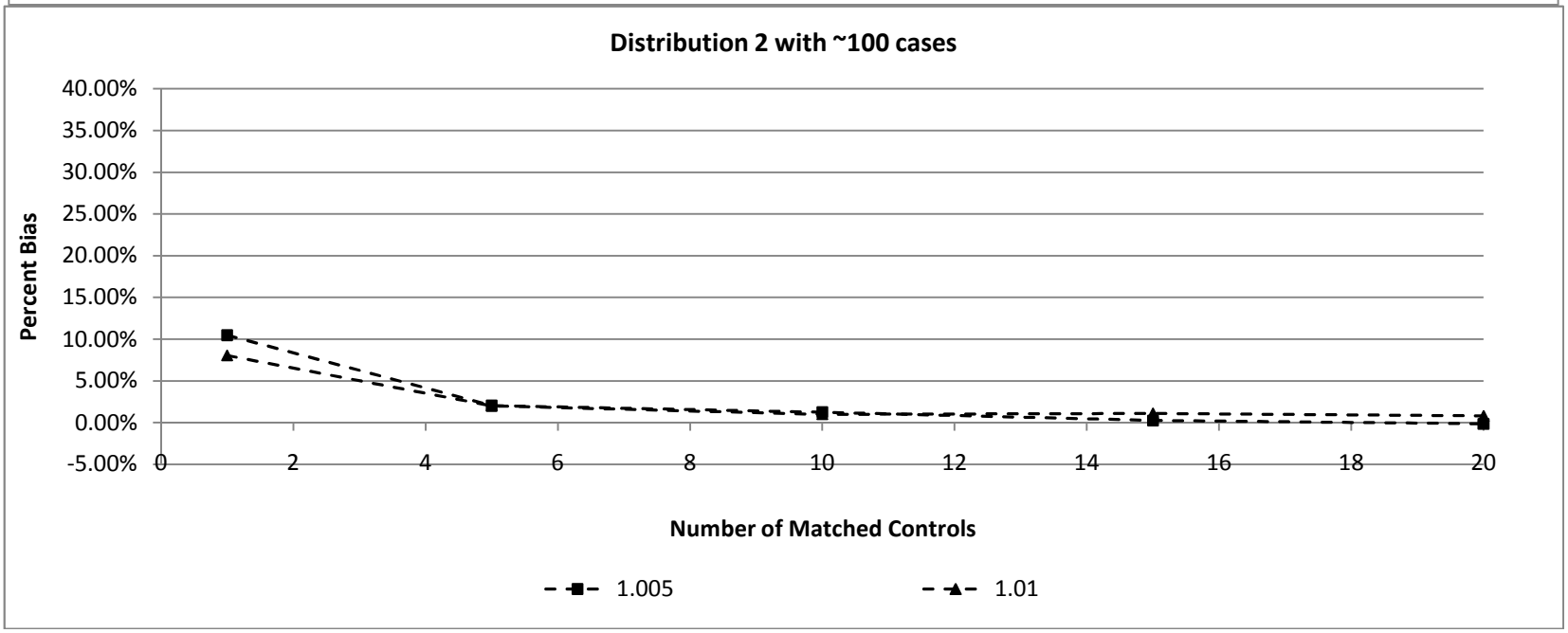

Distribution 0 with ~300 cases
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Percent Bias
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01
1.015

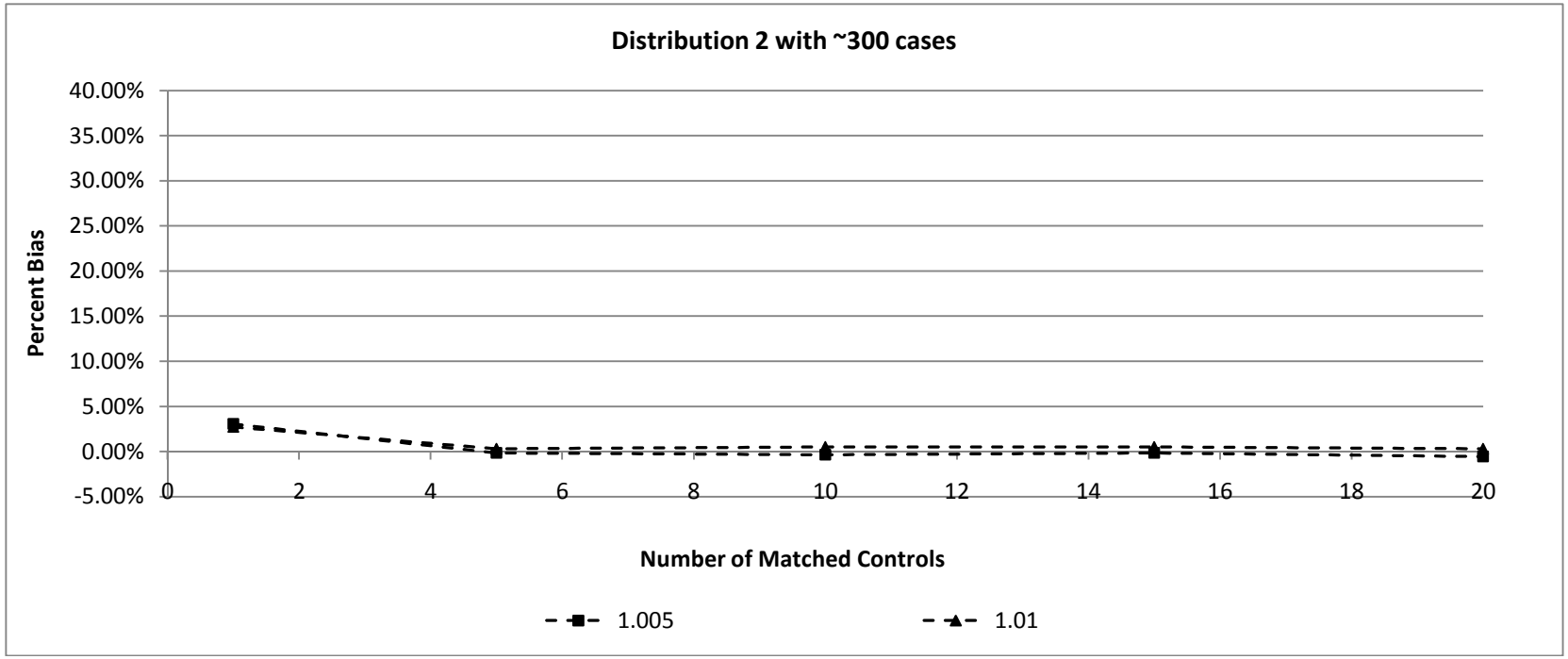

Distribution 1 with ~30 cases
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
Percent Bias
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.05
1.01
1.015


Distribution 1 with ~100 cases
Percent Bias
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01
1.015


Distribution 1 with ~300 cases
Percent Bias
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01
1.015


Distribution 2 with ~30 cases
45.00%
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
-10.00%
Percent Bias
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01


Distribution 2 with ~100 cases
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
Percent Bias
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01


Distribution 2 with ~300 cases
Percent Bias
40.00%
35.00%
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
-5.00%
0
2
4
6
8
10
12
14
16
18
20
Number of Matched Controls
1.005
1.01

Section 2 Figures and Graphs

Tables 2-6: Summary statistics for within cohort percent bias for each simulation scenario. There were 100 cohorts simulated for each scenario and therefore, there were 100 estimates of the within cohort bias. These tables give summary statistics for those 100 estimates.

a)

Distribution 0

True Hazard Ratio	Match	~30 cases					~100 cases					~300 cases				
		Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]
1.005	1:1	9.48%	8.21%	-21.42%	52.89%	99.18%	3.62%	3.78%	-13.68%	17.08%	98.98%	0.72%	1.02%	-7.88%	14.24%	98.75%
	1:5	1.63%	1.77%	-15.87%	14.42%	99.99%	0.91%	1.40%	-7.16%	6.76%	99.99%	0.16%	0.29%	-4.23%	5.09%	99.99%
	1:10	0.95%	1.14%	-9.64%	9.52%	100.00%	0.50%	0.70%	-4.18%	3.76%	100.00%	0.13%	0.19%	-2.91%	2.64%	100.00%
	1:15	0.57%	0.75%	-7.45%	5.67%	100.00%	0.30%	0.33%	-2.99%	2.85%	100.00%	0.08%	0.09%	-2.09%	1.88%	100.00%
	1:20	0.40%	0.64%	-5.96%	4.29%	100.00%	0.25%	0.31%	-2.15%	2.02%	100.00%	0.07%	0.09%	-1.64%	1.40%	100.00%
1.01	1:1	13.18%	12.98%	-22.34%	66.94%	98.28%	2.90%	3.25%	-12.83%	23.66%	97.40%	1.16%	0.87%	-10.20%	8.91%	97.29%
	1:5	2.32%	1.87%	-10.11%	22.55%	99.74%	0.39%	0.75%	-9.60%	10.07%	99.44%	0.32%	0.30%	-5.55%	5.12%	99.53%
	1:10	1.27%	0.87%	-7.92%	16.29%	99.95%	0.11%	0.36%	-7.74%	6.00%	99.91%	0.21%	0.17%	-3.89%	3.71%	99.96%
	1:15	0.85%	0.64%	-6.64%	12.03%	99.99%	0.05%	0.08%	-6.82%	4.66%	99.98%	0.18%	0.17%	-3.03%	2.81%	99.99%
	1:20	0.71%	0.43%	-5.04%	10.02%	100.00%	0.05%	0.33%	-5.43%	3.56%	100.00%	0.11%	0.08%	-2.70%	2.21%	100.00%
1.015	1:1	29.01%	26.84%	-8.79%	77.18%	97.25%	7.05%	6.05%	-8.13%	25.68%	97.57%	1.42%	1.89%	-8.21%	12.22%	96.49%
	1:5	5.01%	4.26%	-10.68%	27.84%	98.91%	2.09%	2.09%	-8.62%	12.04%	98.50%	0.41%	0.43%	-6.47%	7.52%	98.52%
	1:10	2.72%	2.85%	-8.11%	18.27%	99.49%	1.34%	1.65%	-6.87%	9.53%	99.34%	0.37%	0.37%	-4.76%	5.38%	99.53%
	1:15	2.02%	1.92%	-7.13%	16.34%	99.75%	1.03%	1.64%	-5.26%	7.76%	99.70%	0.26%	0.32%	-4.11%	4.53%	99.77%
	1:20	1.62%	1.59%	-5.98%	14.32%	99.85%	0.84%	1.08%	-5.01%	6.67%	99.87%	0.23%	0.22%	-3.45%	3.86%	99.91%

[a] – Percent of the 50,000 confidence intervals calculated (100 cohorts sampled 500 times) that contained the full cohort estimate

b)

		Distribution 1														
		~30 cases					~100 cases					~300 cases				
True Hazard Ratio	Match	Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]
1.005	1:1	20.51%	18.35%	-18.54%	85.08%	98.88%	4.45%	4.26%	-19.44%	23.12%	98.01%	2.53%	2.71%	-9.41%	15.40%	97.49%
	1:5	5.54%	6.06%	-11.51%	29.54%	99.80%	1.39%	1.64%	-11.52%	11.42%	99.71%	0.96%	1.00%	-6.11%	8.17%	99.72%
	1:10	3.28%	3.81%	-6.94%	18.70%	99.96%	0.83%	0.95%	-7.70%	7.18%	99.95%	0.56%	0.53%	-4.35%	5.84%	99.97%
	1:15	2.47%	2.76%	-5.73%	12.64%	99.98%	0.57%	0.65%	-5.86%	5.03%	99.99%	0.45%	0.53%	-3.64%	4.25%	99.99%
	1:20	1.92%	2.40%	-4.71%	9.00%	100.00%	0.45%	0.46%	-4.73%	4.10%	100.00%	0.36%	0.41%	-2.82%	3.68%	100.00%
1.01	1:1	22.08%	16.97%	-18.89%	79.30%	97.11%	4.52%	3.86%	-13.83%	23.71%	96.35%	1.31%	0.61%	-6.70%	10.49%	96.45%
	1:5	4.75%	3.66%	-15.38%	30.54%	98.46%	0.60%	0.11%	-9.55%	11.43%	98.05%	0.27%	0.33%	-5.90%	6.39%	98.46%
	1:10	2.97%	2.89%	-11.48%	21.20%	99.08%	0.21%	-0.10%	-7.08%	8.92%	98.99%	0.17%	0.05%	-4.60%	4.89%	99.32%
	1:15	2.26%	2.09%	-8.84%	17.75%	99.47%	0.10%	-0.24%	-5.69%	7.42%	99.38%	0.11%	-0.01%	-3.93%	4.23%	99.63%
	1:20	1.89%	1.81%	-7.83%	14.48%	99.66%	-0.05%	-0.31%	-4.79%	6.56%	99.68%	0.09%	0.07%	-3.40%	3.82%	99.81%
1.015	1:1	37.75%	40.01%	-19.60%	79.76%	95.24%	10.58%	10.51%	-14.89%	47.69%	95.68%	2.87%	3.46%	-8.51%	14.92%	96.29%
	1:5	14.88%	10.42%	-20.11%	72.20%	97.07%	0.95%	0.47%	-10.72%	14.84%	96.57%	1.17%	1.10%	-6.64%	7.38%	97.54%
	1:10	6.76%	5.54%	-18.69%	47.56%	97.74%	0.35%	0.30%	-8.40%	10.91%	97.67%	0.86%	0.86%	-5.13%	5.82%	98.57%
	1:15	4.41%	3.59%	-16.66%	38.96%	98.26%	0.06%	-0.26%	-8.04%	8.25%	98.21%	0.73%	0.74%	-4.49%	5.58%	99.09%
	1:20	3.24%	2.94%	-14.79%	29.62%	98.54%	0.00%	-0.31%	-7.06%	8.50%	98.70%	0.64%	0.59%	-3.83%	5.40%	99.33%

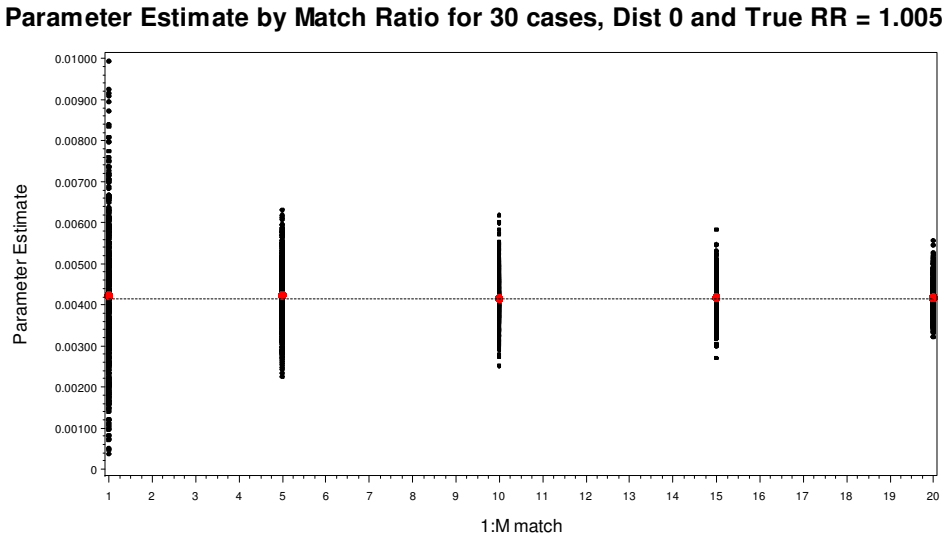
[a] – Percent of the 50,000 confidence intervals calculated (100 cohorts sampled 500 times) that contained the full cohort estimate

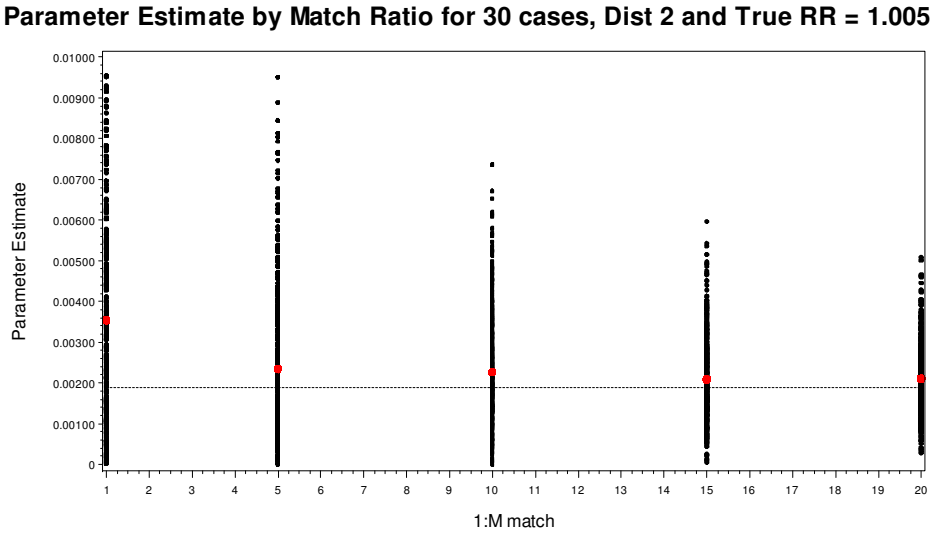
c)

		Distribution 2														
		~30 cases					~100 cases					~300 cases				
True Hazard Ratio	Match	Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]	Mean	Median	Min	Max	95% CI Coverage[a]
1.005	1:1	56.40%	44.47%	-25.96%	220.19%	98.66%	12.60%	11.13%	-20.80%	85.75%	97.44%	4.83%	3.23%	-13.62%	32.64%	97.04%
	1:5	17.57%	15.91%	-24.78%	83.53%	99.18%	3.90%	1.87%	-15.04%	32.29%	99.01%	1.50%	0.90%	-11.84%	15.57%	99.11%
	1:10	11.41%	10.78%	-19.97%	58.97%	99.46%	2.46%	0.80%	-11.23%	21.91%	99.56%	0.92%	0.07%	-8.44%	10.40%	99.68%
	1:15	8.57%	8.46%	-16.82%	42.24%	99.70%	1.83%	0.40%	-8.91%	17.53%	99.78%	0.68%	0.08%	-7.34%	9.10%	99.91%
	1:20	7.24%	7.08%	-12.11%	36.90%	99.79%	1.52%	0.16%	-7.22%	15.10%	99.86%	0.53%	0.26%	-6.12%	7.54%	99.96%
1.01	1:1	37.38%	30.78%	-20.34%	277.00%	95.93%	6.70%	6.28%	-16.23%	30.70%	96.49%	3.14%	3.03%	-12.89%	20.33%	95.82%
	1:5	7.66%	7.94%	-20.45%	48.11%	96.57%	1.76%	1.98%	-13.99%	16.87%	97.67%	1.26%	0.77%	-10.30%	15.07%	97.08%
	1:10	4.48%	4.46%	-20.19%	36.10%	97.18%	1.19%	1.04%	-10.86%	13.61%	98.29%	0.86%	0.93%	-8.46%	12.23%	97.98%
	1:15	3.37%	3.65%	-18.73%	29.58%	97.58%	1.05%	0.97%	-9.15%	13.91%	98.67%	0.73%	0.80%	-7.28%	10.59%	98.46%
	1:20	2.74%	2.84%	-17.29%	28.36%	97.83%	0.95%	0.69%	-8.36%	12.15%	98.94%	0.65%	0.65%	-6.57%	9.64%	98.90%
1.015	1:1	63.83%	55.04%	-5.46%	205.89%	95.84%	8.51%	7.40%	-8.28%	48.05%	96.31%	1.91%	1.65%	-10.74%	13.59%	96.35%
	1:5	13.15%	8.18%	-16.11%	90.40%	96.92%	1.83%	0.58%	-9.84%	18.92%	96.77%	0.63%	0.13%	-6.88%	8.57%	97.59%
	1:10	7.93%	5.53%	-15.35%	94.94%	97.05%	0.81%	0.04%	-8.68%	15.63%	97.26%	0.45%	0.04%	-5.18%	7.60%	98.18%
	1:15	6.09%	3.45%	-15.00%	80.27%	97.16%	0.48%	0.00%	-7.98%	13.36%	97.86%	0.40%	0.09%	-4.13%	6.65%	98.79%
	1:20	5.08%	3.04%	-13.98%	69.54%	97.41%	0.33%	0.02%	-7.61%	11.61%	98.14%	0.39%	0.18%	-3.64%	6.11%	99.01%

[a] – Percent of the 50,000 confidence intervals calculated (100 cohorts sampled 500 times) that contained the full cohort estimate

Figures 2-4: Plot of Parameter Estimates by Match Ratio for example cohorts. Each black dot represents one of the 500 estimates for each match ratio, the red dot is the mean of the 500 estimates, and the dotted line is the value from the analysis of the full cohort.





Tables 2-7: Summary of the within cohort and between cohort sum of squares for each simulation as described by Equations 2.3 – 2.7

a)

Distribution 0

		~30 cases					~ 100 cases					~ 300 cases				
True Hazard Ratio	Match	Within SS	Between SS	Total SS	Within %	Between %	Within SS	Between SS	Total SS	Within %	Between %	Within SS	Between SS	Total SS	Within %	Between %
1	1:1	0.2038	0.1622	0.3660	55.69%	44.31%	0.0319	0.0296	0.0616	51.88%	48.12%	0.0104	0.0106	0.0210	49.60%	50.40%
	1:5	0.2229	0.2673	0.4902	45.47%	54.53%	0.0061	0.0283	0.0344	17.82%	82.18%	0.0020	0.0104	0.0124	16.34%	83.66%
	1:10	0.0434	0.2014	0.2448	17.72%	82.29%	0.0030	0.0281	0.0311	9.67%	90.33%	0.0010	0.0100	0.0110	9.13%	90.87%
	1:15	0.0176	0.1774	0.1950	9.02%	90.98%	0.0020	0.0281	0.0301	6.64%	93.36%	0.0007	0.0102	0.0109	6.11%	93.89%
	1:20	0.0174	0.1765	0.1939	8.98%	91.02%	0.0015	0.0281	0.0296	5.05%	94.96%	0.0005	0.0102	0.0107	4.65%	95.35%
1.005	1:1	0.2320	0.1514	0.3834	60.51%	39.49%	0.0562	0.0452	0.1014	55.45%	44.55%	0.0168	0.0137	0.0305	54.99%	45.01%
	1:5	0.0336	0.1123	0.1459	23.01%	76.99%	0.0108	0.0381	0.0489	22.10%	77.91%	0.0034	0.0121	0.0154	21.81%	78.19%
	1:10	0.0169	0.1078	0.1247	13.55%	86.45%	0.0056	0.0372	0.0428	13.00%	87.00%	0.0018	0.0119	0.0136	12.92%	87.08%
	1:15	0.0114	0.1059	0.1172	9.69%	90.31%	0.0038	0.0368	0.0406	9.28%	90.73%	0.0012	0.0119	0.0131	9.12%	90.89%
	1:20	0.0084	0.1046	0.1130	7.45%	92.55%	0.0028	0.0366	0.0394	7.11%	92.89%	0.0009	0.0119	0.0128	7.00%	93.00%
1.01	1:1	1.8962	0.3210	2.2172	85.52%	14.48%	0.1613	0.0723	0.2337	69.04%	30.96%	0.0449	0.0225	0.0674	66.64%	33.36%
	1:5	0.0918	0.1432	0.2350	39.07%	60.93%	0.0287	0.0475	0.0762	37.64%	62.36%	0.0091	0.0172	0.0263	34.75%	65.25%
	1:10	0.0461	0.1264	0.1726	26.74%	73.26%	0.0154	0.0429	0.0584	26.45%	73.55%	0.0049	0.0167	0.0216	22.70%	77.30%
	1:15	0.0315	0.1193	0.1507	20.88%	79.13%	0.0107	0.0409	0.0516	20.72%	79.28%	0.0035	0.0162	0.0197	17.60%	82.40%
	1:20	0.0245	0.1153	0.1397	17.50%	82.50%	0.0083	0.0403	0.0487	17.12%	82.88%	0.0027	0.0158	0.0185	14.52%	85.48%
1.015	1:1	12.0315	0.7675	12.7990	94.00%	6.00%	0.5829	0.1404	0.7233	80.58%	19.42%	0.1255	0.0362	0.1618	77.60%	22.40%
	1:5	0.3850	0.2560	0.6410	60.06%	39.94%	0.0778	0.0724	0.1502	51.81%	48.19%	0.0237	0.0238	0.0475	49.87%	50.13%
	1:10	0.1595	0.1901	0.3496	45.63%	54.37%	0.0402	0.0584	0.0986	40.76%	59.25%	0.0127	0.0186	0.0313	40.55%	59.45%
	1:15	0.1046	0.1664	0.2710	38.62%	61.39%	0.0280	0.0527	0.0807	34.71%	65.30%	0.0090	0.0180	0.0270	33.39%	66.61%
	1:20	0.0801	0.1536	0.2337	34.28%	65.72%	0.0220	0.0499	0.0718	30.61%	69.40%	0.0071	0.0169	0.0241	29.64%	70.36%

b)

Distribution 1

True Hazard Ratio	Match	~30 cases Within SS	~30 cases Between SS	~30 cases Total SS	~30 cases Within %	~30 cases Between %	~ 100 cases Within SS	~ 100 cases Between SS	~ 100 cases Total SS	~ 100 cases Within %	~ 100 cases Between %	~ 300 cases Within SS	~ 300 cases Between SS	~ 300 cases Total SS	~ 300 cases Within %	~ 300 cases Between %
1	1:1	0.2244	0.1755	0.4000	56.11%	43.89%	0.0532	0.0557	0.1088	48.85%	51.15%	0.0166	0.0155	0.0321	51.80%	48.20%
	1:5	0.0330	0.1400	0.1730	19.09%	80.91%	0.0099	0.0526	0.0625	15.81%	84.19%	0.0032	0.0150	0.0182	17.48%	82.52%
	1:10	0.0157	0.1350	0.1507	10.44%	89.57%	0.0048	0.0521	0.0568	8.39%	91.62%	0.0016	0.0148	0.0163	9.59%	90.41%
	1:15	0.0101	0.1332	0.1433	7.07%	92.93%	0.0032	0.0519	0.0551	5.72%	94.28%	0.0011	0.0147	0.0158	6.66%	93.34%
	1:20	0.0076	0.1319	0.1395	5.45%	94.55%	0.0024	0.0518	0.0541	4.39%	95.61%	0.0008	0.0143	0.0151	5.16%	94.84%
1.005	1:1	0.5362	0.1747	0.7109	75.42%	24.58%	0.0758	0.0430	0.1188	63.82%	36.18%	0.0239	0.0127	0.0366	65.24%	34.76%
	1:5	0.0641	0.1120	0.1760	36.41%	63.59%	0.0157	0.0332	0.0488	32.07%	67.93%	0.0051	0.0100	0.0150	33.71%	66.29%
	1:10	0.0325	0.1035	0.1360	23.87%	76.13%	0.0084	0.0313	0.0397	21.19%	78.81%	0.0028	0.0093	0.0121	22.96%	77.04%
	1:15	0.0222	0.1005	0.1227	18.08%	81.92%	0.0058	0.0307	0.0365	15.92%	84.08%	0.0019	0.0091	0.0110	17.23%	82.77%
	1:20	0.0171	0.0991	0.1163	14.73%	85.27%	0.0044	0.0303	0.0347	12.73%	87.27%	0.0015	0.0088	0.0102	14.47%	85.53%
1.01	1:1	3.7882	0.3303	4.1185	91.98%	8.02%	0.2015	0.0426	0.2441	82.56%	17.44%	0.0533	0.0195	0.0728	73.26%	26.74%
	1:5	0.1356	0.0840	0.2196	61.74%	38.27%	0.0340	0.0219	0.0558	60.85%	39.15%	0.0111	0.0123	0.0234	47.53%	52.47%
	1:10	0.0684	0.0629	0.1313	52.09%	47.91%	0.0187	0.0187	0.0374	50.02%	49.98%	0.0062	0.0101	0.0164	38.04%	61.96%
	1:15	0.0473	0.0550	0.1022	46.24%	53.76%	0.0135	0.0173	0.0307	43.79%	56.21%	0.0045	0.0092	0.0137	33.00%	67.00%
	1:20	0.0365	0.0505	0.0870	41.98%	58.02%	0.0106	0.0166	0.0272	38.98%	61.02%	0.0036	0.0088	0.0124	29.18%	70.82%
1.015	1:1	15.9885	0.3717	16.3602	97.73%	2.27%	2.7746	0.2613	3.0359	91.39%	8.61%	0.1518	0.0322	0.1840	82.50%	17.50%
	1:5	4.3316	0.6778	5.0094	86.47%	13.53%	0.1237	0.0748	0.1985	62.32%	37.68%	0.0282	0.0171	0.0454	62.25%	37.75%
	1:10	1.5052	0.3340	1.8392	81.84%	18.16%	0.0613	0.0549	0.1162	52.73%	47.27%	0.0151	0.0131	0.0282	53.42%	46.59%
	1:15	0.4013	0.2154	0.6167	65.08%	34.92%	0.0417	0.0472	0.0888	46.90%	53.10%	0.0108	0.0116	0.0224	48.24%	51.76%
	1:20	0.2102	0.1705	0.3808	55.21%	44.79%	0.0328	0.0428	0.0756	43.42%	56.58%	0.0086	0.0108	0.0193	44.23%	55.77%

c)

		Distribution 2														
		~30 cases					~ 100 cases					~ 300 cases				
True Hazard Ratio	Match	Within SS	Between SS	Total SS	Within %	Between %	Within SS	Between SS	Total SS	Within %	Between %	Within SS	Between SS	Total SS	Within %	Between %
1	1:1	1.5065	1.6097	3.1162	48.35%	51.66%	0.2326	0.1790	0.4116	56.51%	43.49%	0.0571	0.0511	0.1082	52.80%	47.21%
	1:5	0.1697	1.0823	1.2520	13.55%	86.45%	0.0361	0.1527	0.1888	19.14%	80.87%	0.0107	0.0481	0.0588	18.15%	81.85%
	1:10	0.0778	1.0029	1.0807	7.20%	92.81%	0.0172	0.1482	0.1655	10.41%	89.59%	0.0052	0.0475	0.0527	9.82%	90.18%
	1:15	0.0495	0.9711	1.0206	4.85%	95.15%	0.0113	0.1466	0.1579	7.13%	92.87%	0.0035	0.0473	0.0507	6.83%	93.17%
	1:20	0.0366	0.9583	0.9950	3.68%	96.32%	0.0083	0.1464	0.1547	5.39%	94.61%	0.0026	0.0471	0.0497	5.20%	94.80%
1.005	1:1	1.9770	0.4906	2.4676	80.12%	19.88%	0.2257	0.0756	0.3014	74.91%	25.09%	0.0719	0.0245	0.0964	74.54%	25.46%
	1:5	0.2154	0.2621	0.4775	45.11%	54.89%	0.0454	0.0573	0.1027	44.22%	55.78%	0.0157	0.0175	0.0332	47.31%	52.69%
	1:10	0.1096	0.2325	0.3421	32.03%	67.98%	0.0251	0.0525	0.0775	32.35%	67.65%	0.0087	0.0159	0.0245	35.32%	64.68%
	1:15	0.0731	0.2189	0.2920	25.03%	74.97%	0.0179	0.0497	0.0676	26.49%	73.51%	0.0061	0.0152	0.0213	28.81%	71.19%
	1:20	0.0564	0.2112	0.2676	21.09%	78.91%	0.0140	0.0484	0.0624	22.49%	77.51%	0.0048	0.0148	0.0196	24.52%	75.48%
1.01	1:1	8.1215	0.9882	9.1097	89.15%	10.85%	0.3756	0.0690	0.4446	84.48%	15.53%	0.1132	0.0271	0.1403	80.69%	19.31%
	1:5	0.3307	0.1240	0.4547	72.73%	27.27%	0.0684	0.0412	0.1096	62.43%	37.57%	0.0246	0.0156	0.0403	61.20%	38.80%
	1:10	0.1572	0.0725	0.2297	68.44%	31.56%	0.0382	0.0342	0.0724	52.74%	47.26%	0.0138	0.0126	0.0264	52.34%	47.66%
	1:15	0.1082	0.0553	0.1635	66.21%	33.80%	0.0279	0.0312	0.0591	47.16%	52.84%	0.0103	0.0114	0.0217	47.41%	52.60%
	1:20	0.0837	0.0485	0.1323	63.31%	36.69%	0.0224	0.0293	0.0518	43.36%	56.64%	0.0083	0.0108	0.0191	43.68%	56.32%
1.015	1:1	44.3054	2.5152	46.8205	94.63%	5.37%	0.8612	0.1406	1.0019	85.96%	14.04%	0.1718	0.0292	0.2010	85.46%	14.54%
	1:5	1.9373	0.5624	2.4996	77.50%	22.50%	0.1261	0.0638	0.1899	66.41%	33.59%	0.0365	0.0204	0.0569	64.21%	35.79%
	1:10	0.6692	0.3959	1.0651	62.83%	37.17%	0.0659	0.0490	0.1150	57.34%	42.66%	0.0203	0.0167	0.0370	54.81%	45.20%
	1:15	0.4441	0.3075	0.7516	59.09%	40.91%	0.0465	0.0430	0.0895	51.93%	48.07%	0.0147	0.0147	0.0294	49.99%	50.02%
	1:20	0.3022	0.2559	0.5581	54.15%	45.86%	0.0370	0.0392	0.0762	48.62%	51.38%	0.0119	0.0141	0.0260	45.88%	54.12%

Table 2-8: Descriptive statistics of the original cumulative exposure and log of cumulative exposure for the Gold Miners data risk-sets

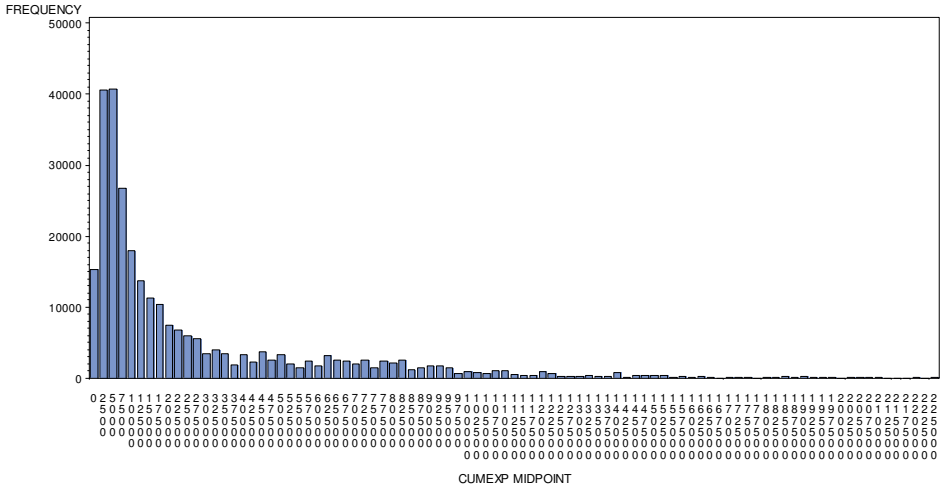
Gold Miner Data						
Exposure Metric	Group	Mean	Variance	Skew	Min	Max
Cumulative Exposure	case	9.39E+04	2.29E+09	0.18	521.97	2.19E+05
	cont	2.62E+04	1.16E+09	2.10	275.48	2.26E+05
Log Cumulative Exposure	case	11.24	0.68	-2.37	6.26	12.30
	cont	9.37	1.83	-0.07	5.62	12.33

Table 2-9: Descriptive statistics of cumulative exposure and log of cumulative exposure after being scaled to match the range of the simulations for the Gold Miners data risk-sets

Gold Miner Adjusted Data						
Exposure Metric	Group	Mean	Variance	Skew	Min	Max
Adjusted Cumulative Exposure	case	290.96	2.22E+04	0.18	0.77	679.99
	cont	80.56	1.12E+04	2.10	0.00	700.00
Adjusted Log Cumulative Exposure	case	628.02	8.51E+03	-2.37	71.46	746.76
	cont	419.19	2.29E+04	-0.07	0.00	750.00

Figure 2-5: Histogram of a) Cumulative Exposure and b) Log of Cumulative Exposure for the Gold Miner data risk-sets

a)



b)

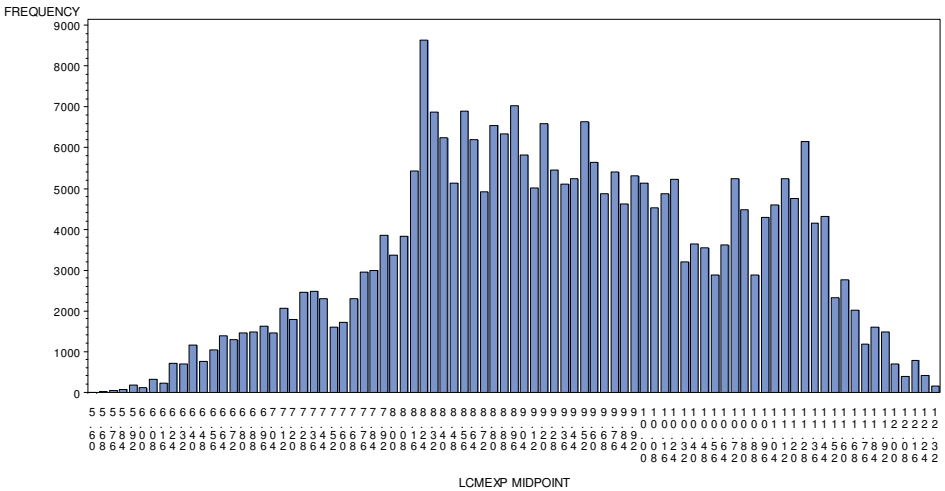


Table 2-10: Results from Cox proportional hazards analysis for the full cohort for each exposure metric. The model was fitted with the original exposure data and on the data where the exposure variables were scaled to have a range of 0 – 750.

Gold Miner Data Analysis for Full Cohort

	Analysis on original data			Analysis on adjusted data		
Exposure Metric	Parameter Estimate	Standard Error	Hazard Ratio (per unit increase)	Parameter Estimate	Standard Error	Hazard Ratio (per unit increase)
Cumulative Exposure	0.00002	1.1935E-06	1.00002	0.00739	0.00038	1.00742
Log Cumulative Exposure	1.55565	0.11007	4.73817	0.01391	0.0009844	1.014007194

Table 2-11: Summary statistics for the matched nested case-control analysis of the Gold Miner adjusted data

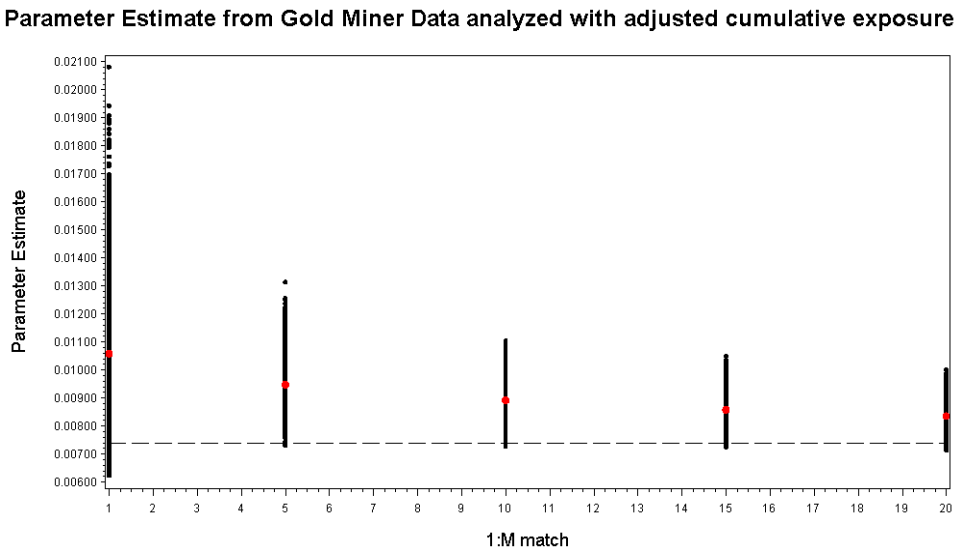
	Analysis on Adjusted Cumulative Exposure			Analysis on Adjusted Log Cumulative Exposure		
Match	Mean	Var	Average of Estimated Var	Mean	Var	Average of Estimated Var
1:1	0.0106	2.86E-06	2.58E-06	0.0124	2.28E-06	3.80E-06
1:5	0.0095	5.15E-07	6.43E-07	0.0134	5.18E-07	1.59E-06
1:10	0.0089	2.73E-07	4.12E-07	0.0137	2.67E-07	1.29E-06
1:15	0.0086	1.79E-07	3.29E-07	0.0138	1.86E-07	1.19E-06
1:20	0.0084	1.30E-07	2.85E-07	0.0138	1.44E-07	1.14E-06

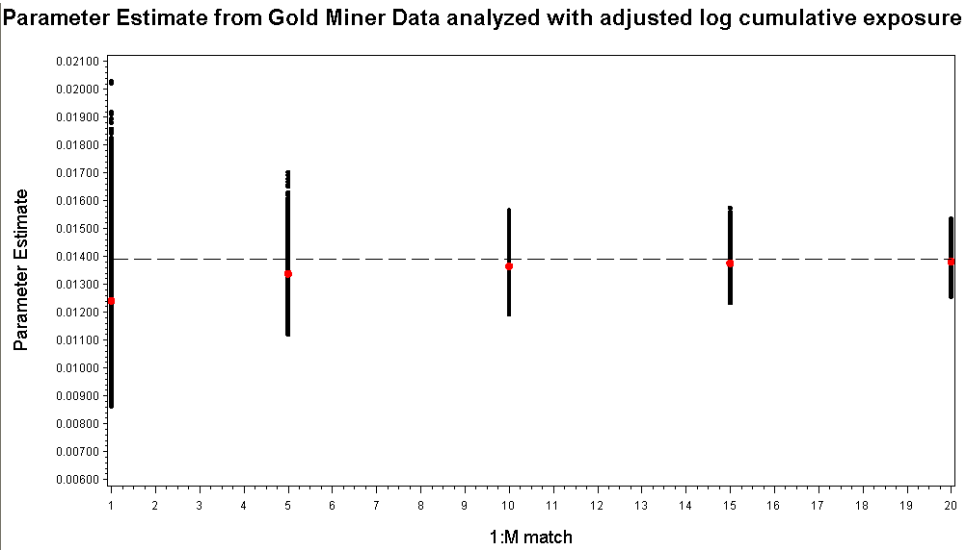
Table 2-12: The within cohort percent bias, efficiency and MSE as compared to the full cohort estimates for matched nested case-control analysis of the Gold Miner adjusted data

	Analysis on Adjusted Cumulative Exposure			Analysis on Adjusted Log Cumulative Exposure		
Match	Percent Bias	Efficiency[a]	MSE	Percent Bias	Efficiency[a]	MSE
1:1	43.30%	5.71%	1.31E-05	-10.73%	25.50%	4.50E-06
1:5	28.29%	22.94%	4.88E-06	-3.76%	60.85%	7.88E-07
1:10	20.72%	35.83%	2.61E-06	-1.82%	75.06%	3.30E-07
1:15	16.18%	44.80%	1.62E-06	-1.03%	81.46%	2.06E-07
1:20	13.09%	51.76%	1.07E-06	-0.60%	85.22%	1.50E-07

[a] - Efficiency is defined as the estimated variance from the full cohort analysis divided by the average of the estimated variances from the nested case-control analysis

Figures 2-6: Plot of the parameter estimates by matched ratio for the Gold Miner adjusted data for a) Cumulative Exposure and b) Log of Cumulative Exposure. Each black dot represents one of the 500 estimates for each match ratio, the red dot is the mean of the 500 estimates, and the dotted line is the value from the analysis of the full cohort





Chapter 3:

Evaluation of Chen's Estimator in Nested Case-Control Study

Section 1

Introduction:

Thomas' partial likelihood estimator (Liddell et al, 1977) is the most common method of analysis for nested case-control studies. Alternative methods have been proposed for the analysis of nested case-control study data in the hopes of improving efficiency (Chen, 2001; Langholz and Goldstein, 1996; Robins et al, 1994; Samuelsen, 1997); however, these methods involve collecting additional data than that needed for Thomas' estimator. For example, the idea of counter-matching has been proposed which involves selecting controls based on knowledge of a surrogate variable related to the covariate of interest (Langholz and Goldstein, 1996).

Chen (2004) proposed an alternative estimator in the analysis of nested case-control study data which involves only the data collected for Thomas' estimator. Chen showed that his estimator is consistent and asymptotically normal. Furthermore, he was able to show that his estimator has a smaller asymptotic variance than that of Thomas' estimator and therefore has a greater asymptotic relative efficiency compared to the full cohort analysis.

Description of Chen's Estimator:

The motivation for Chen's estimator is that heuristically, gains in efficiency could be made by not only including those controls sampled from the risk-set associated with

the failure at time *t*, R_t , but also considering controls sampled from risk-sets R_s where *s* is in some neighborhood of *t.*

In particular, to define Chen's estimator, let R_s^* be the *m* sampled controls from risk-set R_s and let $\varphi(t)$ be an infinitely differentiable nonnegative even function with bounded support. Then, using the counting process notation of the Cox model, define:

$$\bar{N}(s) := \sum_{i=1}^{n} \frac{1}{n} N_i(s) \tag{3.1}$$

$$\varphi_n(t) = \varphi\left(n^{\frac{1}{3}} t\right) \tag{3.2}$$

$$b(t) := \frac{\int_0^\tau \sum_{j \in R_S^*} \varphi_n(t-s) e^{\beta_P' X_j(s)} d\bar{N}(s)}{m \int_0^\tau \varphi_n(t-s) d\bar{N}(s)} \tag{3.3}$$

$$w_i(t) := \frac{mb(t)}{e^{\beta_P' X_i(t)} + mb(t)} \tag{3.4}$$

where β_P is Thomas' estimator from the data. Also, define:

$$S_k(t,\beta) := \int_0^\tau \sum_{j \in R_s^*} \varphi_n(t-s) w_j(s) X_j^k(s) e^{\beta' X_j(s)} \, d\bar{N}(s) \qquad k = 0, 1, 2 \tag{3.5}$$

where the power 2 on covariate $X_j(s)$ indicates the outer product $\otimes 2$. Then Chen's estimator is the solution of:

$$U(\beta) := \sum_{i=1}^{n} \int_0^\tau w_i(t) \left\{ X_i(t) - \frac{S_1(t,\beta)}{S_0(t,\beta)} \right\} dN_i(t) = 0 \tag{3.6}$$

and:

$$\dot{U}(v)|_{v=\hat{\beta}} = \int_0^\tau \left[\frac{S_2(t,\hat{\beta})}{S_0(t,\hat{\beta})} - \left\{ \frac{S_1(t,\hat{\beta})}{S_0(t,\hat{\beta})} \right\}^{\otimes 2} \right] \sum_{i=1}^{n} w_i(t) dN_i(t) \tag{3.7}$$

gives a consistent estimator of the inverse of the covariance matrix.

Chen's estimator asymptotically outperforms that of Thomas' estimator in the sense that its asymptotic variance is smaller than that of Thomas' estimator. However, it is not clear how well it will perform with few or moderate number of cases. Also, Chen's

estimator requires defining a function $\varphi(t)$, and it is not clear how different definitions of $\varphi(t)$ will affect the estimate. The function $\varphi(t)$ has bounded support which guarantees that the asymptotic support of $\varphi_n(t) := \varphi\left(n^{\frac{1}{3}}t\right)$ converges to {0} at a rate of $n^{\frac{1}{3}}$. The support of $\varphi_n(t)$ defines the neighborhood around t whose controls will also be considered.

Objective:

Occupational cohorts were simulated to get a better understanding of the following questions:

1. How does the efficiency and bias of Chen's estimate compare to Thomas' estimate with few and moderate number of cases?
2. How sensitive is Chen's estimate to the definition of $\varphi(t)$?

Method:

The cohorts were simulated identically as they were in Section 1 of Chapter 2. In particular, eight simulation scenarios were performed defined by number of cases in the cohort (~30 or ~100), the exposure-response relationship (hazard ratio per unit exposure = 1.005 or 1.01), and the distribution of the exposure intensity [Distribution 0: Normal(25, 64) - Truncated(0, 50), Distribution 2: Log-Normal(.75, 1) – Truncated(0, 50)]. For each scenario with ~30 cases, 2,000 cohorts were simulated and for the scenarios with ~100 cases, 1,000 cohorts were simulated. All other details are described in Chapter 2 methods.

Analysis:

Risk-sets were created for each cohort, with age as the time scale. Then for each case, 5 controls were randomly sampled from the risk-sets. The full cohort was analyzed using Cox proportional hazards regression (procedure PHREG in SAS) to obtain estimates of the exposure-response parameter. The sampled risk-sets were analyzed using Thomas' estimation procedure (which is identical to conditional logistic regression; procedure PHREG in SAS) and using Chen's procedure. For Chen's procedure, 3 different functions for $\varphi_n(t)$ were defined as follows:

$$\text{Phi 1: } \varphi_n(t) = (1.5^2 - t^2) I(|t| \leq 1.5)$$

$$\text{Phi 2: } \varphi_n(t) = (2.5^2 - t^2) I(|t| \leq 2.5) \qquad (3.7)$$

$$\text{Phi 3: } \varphi_n(t) = (5.5^2 - t^2) I(|t| \leq 5.5)$$

where *t* is measured in years and $I(.)$ is an indicator function. Figures 3-1 graph the above functions. For Phi 1, Phi 2 and Phi 3, each risk-set considers controls sampled for cases with failure age within 1.5, 2.5 and 5.5 year, respectively.

A SAS macro was written to obtain Chen's estimate. Chen's method involves solving equation 3.6 for β. The Newton-Raphson method was used and iterations were stopped when the solution to equation 3.6 evaluated at the current estimate was less than 0.0001. If this condition was not satisfied after 10 iterations, the estimate was said to have not converged.

For each scenario, 2,000 and 1,000 estimates of the exposure-response parameter were obtained for the analysis of full risk-sets and for each of the sampled risk-sets from the cohorts with ~30 and ~100 cases, respectively. The sample variance of these estimates was obtained. The relative efficiency was estimated by dividing the

sample variance obtained from the full risk-set analysis by the sample variance obtained from the sampled risk-set analysis. The bias was estimated by subtracting the true exposure-response parameter (i.e. the log of the true hazard ratio) from the mean of the estimated parameters.

Results and Discussion:

Tables 3-1 give summary statistics of the simulations and Tables 3-2 give a summary of the relative efficiency and percent bias for the various analyses. Only those cohorts whose estimates converged for all three Chen estimators and Thomas' estimate were included in the summary tables. For example, for the scenario using exposure intensity Distribution 0, with ~30 cases and a true hazard ratio of 1.005, two thousand cohorts were simulated but only 1992 cohorts were summarized because for 8 of the cohorts, one of the three Chen estimators did not converge.

Notice that the estimated variance of the Chen estimate generally overestimated the sample variance of the parameter estimates. This is especially true for Distribution 2. This would result in conservative hypothesis tests and confidence intervals.

Also note that the relative efficiency for Chen's estimator using Phi 3 is always greater than that of Thomas' estimator for all simulations scenarios. Also, the magnitude of the percent bias is either similar or smaller than that of Thomas'. Phi 3 had the largest support and therefore included the controls of more neighboring risk-sets compared to the Phi 1 and Phi 2 analyses.

In addition, notice that generally, for the simulations with ~30 cases, Chen's estimator using Phi 1 performs worse (i.e. has greater bias and smaller efficiency) than

that of Thomas'. This indicates the importance of how $\varphi(t)$ is defined in the analysis, especially in cohorts with few cases. In particular, if the support of $\varphi(t)$ is not large enough then Chen's estimate will perform poorly. In fact, a preliminary simulation was ran (the results are not shown) in which the support of $\varphi(t)$ was defined to be smaller than the difference in time between the two closest risk-sets and the bias and relative efficiency of Chen's estimate was much worse than that of Thomas' estimate.

Therefore, from these simulations, it appears that Chen's estimate may outperform Thomas' estimate, even with few or moderate number of cases. However, for Chen's estimate to be effective a sufficient number of neighboring risk-sets must be grouped for each failure time. This will happen if the support of $\varphi(t)$ is sufficiently large and/or when the failure times are sufficiently dense (which will occur as the number of cases increases). However, from these simulations, the improvement was moderate.

Section 2:

Simulation with Gold Miner Cohort

Chen's estimator was used on the Gold Miner data set discussed in Section 2 of Chapter 2 and compared with Thomas's estimator. The Gold Miner data set contained 170 cases. To further study the effect of the number of cases on Chen's estimator, the Gold Miner data set was edited. Thirty cases were randomly sampled from the 170 cases, along with their risk-sets. Therefore, two sets of simulations were ran; one on the original Gold Miner data set and one on the sampled Gold Miner data set.

Figures 3-2 give plots of the cases vs their failure age. This was used to help build our definition of $\varphi(t)$ used in Chen's estimator. From this it appears that the risk-sets are very densely populated, and that groupings of 1 year would be sufficient. However, to study the effect of $\varphi(t)$, the following seven functions will be investigated:

$$\text{Phi 0: } \varphi_n(t) = (.0005^2 - t^2) I(|t| \le .0005)$$
$$\text{Phi 1: } \varphi_n(t) = (.25^2 - t^2) I(|t| \le .25)$$
$$\text{Phi 2: } \varphi_n(t) = (.5^2 - t^2) I(|t| \le .5)$$
$$\text{Phi 3: } \varphi_n(t) = (.75^2 - t^2) I(|t| \le .75)$$
$$\text{Phi 4: } \varphi_n(t) = (1^2 - t^2) I(|t| \le 1)$$
$$\text{Phi 5: } \varphi_n(t) = (2^2 - t^2) I(|t| \le 2)$$
$$\text{Phi 6: } \varphi_n(t) = (3^2 - t^2) I(|t| \le 3) \qquad (3.8)$$

which corresponds to grouping risk-sets within 0.0005, 0.25, 0.5, 0.75, 1, 2, and 3 years. Phi 0 was chosen so that no risk-sets would be grouped together. The closest two risk-sets were 2 days apart (i.e. ~.005 years apart).

Risk-sets were formed based on cases age, and 5 controls from each risk-set were randomly sampled 500 times. Risk was analyzed with respect to cumulative exposure and log of cumulative exposure. Eight exposure-response estimates were obtained for each exposure metric by analyzing the sampled risk-sets; one from Thomas' method and seven from Chen's method corresponding to the seven $\varphi_n(t)$ listed above.

The results from analyzing the full original cohort are summarized in Table 3-3. The results from analyzing the sampled risk-sets of the original cohort are summarized in Table 3-4. Note that Chen's estimate with Phi 0 performed very poorly when

compared to Thomas' estimate. The estimated and sample variances were much larger as well as the bias. This demonstrates the fact that Chen's estimate performs poorly when the definition of $\varphi_n(t)$ has a support that is smaller than the two closest risk-sets. The remaining discussion will exclude the results of Chen's estimate based on Phi 0.

As was seen in the Chapter 2, Thomas' estimate tended to overestimate the exposure response parameter of the full cohort when risk was based on cumulative exposure and underestimate the exposure response parameter when risk was based on the log of cumulative exposure. Chen's estimate also followed this same trend. Also note that the mean of the estimated variances always decreased as the support of $\varphi_n(t)$ increased and that the sample variance for Chen's estimates were always smaller than that of Thomas' estimate. However, the average estimated variance for Chen's estimate was not always smaller than that of Thomas' estimate when the log of cumulative exposure was modeled.

Table 3-5 gives the percent bias and mean squared error (mse) for the estimates as compared to the estimate from analyzing the full cohort. It also gives the estimated relative efficiency which is estimated as the variance estimate from the full cohort analysis divided by the average of the variance estimates from the sampled risk-sets analysis. For the analysis based on cumulative exposure, Chen's estimate outperformed Thomas's estimate regardless of the definition of $\varphi_n(t)$. The percent bias and mse decreased significantly and the relative efficiency was always larger. However, for the analysis based on log of cumulative exposure, the bias increased and became larger than the bias of Thomas' estimate as the support of $\varphi_n(t)$ increased. Also, the

estimated relative efficiency increased as the support of $\varphi_n(t)$ increased, although the estimates were similar.

The results from analyzing the full edited cohort are summarized in Table 3-6. The results from analyzing the sampled risk-sets of the edited cohort are summarized in Table 3-7 and the percent bias, relative efficiency and mse are summarized in Table 3-8. It appears that Chen's estimate continues to improve as the support of $\varphi_n(t)$ increases, and perhaps larger supports should have been considered. The bias and sample variance decreased and the relative efficiency increased.

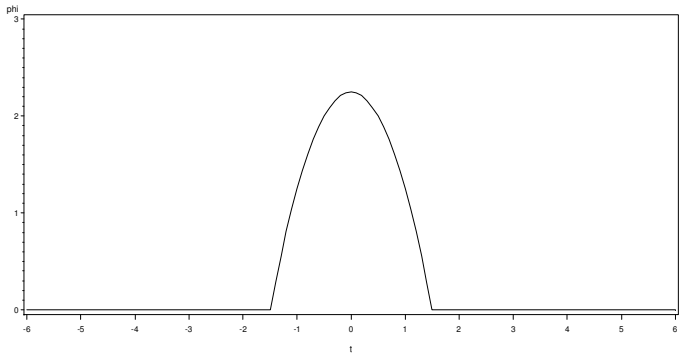
When cumulative exposure was used as the exposure metric, the bias and mse were always smaller for Chen's estimate when compared to Thomas' estimate. However, when the log of cumulative exposure was used as the metric, improvement over Thomas' estimate wasn't seen until the larger supports of $\varphi_n(t)$.

Overall, the Chen estimate seems to be comparable to the Thomas estimator. The definition of $\varphi_n(t)$ seems to present an issue and it is unclear at this point how to determine an optimal support for $\varphi_n(t)$ in order to see any real significant benefit from Chen's estimate. However, it is clear that with fewer cases, larger supports for $\varphi_n(t)$ are required.

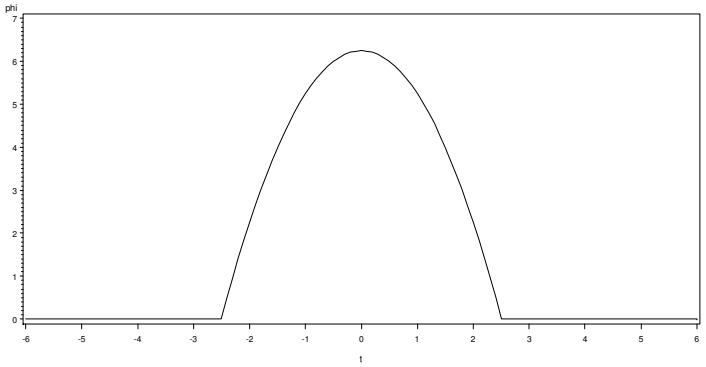
Figures and Tables

Figures 3-1: Graphs of each of the three phi functions used for Chen's Estimator in Section 1. t is measured in years.

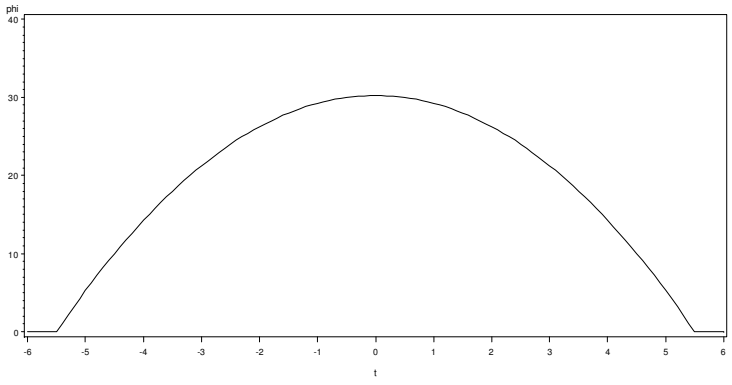
Phi 1



Phi 2



Phi 3



Tables 3-1: Summary statistics of the estimates from the simulations with 1:5 matching for each scenario by analysis method

Distribution 0

True Hazard Ratio	Method	~30 cases N[a]	~30 cases Mean[b]	~30 cases Variance[c]	~30 cases Average Estimated Variance[d]	~ 100 cases N[a]	~ 100 cases Mean[b]	~ 100 cases Variance[c]	~ 100 cases Average Estimated Variance[d]
1.005= $e^{0.00499}$	Thomas	1992	5.110E-03	2.722E-06	2.827E-06	993	5.020E-03	9.350E-07	8.900E-07
	Phi 1	1992	5.370E-03	2.885E-06	3.177E-06	993	5.120E-03	9.380E-07	9.150E-07
	Phi 2	1992	5.280E-03	2.776E-06	3.027E-06	993	5.070E-03	9.250E-07	8.940E-07
	Phi 3	1992	5.130E-03	2.629E-06	2.839E-06	993	4.950E-03	8.960E-07	8.640E-07
1.01= $e^{0.00995}$	Thomas	1992	1.024E-02	4.538E-06	4.535E-06	995	1.007E-02	1.527E-06	1.489E-06
	Phi 1	1992	1.038E-02	3.572E-06	4.920E-06	995	1.007E-02	1.244E-06	1.409E-06
	Phi 2	1992	1.025E-02	3.482E-06	4.510E-06	995	1.002E-02	1.226E-06	1.359E-06
	Phi 3	1992	1.009E-02	3.433E-06	4.087E-06	995	9.930E-03	1.218E-06	1.305E-06

Distribution 2

True Hazard Ratio	Method	~30 cases N[a]	~30 cases Mean[b]	~30 cases Variance[c]	~30 cases Average Estimated Variance[d]	~ 100 cases N[a]	~ 100 cases Mean[b]	~ 100 cases Variance[c]	~ 100 cases Average Estimated Variance[d]
1.005= $e^{0.00499}$	Thomas	1996	5.210E-03	1.072E-05	9.258E-06	989	5.060E-03	1.866E-06	1.934E-06
	Phi 1	1996	5.400E-03	1.140E-05	1.297E-05	989	5.040E-03	1.814E-06	2.198E-06
	Phi 2	1996	5.300E-03	1.095E-05	1.215E-05	989	5.030E-03	1.815E-06	2.130E-06
	Phi 3	1996	5.180E-03	1.046E-05	1.118E-05	989	5.010E-03	1.825E-06	2.071E-06
1.01= $e^{0.00995}$	Thomas	1979	1.064E-02	8.789E-06	8.768E-06	990	1.011E-02	2.296E-06	2.256E-06
	Phi 1	1979	1.028E-02	8.346E-06	1.326E-05	990	9.530E-03	1.948E-06	2.324E-06
	Phi 2	1979	1.009E-02	7.668E-06	1.158E-05	990	9.610E-03	1.890E-06	2.192E-06
	Phi 3	1979	1.001E-02	6.700E-06	9.927E-06	990	9.760E-03	1.791E-06	2.089E-06

[a] – The number of sampled risk-sets such that all 4 parameter estimates converged
[b] – Mean of the parameter estimates
[c] – Sample variance of the parameter estimates
[d] – Average of the variance estimates

Tables 3-2: Percent bias, relative efficiency, and 95% confidence interval coverage probabilities for the simulation scenarios with 1:5 matching

Distribution 0

		~30 cases				~ 100 cases			
True Hazard Ratio	Method	N[a]	Percent Bias[b]	Relative Efficiency[c]	95% CI Captures True Value[d]	N[a]	Percent Bias[b]	Relative Efficiency[c]	95% CI Captures True Value[d]
1.005	Thomas	1992	2.47%	71.37%	95.84%	993	0.65%	74.89%	95.17%
	Phi 1	1992	7.73%	67.33%	96.59%	993	2.57%	74.62%	95.47%
	Phi 2	1992	5.82%	69.99%	96.34%	993	1.61%	75.65%	94.76%
	Phi 3	1992	2.79%	73.88%	96.14%	993	-0.71%	78.13%	94.76%
1.01	Thomas	1992	2.87%	45.30%	95.94%	995	1.21%	46.52%	95.08%
	Phi 1	1992	4.36%	57.55%	98.09%	995	1.19%	57.09%	96.68%
	Phi 2	1992	2.99%	59.04%	97.44%	995	0.75%	57.95%	96.18%
	Phi 3	1992	1.42%	59.87%	96.94%	995	-0.19%	58.33%	95.98%

Distribution 2

		~30 cases				~ 100 cases			
True Hazard Ratio	Method	N[a]	Percent Bias[b]	Relative Efficiency[c]	95% CI Captures True Value[d]	N[a]	Percent Bias[b]	Relative Efficiency[c]	95% CI Captures True Value[d]
1.005	Thomas	1996	4.55%	44.88%	96.95%	989	1.44%	46.98%	96.67%
	Phi 1	1996	8.32%	42.21%	98.10%	989	1.14%	48.34%	96.66%
	Phi 2	1996	6.17%	43.97%	98.00%	989	0.83%	48.31%	96.26%
	Phi 3	1996	3.81%	46.01%	97.65%	989	0.41%	48.05%	95.96%
1.01	Thomas	1979	6.94%	8.70%	95.81%	990	1.58%	13.79%	94.65%
	Phi 1	1979	3.34%	9.16%	95.76%	990	-4.23%	16.24%	92.32%
	Phi 2	1979	1.41%	9.97%	95.15%	990	-3.41%	16.75%	93.23%
	Phi 3	1979	0.61%	11.41%	94.54%	990	-1.94%	17.67%	94.14%

[a] – The number of simulated cohorts such that all 4 parameter estimates converged

[b] – Percent bias is defined as the percent difference between the mean of the estimates and the true value of β

[c] – Relative efficiency is defined as the sample variance of the full cohort estimates divided by the sample variance of the sample risk-set estimates

[d] – The percent of calculated Wald based confidence intervals that contained the true value of β

Figures 3-2: Plot of each case's failure age for the original Gold Miner Data set with 170 cases and the sampled Gold Miner data set with 30 cases

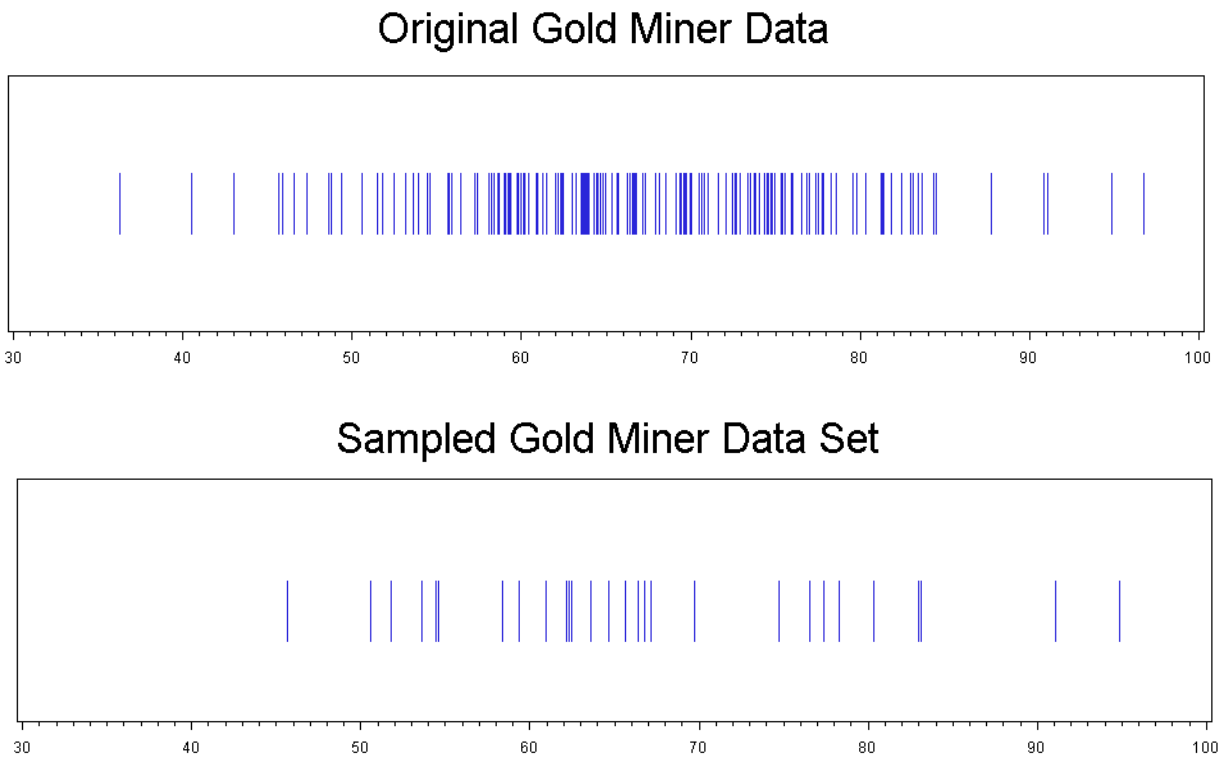


Table 3-3: Summary of analysis on the full cohort of the original Gold Miner Data

Gold Miner Data Analysis			
	Analysis on full cohort		
Exposure Metric	Parameter Estimate	Variance Estimate	Hazard Ratio (per unit increase)
Cumulative Exposure	0.00740	1.47E-07	1.00743
Log Cumulative Exposure	0.01391	9.69E-07	1.01401

Table 3-4: Summary statistics of the analysis on the 500 1:5 matched sampled risk-sets from the original Gold Miner Data

	Analysis on Sampled Risk-Sets							
	Cumulative Exposure				Log Cumulative Exposure			
Method	n	Mean	Var	Average Estimated Variance	n	Mean	Var	Average Estimated Variance
Thomas	494	0.00945	5.29E-07	6.38E-07	500	0.01339	4.84E-07	1.59E-06
Phi 0	494	0.01118	7.30E-07	1.73E-06	500	0.01577	6.82E-07	3.53E-06
Phi 1	494	0.00864	3.69E-07	5.26E-07	500	0.01358	3.04E-07	1.76E-06
Phi 2	494	0.00855	3.41E-07	4.30E-07	500	0.01338	2.82E-07	1.60E-06
Phi 3	494	0.00851	3.19E-07	3.88E-07	500	0.01323	2.54E-07	1.51E-06
Phi 4	494	0.00858	3.15E-07	3.70E-07	500	0.01317	2.37E-07	1.47E-06
Phi 5	494	0.00871	3.39E-07	3.42E-07	500	0.01314	2.34E-07	1.42E-06
Phi 6	494	0.00877	3.62E-07	3.34E-07	500	0.01313	2.30E-07	1.40E-06

Table 3-5: Summary of percent bias, estimated relative efficiency and mean square error (compared to the full cohort estimate) for each analysis on the original Gold Miner data set with 1:5 matching.

	Analysis on Sampled Risk-Sets					
	Cumulative Exposure			Log Cumulative Exposure		
Method	Percent Bias	Relative Efficiency	MSE	Percent Bias	Relative Efficiency	MSE
Thomas	27.97%	23.05%	4.71E-06	-3.78%	60.83%	3.63E-05
Phi 0	51.46%	8.50%	1.50E-05	13.34%	27.48%	7.07E-05
Phi 1	17.16%	27.94%	1.91E-06	-2.39%	55.06%	3.85E-05
Phi 2	15.80%	34.20%	1.65E-06	-3.84%	60.48%	3.60E-05
Phi 3	15.51%	37.88%	1.56E-06	-4.88%	64.05%	3.43E-05
Phi 4	16.13%	39.75%	1.70E-06	-5.37%	65.97%	3.35E-05
Phi 5	17.97%	43.00%	2.05E-06	-5.59%	68.22%	3.31E-05
Phi 6	19.03%	44.07%	2.24E-06	-5.61%	69.31%	3.31E-05

Table 3-6: Summary of analysis on the full cohort of the sampled Gold Miner Data with 30 cases

Gold Miner Data Analysis			
	Analysis on full cohort		
Exposure Metric	Parameter Estimate	Variance Estimate	Hazard Ratio (per unit increase)
Cumulative Exposure	8.94E-03	8.48E-07	1.00898
Log Cumulative Exposure	0.02215	0.00001	1.02240

Table 3-7: Summary statistics of the analysis on the 500 1:5 matched sampled risk-sets from the sampled Gold Miner Data with 30 cases

Analysis on Sampled Risk-Sets								
	Cumulative Exposure				Log Cumulative Exposure			
Method	n	Mean	Var	Average Estimated Variance	n	Mean	Var	Average Estimated Variance
Thomas	476	0.01263	6.51E-06	7.13E-06	466	0.02419	1.47E-05	3.81E-05
Phi 0	476	0.01488	8.93E-06	2.62E-05	466	0.02873	2.13E-05	1.36E-04
Phi 1	476	0.01250	7.71E-06	1.24E-05	466	0.02543	1.83E-05	7.42E-05
Phi 2	476	0.01231	7.64E-06	1.10E-05	466	0.02547	1.82E-05	6.87E-05
Phi 3	476	0.01234	7.69E-06	1.10E-05	466	0.02559	1.85E-05	6.87E-05
Phi 4	476	0.01170	7.31E-06	8.17E-06	466	0.02448	1.30E-05	5.11E-05
Phi 5	476	0.01121	6.33E-06	5.38E-06	466	0.02329	1.10E-05	3.54E-05
Phi 6	476	0.01099	5.34E-06	4.55E-06	466	0.02272	8.82E-06	3.08E-05

Table 3-8: Summary of percent bias, estimated relative efficiency and mean square error (compared to the full cohort estimate) for each analysis on the sampled Gold Miner data set with 30 cases with 1:5 matching.

	Analysis on Sampled Risk-Sets					
	Cumulative Exposure			Log Cumulative Exposure		
Method	Percent Bias	Relative Efficiency	MSE	Percent Bias	Relative Efficiency	MSE
Thomas	41.47%	11.86%	2.01E-05	9.22%	28.95%	1.89E-05
Phi 0	66.68%	3.23%	4.42E-05	29.70%	8.14%	6.46E-05
Phi 1	40.12%	6.83%	2.04E-05	14.82%	14.89%	2.91E-05
Phi 2	37.90%	7.68%	1.90E-05	15.00%	16.09%	2.92E-05
Phi 3	38.28%	7.70%	1.93E-05	15.53%	16.07%	3.03E-05
Phi 4	31.18%	10.35%	1.49E-05	10.51%	21.61%	1.84E-05
Phi 5	25.49%	15.72%	1.15E-05	5.13%	31.17%	1.23E-05
Phi 6	23.04%	18.59%	9.55E-06	2.56%	35.85%	9.14E-06

Chapter 4:

Effect of Classical Measurement Error on the Cox Proportional Hazard Model

Introduction:

Often, the estimated exposure-response curve (i.e. the curve reflecting how $\frac{h(t|\boldsymbol{X})}{h_0(t)}$ changes with respect to cumulative exposure) from occupational cohort studies tends to "level off", or even decrease, at high cumulative exposure levels. There have been many explanations, including the healthy worker survivor effect, a saturation effect, and/or misclassification or mismeasurement of exposure (Stayner 2003).

It is well known that mismeasurement of exposure under a classical error model leads to bias of the exposure-response parameter, and this bias tends to be towards the null (Hu and Lin, 2002). A classical error model assumes the error term and the true exposure are independent. The error model can be additive, in which case the error term and the true exposure are added to obtain the observed exposure, or multiplicative, in which case the error term and true exposure variable are multiplied to obtain the observed exposure.

Hu and Lin (2002), through a simulation study, showed that the introduction of an additive classical error model caused the exposure-response parameter estimate to be biased towards the null, and the bias increased as the standard deviation of the distribution of the error term increases. They further proposed estimators which corrected this bias; however, this required some knowledge of the form of the

distribution of the error term.

It is also possible that measurement error may cause the perceived shape of the exposure-response curve to change resulting in the leveling off of the curve that is often seen at higher exposures.

In practice, the shape of the exposure-response curve is estimated by fitting different models and comparing model fit statistics such as the Akaike information criterion (AIC) value. The most common models fit are:

Log-Linear: $h(t|Exp) = h_0(t)e^{\beta * Exp}$

Power/ Log-Log: $h(t|Exp) = h_0(t)Exp^{\beta} = h_0(t)e^{\beta * \ln(Exp)}$

Linear: $h(t|Exp) = h_0(t) * (1 + \beta * Exp)$ (4.1)

where Exp represents the exposure metric of interest, such as cumulative exposure. Also, fitting splines or a categorical model will give good visuals of how the hazard ratio varies with exposure, however for parsimonious reasons and ease of interpretation, one of the above models is usually reported. Note that the shape of the exposure-response curves associated with the log-linear and linear models do not level off at the higher exposures, whereas, the shape of the power model does level-off at higher exposures if $\beta < 1$. In practice, the log-linear model frequently doesn't fit well and in radiation epidemiology studies, the linear model is the preferred model.

The PHREG procedure in SAS is often used to perform Cox proportional hazards regression. However, this procedure assumes the hazard ratio is log-linear and therefore can only fit the log-linear and power models. The power model can be fit by fitting a log-linear model based on the log of exposure. However, for the power model, further consideration must be made if there are exposure values equal to 0. Often, a

constant, *k*, is added to the exposure value before taking the log and thus the hazard function will have the following form:

$$h(t|Exp) = h_0(t)(k + Exp)^{\beta} = h_0(t)e^{\beta * \ln(k + Exp)} \quad (4.2)$$

This constant can be interpreted as a background level of exposure or is simply present to avoid taking the logarithm of 0. It can be assigned a specific value, a priori, or can be considered as an additional parameter to be estimated in the model (Steenland, 2004).

The linear model cannot be fit using the PHREG procedure. However other programs, such as Epicure, allow for a linear model. Also, recently Langholz and Richardson (2009) proposed a method for the SAS procedure NLMIXED to appropriately handle this model. See Appendix 1 for a detailed description of this method.

Objective:

In occupational cohorts, cumulative exposure of an individual is estimated by multiplying an exposure intensity for a particular job by the duration worked at that job and summing this product across all jobs worked. It seems that there will be little to no error in the measure of duration of a worker, and most of the error in the cumulative exposure estimate would be the result of error in the measurement of the intensity of exposure. The exposure intensity is commonly assumed to follow a log - normal distribution. In particular, the exposure intensity, X, is commonly assumed to satisfy the model:

$$U = \ln(X) = W + \varepsilon$$

or

$$X = e^{W} e^{\varepsilon} \quad (4.3)$$

where

X = the observed exposure intensity

W = the true log-transformed exposure intensity

ε = the measurement error

and W and e are assumed to be independent, following a normal distribution (Kim et al, 2006). Therefore, the error model is multiplicative, or additive on the log scale.

Occupational cohorts were simulated to get a better understanding of the following questions:

1. How does the above measurement error affect the estimated exposure-response parameter?
2. What is the probability of model form misspecification with and without measurement error?
3. How does the perceived shape of the exposure-response curve change with the introduction of measurement error? In particular, is there a tendency for the curve to level off at high exposures in the presence of error?

Method:

Simulations were conducted using SAS Software (version 9.1.3, SAS Institute Inc., Cary, NC). There were six simulation scenarios performed, defined by the true hazard ratio form (log-linear, linear, and power) and the true exposure-response parameter (β = log(1.005) or log (1.01) for log-linear model, β = 0.01 or 0.02 for linear model, and β =

0.01 or 0.02 for power model). The method of simulating cohorts is as described in the Methods section of Chapter 2. In particular, the true exposure intensity Distribution 2 (Log-Normal(.75, 1) – Truncated(0, 50)) was used in each simulation and at each year of follow-up, the probability of mortality from the outcome of interest, *h*, was assigned to each worker based on the workers age and cumulative exposure, *cumexp*, by the following formulas:

$$h = e^{\alpha_0 + 1.5 * \ln\left(\frac{age}{55}\right) + \beta * cumexp} \quad (4.4)$$

for the log-linear model,

$$h = e^{\alpha_0 + 1.5 * \ln\left(\frac{age}{55}\right)} (1 + \beta * cumexp) \quad (4.5)$$

for the linear model, and

$$h = e^{\alpha_0 + 1.5 * \ln\left(\frac{age}{55}\right) + \beta * \ln(cumexp)} \quad (4.6)$$

for the power model, where β is the true exposure-response parameter. The parameter α_0 was chosen so that there would be ~300 cases per cohort.

Once the cohorts were generated, three different observed exposure intensity values were assigned to each worker by multiplying the true exposure intensity value by a random Log-Normal(0, std^2) variable where *std*= 0.1, 0.3 or 0.5. Altogether, there were six simulation scenarios performed with 1,000 cohorts of 5,000 workers.

Analysis:

Each full cohort was analyzed using Cox proportional hazards regression with attained age as the time scale using the PHREG procedure for log-linear and power models and the NLMIXED procedure for linear models. For each cohort, the true cumulative exposure and each of the three observed cumulative exposures were

modeled as a log-linear, power, and linear model resulting in a total of 12 parameter estimates for each cohort. The AIC value was obtained for all regression models to assess model fit.

For each model, the average of the 1,000 parameter estimates was obtained and compared to assess the effect of measurement error on the parameter estimate. Furthermore, the AIC values of the models were compared and the model with the smallest AIC value was said to give the best fit to the data.

Results and Discussion:

For each cohort, summary statistics were collected for the true and observed cumulative exposure of each worker and the average of each of these summary statistic values for the 1,000 cohorts per simulation scenario are summarized in Tables 4-1. Note that the worst case error model, error model with standard deviation 0.05, had a range of observed exposure values more than double the size of the true exposure values.

The summary statistics for the 1,000 parameter estimates are summarized in Tables 4-2. Notice that the measurement error introduced caused attenuation in the exposure response parameter in every scenario as can be seen by comparing the mean value of the estimates for the model based on the true cumulative exposure and the corresponding models based on the observed cumulative exposures. The attenuation became more severe as the error standard deviation increased. However, the estimated standard errors also decreased and therefore, in these simulations, all parameter

estimates remained significant based on a Wald significance test with an $\alpha = 0.05$ level of significance (results not shown).

Table 4-3 gives the results of comparing the AIC values from fitting a log-linear, linear and power model to each cohort and lists the percentage of cohorts for which that model gave the best fit. For example, when the true hazard ratio function was log-linear and had a true hazard ratio of 1.01 the linear model fit the observed data (with *std* = 0.5) best for 79.6% of the 1,000 cohorts simulated. There does appear to be some model misspecification even when there is no measurement error. However, as the exposure response relationship was increased, the probability of misspecification decreased in all models.

Also, notice that when the true hazard ratio had a log-linear form, the linear model fit the observed data best more often as the error standard deviation increased. Furthermore, when the true hazard ratio had a linear form, the power model fit the data best more often as the error standard deviation increased. However, when the true hazard function had a Power form, the Power model fit best more often as the error standard deviation increased. This suggests that the introduction of this measurement error does change the perceived shape of the exposure-response curve. In fact, there appears to be more attenuation in the high exposure end of the curve causing the curves to “level-off” as evidence by the fact that the power model tends to fit the data more often.

To illustrate how well categorical analysis and spline analysis gives a good visualization of the shape exposure-response curve, Figures 4-1 are graphs of the results of a categorical and a restricted cubic spline analysis (Harrell, 2001) as well as

the corresponding true model analysis on the true exposure data for three example cohorts. Five categories were selected based on the quintiles of the distribution of the cases exposure variable. For the restricted cubic spline analysis, three knots were chosen at the 10^{th}, 50^{th}, and 90^{th} percentiles of the overall exposure distribution of the risk-sets. Notice that the spline graph matches the underlying true model and therefore gives a good representation of the shape without imposing a particular parametric model.

To illustrate the effect of the introduction of measurement error, Figures 4-2 graph the results of a categorical and a restricted cubic spline analysis on the same cohorts based on the true and observed exposures. Note that the true and observed curves almost agree in the lower exposure end of the curve, and deviated more as the exposure increased. This may be due to the fact that under this error model, the measurement error is more severe in high exposure categories, which is a common assumption in exposure assessment.

In summary, the introduction of this multiplicative error model caused attenuation of the estimated exposure-response parameter if the true model is fit. The attenuation was more severe as the variance of the error term was increased. In addition, the introduction of this error caused the perceived shape of the exposure response curve to change, resulting in the leveling off of the curve in the high exposure range.

Figures and Graphs

Tables 4-1: Average summary statistics for the true and observed cumulative exposures of each cohort in each simulation scenario. The summary statistics were calculated for each cohort, then averaged across the 1,000 cohorts simulated for each scenario.

Log-Linear True Model

True Hazard Ratio	Error Model	Mean	Var	Skewness	Min	Max
1.005	True	50.49	3761.06	3.66	0.41	697.95
	Error Std = 0.1	50.75	3865.92	3.73	0.41	740.32
	Error Std = 0.3	52.85	4790.17	4.29	0.38	1003.41
	Error Std = 0.5	57.25	7158.83	5.42	0.31	1496.64
1.01	True	50.43	3632.89	3.45	0.44	650.52
	Error Std = 0.1	50.68	3731.88	3.51	0.43	686.37
	Error Std = 0.3	52.76	4614.23	4.03	0.40	927.15
	Error Std = 0.5	57.14	6928.36	5.19	0.32	1432.49

Linear True Model

True β	Error Model	Mean	Var	Skewness	Min	Max
0.01	True	50.57	3778.13	3.67	0.43	701.63
	Error Std = 0.1	50.82	3877.62	3.73	0.42	742.68
	Error Std = 0.3	52.90	4787.49	4.27	0.38	999.29
	Error Std = 0.5	57.29	7156.96	5.41	0.32	1496.20
0.02	True	50.50	3748.91	3.67	0.43	701.55
	Error Std = 0.1	50.76	3850.36	3.73	0.43	741.12
	Error Std = 0.3	52.84	4751.87	4.27	0.39	989.14
	Error Std = 0.5	57.25	7113.80	5.37	0.32	1479.52

Power True Model

True β	Error Model	Mean	Var	Skewness	Min	Max
0.25	True	50.52	3782.49	3.70	0.45	704.87
	Error Std = 0.1	50.78	3889.59	3.77	0.45	752.48
	Error Std = 0.3	52.85	4789.10	4.29	0.40	1001.59
	Error Std = 0.5	57.22	7136.19	5.40	0.33	1490.12
0.5	True	50.53	3769.04	3.69	0.48	703.41
	Error Std = 0.1	50.79	3871.43	3.76	0.47	746.38
	Error Std = 0.3	52.87	4781.06	4.30	0.43	1003.07
	Error Std = 0.5	57.22	7107.78	5.36	0.35	1470.01

Tables 4-2: Summary statistics of parameter estimates from modeling the true cumulative exposure, the observed cumulative exposure using a log-linear, linear, and power model.

a)

		Log-Linear True Model								
		Log-Linear			Linear			Power		
True Hazard Ratio	Error Standard Deviation	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance
$1.005 = e^{0.00499}$	No Error	0.0050	3.27E-07	3.19E-07	0.0115	9.58E-06	9.94E-06	0.2942	4.19E-03	3.15E-03
	.1	0.0048	3.15E-07	3.05E-07	0.0112	9.17E-06	9.55E-06	0.2912	4.12E-03	3.12E-03
	.3	0.0038	3.11E-07	2.19E-07	0.0097	6.80E-06	7.18E-06	0.2713	3.70E-03	2.89E-03
	.5	0.0026	2.70E-07	1.32E-07	0.0073	4.25E-06	4.34E-06	0.2391	3.11E-03	2.53E-03
$1.01 = e^{0.00995}$	No Error	0.0100	1.74E-07	1.92E-07	0.0583	2.42E-04	2.79E-04	0.7776	6.52E-03	3.74E-03
	.1	0.0096	2.09E-07	1.79E-07	0.0569	2.26E-04	2.61E-04	0.7692	6.37E-03	3.69E-03
	.3	0.0071	5.40E-07	1.13E-07	0.0470	1.34E-04	1.54E-04	0.7082	5.52E-03	3.34E-03
	.5	0.0045	6.46E-07	6.10E-08	0.0336	6.31E-05	6.62E-05	0.6135	4.45E-03	2.83E-03

b)

		Linear True Model								
		Log-Linear			Linear			Power		
True β	Error Standard Deviation	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance
0.01	No Error	0.0039	3.85E-07	4.05E-07	0.0104	1.08E-05	1.02E-05	0.2840	4.30E-03	3.31E-03
	.1	0.0038	3.74E-07	3.89E-07	0.0102	1.04E-05	9.82E-06	0.2814	4.26E-03	3.27E-03
	.3	0.0031	3.23E-07	2.90E-07	0.0087	7.67E-06	7.31E-06	0.2628	3.85E-03	3.04E-03
	.5	0.0021	2.62E-07	1.80E-07	0.0064	4.85E-06	4.32E-06	0.2322	3.36E-03	2.67E-03
0.02	No Error	0.0050	2.95E-07	2.76E-07	0.0210	3.28E-05	3.22E-05	0.4440	4.19E-03	3.23E-03
	.1	0.0049	2.98E-07	2.64E-07	0.0205	3.11E-05	3.05E-05	0.4394	4.14E-03	3.20E-03
	.3	0.0039	3.40E-07	1.93E-07	0.0171	2.07E-05	2.05E-05	0.4084	3.76E-03	2.96E-03
	.5	0.0027	2.85E-07	1.17E-07	0.0122	1.03E-05	1.04E-05	0.3592	3.09E-03	2.58E-03

c)

		Power True Model								
		Log-Linear			Linear			Power		
True β	Error Standard Deviation	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance	Mean	Empirical Variance	Mean of Estimated Variance
0.25	No Error	0.0028	4.19E-07	5.15E-07	0.0067	6.25E-06	6.25E-06	0.2519	2.85E-03	3.02E-03
	.1	0.0027	3.99E-07	4.96E-07	0.0066	5.98E-06	6.00E-06	0.2493	2.82E-03	2.99E-03
	.3	0.0022	3.26E-07	3.86E-07	0.0055	4.63E-06	4.52E-06	0.2329	2.66E-03	2.79E-03
	.5	0.0015	2.21E-07	2.44E-07	0.0041	2.87E-06	2.73E-06	0.2063	2.29E-03	2.45E-03
0.5	No Error	0.0046	2.53E-07	2.92E-07	0.0255	5.37E-05	5.70E-05	0.5006	2.96E-03	3.15E-03
	.1	0.0045	2.54E-07	2.79E-07	0.0248	5.12E-05	5.33E-05	0.4959	2.95E-03	3.12E-03
	.3	0.0036	2.88E-07	2.05E-07	0.0197	3.06E-05	3.12E-05	0.4606	2.72E-03	2.89E-03
	.5	0.0025	2.53E-07	1.29E-07	0.0132	1.36E-05	1.33E-05	0.4030	2.40E-03	2.51E-03

Tables 4-3: Percent of cohorts that were fit best by each model

		Log-Linear True Model			
		No Error	**Error std = 0.1**	**Error std = 0.3**	**Error std = 0.5**
True Hazard Ratio	Model				
1.005	Log-Linear	91.0%	88.6%	63.2%	28.8%
	Linear	9.0%	11.4%	36.3%	70.1%
	Power	0.0%	0.0%	0.5%	1.1%
1.01	Log-Linear	99.8%	99.9%	90.9%	20.4%
	Linear	0.2%	0.1%	9.1%	79.6%
	Power	0.0%	0.0%	0.0%	0.0%

		Linear True Model			
		No Error	**Error std = 0.1**	**Error std = 0.3**	**Error std = 0.5**
True β	Model				
0.01	Log-Linear	16.7%	14.9%	9.1%	4.6%
	Linear	73.0%	73.9%	75.6%	68.8%
	Power	10.3%	11.2%	15.3%	26.6%
0.02	Log-Linear	3.1%	2.7%	0.6%	0.5%
	Linear	90.7%	90.7%	85.8%	70.7%
	Power	6.2%	6.6%	13.6%	28.8%

		Power True Model			
		No Error	**Error std = 0.1**	**Error std = 0.3**	**Error std = 0.5**
True β	Model				
0.25	Log-Linear	0.7%	0.8%	0.9%	0.9%
	Linear	9.1%	8.4%	8.2%	7.6%
	Power	90.2%	90.8%	90.9%	91.5%
0.5	Log-Linear	0.0%	0.0%	0.0%	0.0%
	Linear	4.7%	5.0%	2.9%	1.5%
	Power	95.3%	95.0%	97.1%	98.5%

Figure 4-1: Graph of categorical, true model (solid curve) and restricted cubic spline (dashed curve) on the true exposure for an example cohort where true hazard ratio shape is log-linear, linear, and power

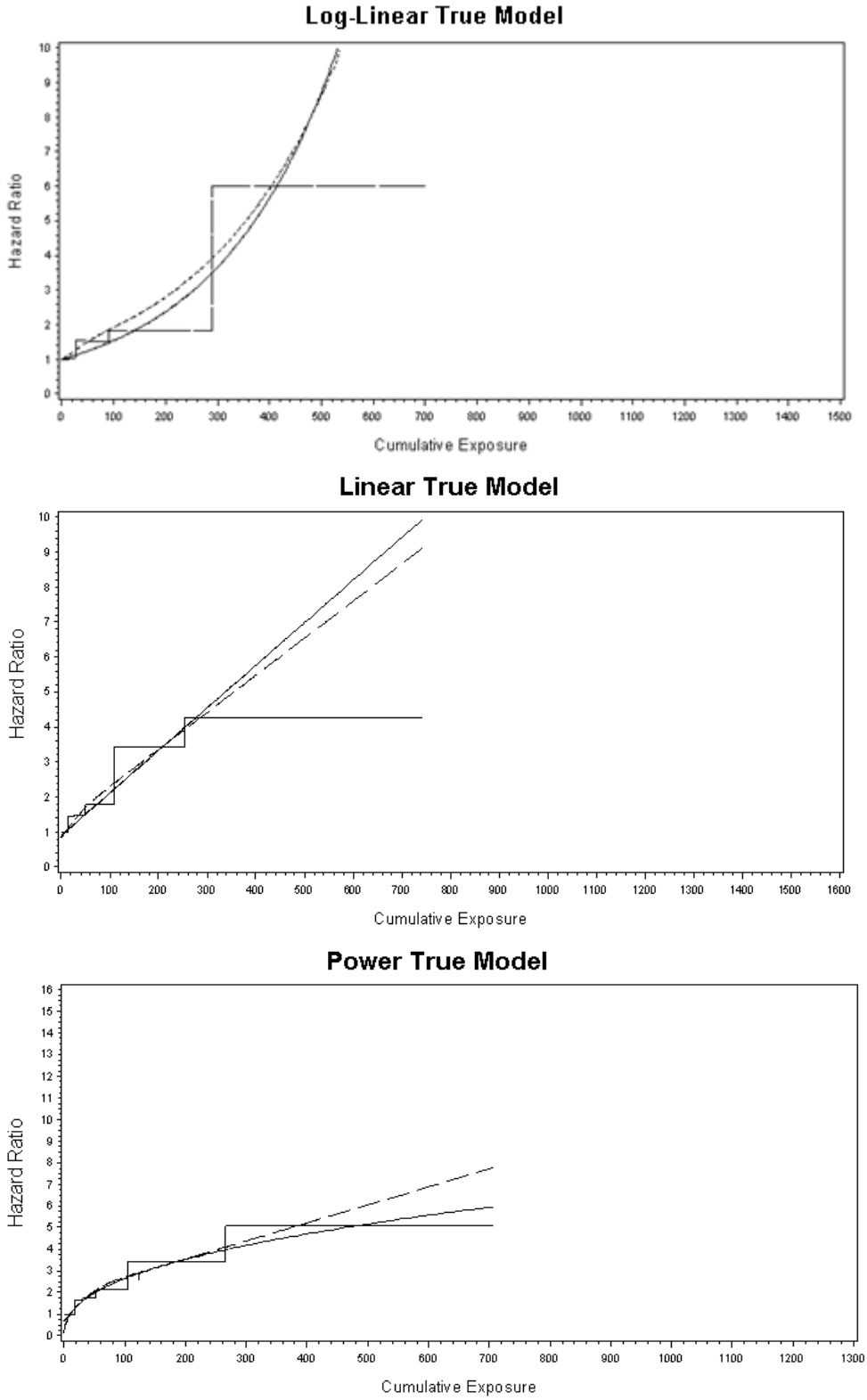
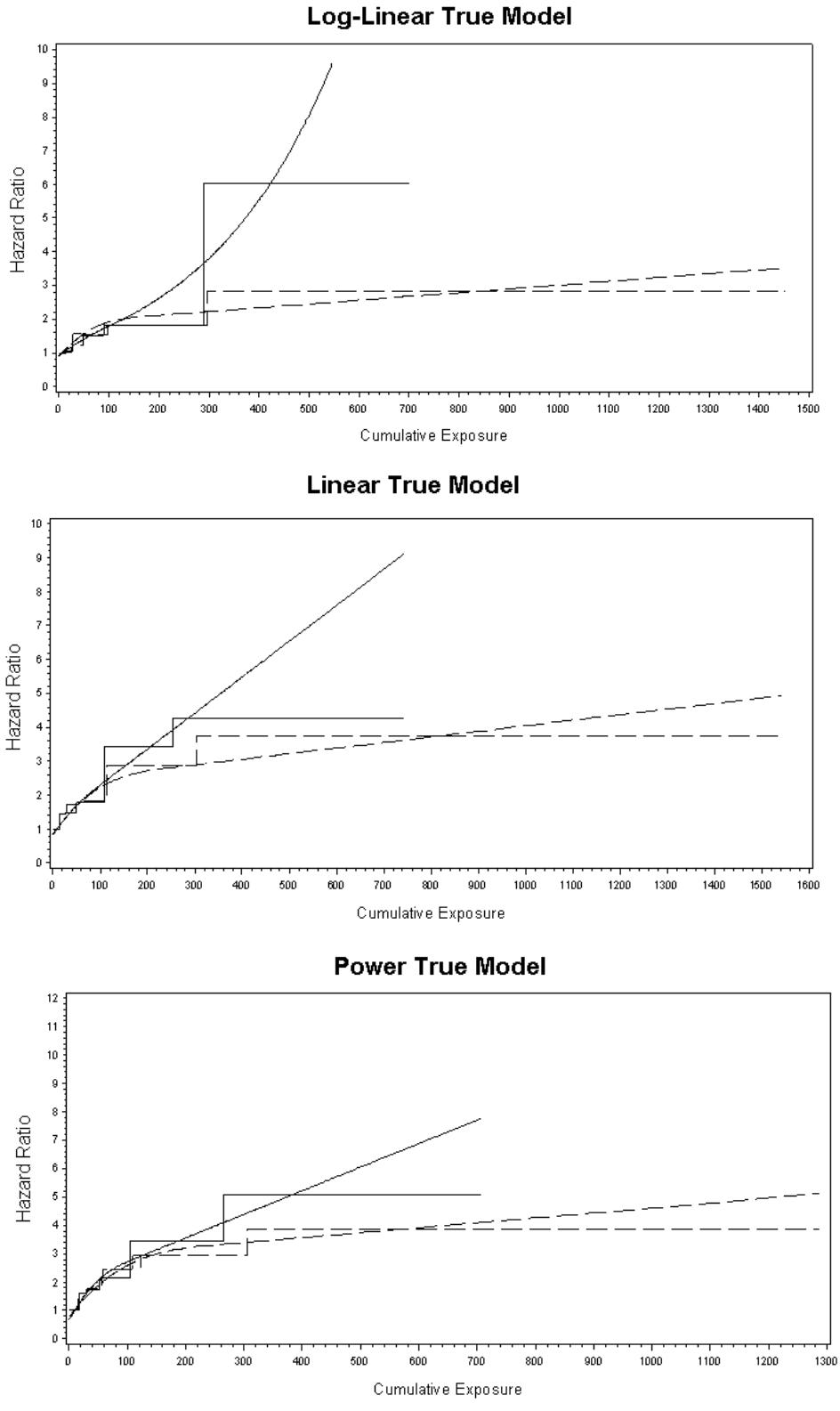


Figure 4-2: Graph of categorical analysis and restricted cubic spline analysis on the true exposure (solid curves) and observed exposure (std = 0.5) (dashed curves) for example cohorts where true hazard ratio shape follows a log-linear, linear, and power model



References

Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics 1982; 10:1100-1120.

Borgan O, Goldstein L, Langholz B. Methods for the Analysis of sampled Cohort Data in the Cox Proportional Hazards Model. The Annals of Statistics 1995; 23:1749-1778.

Breslow NE, Day NE. Statistical Methods in Cancer Research. Oxford: Oxford University Press, 1987.

Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative Models and Cohort Analysis. Journal of the American Statistical Association 1983; 78(381):1-12.

Chen K. Generalized Case-Cohort Sampling. Journal of the Royal Statistical Society. Series B, Statistical Methodology.. 2001; 63:791-809.

Chen K. Statistical Estimation in the Proportional Hazards Model with Risk Set Sampling. Then Annals of Statistics 2004; 32: 1513-1532.

Hu C, Lin DY. Cox Regression with Covariate Measurement Error. Scandinavian Journal of Statistics 2002; 29:637-655.

Cox DR. Regression Models and Life-Tables (with discussion). Journal of the Royal Statistical Society B 1972; 34:187-220.

Cox DR. Partial likelihood. Biometrika 1975; 62(2):269-76.

Goldstein L, Langholz B. Asymptotic Theory for Nested Case-Control Sampling in the Cox Regression Model. The Annals of Statistics 1992; 20:1903-1928.

Harrell FE. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis. New York: Springer-Verlag New York, Inc, 2001.

Hein MJ, Deddens JA, Schubauer-Berigan MK. Bias From Matching On Age at Death or Censor in Nested Case-Control Studies. Epidemiology 2009; 20(3):330-8.

Hu C, Lin DY. Cox Regression with Covariate Measurement Error. Scandinavian Journal of Statistics 2002; 29:637-655.

Kim HM, Yasui Y, Burstyn I. Attenuation in Risk Estimates in Logistic and Cox Proportional-Hazards Models due to Group-Based Exposure Assessment Strategy. Annals of Occupational Hygiene 2006; 50:623-635.

Kriebel D, Checkoway H, Pearce N. Exposure and Dose Modeling in Occupational Epidemiology. Occupational and Environmental Medicine 2007; 64;492-8.

Kromhout H, Symanski E, Rappaport SM. A Comprehensive Evaluation of Within- and Between-Worker Components of Occupational Exposure to Chemical Agents. Annals of Occupational Hygiene 1993; 37:253-270.

Kubale T, Daniels R, Nowlin S, Yiin J, Schubauer-Berigan M, Hornung R, Waters K, Deddens J. Assessment of the Stability of the Reported Risk Estimate from a Nested Case-Control Study of Leukemia and Ionizing Radiation. Poster presented at the 2006 ASA Conference on Radiation and Health, held in Pacific Grove, CA. June, 2006.

Kubale TL, Daniels RD, Yiin JH, Couch J, Schubauer-Berigan MK, Kinnes, GH, Silver SR, Nowlin SJ, and Chen P. A Nested Case-Control Study of Leukemia Mortality and Ionizing Radiation at the Portsmouth Naval Shipyard. Radiat Res; 164(6): 810-19. 2005

Kumazawa, S. and Numakunai, T. A New Theoretical Analysis of Occupational Dose Distributions Indicating the Effect of Dose Limits. Health Phys 41 (3) 465-75. 1981.

Langholz B, Goldstein L. Risk Set Sampling in Epidemiologic Cohort Studies. Statist. Sci. 1996; 11:35-53.

Langholz B, Richardson D. Fitting General Relative Risk and Rate Models to Individually Matched Case-Control Data and Cohort Risk Set Data Using SAS Proc NLMIXED. Biostatistics Division technical Report 179, University of Southern California, LA, 2009.

Lee ET, Wang JW. Statistical Methods for Survival Data Analysis (Third Edition). New Jersey: John Wiley & Sons, Inc., 2003.

Liddell FDK, McDonald JC, Thomas DC. Methods of Cohort Analysis: Appraisal by Application to Asbestos Mining. Journal of the Royal Statistical Society: Series A 1977; 140:469-491

Richardson DB, Loomis D. The Impact of Exposure Categorization for Grouped Analyses of Cohort Data. Occupational and Environmental Medicine 2004; 61:930-5.

Robins JM, Rotnitzky A, Zhao LP. Estimation of Regression Coefficients When Some Regressors are Not Always Observed. J. Amer. Statist. Assoc. 1994; 89:846-866.

Samuelsen SO. A Pseudo-Likelihood Approach to Analysis of Nested Case-Control Studies. Biometrika 1997; 84:379-394.

Stayner L, Steenland K, Dosemeci M, Hertz-Picciotto I. Attenuation of Exposure-Response Curves in Occupational Cohort Studies at High Exposure Levels. Scand J Work Environ Health 2003; 29:317-324.

Steenland K, Deddens JA. A Practical Guide to Dose-Response Analyses and Risk Assessment in Occupational Epidemiology. Epidemiology 2004; 15:63-70.

Steenland K, Brown D. Silicosis Among Gold Miners: Exposure-Response Analyses and Risk Assessment. American Journal of Public Health 1995; 85:1372-1377.

Thiébaut ACM, Bénichou J. Choice of Time-Scale in Cox's Model Analysis of Epidemiologic Cohort Data: A Simulation Study. Statistics in Medicine 2004; 23:3803-3820.

Yiin JH, Schubauer-Berigan MK, Silver SR, Daniels RD, Kinnes GM, Zaebst DD, Couch JR, Kubale TL, Chen PH. Risk of Lung Cancer and Leukemia from Exposure to Ionizing Radiation and Potential Confounders Among Workers at the Portsmouth Naval Shipyard. Radiat Res; 163(6):603-13. 2005

Appendix A

Performing Linear Cox Proportional Hazard Analysis in SAS

Currently, there is only one procedure written in SAS to explicitly perform Cox proportional hazard regression, namely PHREG. However, this procedure assumes that the hazard function has the following form:

$$h(t|\boldsymbol{X}) = h_0(t)e^{\boldsymbol{\beta}'\prime\prime\boldsymbol{X}}$$

where t is time, X is a vector of the covariates, $h(t|\boldsymbol{X}) = h_0(t)e^{\boldsymbol{\beta}'\prime\prime\boldsymbol{X}}$ and $\boldsymbol{\beta}$ contains the parameters to be estimated. However, it may desirable to generalize the above function in the following manner:

$$h(t|\boldsymbol{X}) = h_0(t)g(\boldsymbol{X}, \boldsymbol{\beta})$$

where *g(.)* is any function only of the covariates, $\boldsymbol{X}$, and the parameters, $\boldsymbol{\beta}$. For example, in radiation studies, the hazard function is often assumed to have a linear form:

$$h(t|Exp) = h_0(t)(1 + \beta * Exp)$$

where *Exp* is exposure. Currently, there is no procedure that will explicitly perform Cox proportional hazards regression for general hazard functions and therefore, other software packages are used.

However, Langholz and Richardson (2009) provided code that will fit general hazard models. To do this, they defined p_i corresponding to risk-set $\boldsymbol{R}_i$ as follows:

$$p_i(\boldsymbol{\beta}) = \frac{g(\boldsymbol{X_i}, \boldsymbol{\beta})}{\sum_{j \in R_i} g(\boldsymbol{X_j}, \boldsymbol{\beta})}$$

Then the likelihood function for Cox proportional hazard would be the product of all *n* risk-sets, i.e.:

$$L(\boldsymbol{\beta}) = \prod p_i(\boldsymbol{\beta})$$

Notice that this is the same as the likelihood from observing *n* "successes" from a Bernoulli trial with probability of successes $p_i(\boldsymbol{\beta})$. Procedure NLMIXED has the capability of fitting a Bernoulli model of this type and can therefore be used to fit the data.

Before NLMIXED can be used, the data set must be formatted so that all information from a risk-set is contained in one line. For example, when Cox proportional hazards regression is performed, the data set is formatted in the following manner:

Risk_Set	Case	X
1	1	25
1	0	14
2	1	19
2	0	4
3	1	72
3	0	27

etc…

where the variable Risk_set indexes each risk-set, case is an indicator variable indicating if the observation is a case or not, and X is the corresponding covariate of interest. Here the case is listed first for each risk-set and each risk-set contains only two observations. The following code may be used on the above data set to appropriately format the data:

```
data analytic (keep= risk_set t x1-x2 ); set risk_sets;
   by Risk_set;
   array x{2};
   retain i x1-x2;

   if first.Risk_set then do;
      i = 0;
```

```
      do t=1 to 2;
         x(t)=0;
      end;
   end;

   i = i + 1;

   x{i}=x;

   if last.Risk_set then do;
      t = 1;
      output;
   end;
run;
```

Thus, the data will be formatted as follows:

Risk_Set	t	X1	X2
1	1	25	14
2	1	19	4
3	1	72	27
etc…			

Where X1 is the covariate of the case of each risk-set and X2 is the covariate of the control and t is a variable that is always 1. With this data set, NLMIXED may be used in the following manner:

```
proc nlmixed data=analytic;
   parms b=0;
   sum=0;
   array x{2};

   do i = 1 to 2;
     sum=sum + (1 + b*x(i));
   end;

   p = (1 + b*x(1))/sum;
   model t ~ binary(p);
run;
```

This will fit the Bernoulli likelihood which is equivalent to the Cox proportional hazards likelihood with a linear hazard function as described above with parameter b. This code

has the obvious extension to the situation when there are more than 2 individuals in each risk-set and there are more than one covariate of interest. Also, additional forms of the hazard function may fit.

Appendix B

Chapter 2 Simulations: Simulating and Analyzing Realistic Occupational Cohorts

```
/*****************************************************************************************************************/
/***                                                                                                            ***/
/*** author:  Stephen Bertke/Misty Hein                                                                         ***/
/*** purpose: To generate occupational cohorts under various scenarios for risk.                                ***/
/***          Save output to analyze Bias and Efficiency of analyzing full cohort and sampled risk sets       ***/
/***                                                                                                            ***/
/*****************************************************************************************************************/

%macro cohort(cohort);
%put; %put; %put CREATING COHORT NUMBER: &i_cohort;

*** Randomly assign age at exposure begin (integer), exposure intensity, maximum exposure duration (integer), and
     maximum follow-up (integer);
data cohort1;
  cohort = 1*&cohort;
  do worker = 1 to &n_workers;
    age_exp_begin    = &beta_1 + ROUND(&beta_2*ranexp(&seed),1);
    age_risk_begin   = age_exp_begin;

    if      &exp_method = 0 then do;
        do until (0<exp_intensity<50);
            exp_intensity = 25 + 8*rannor(0);
                end;
         end;

    else if &exp_method = 1 then do;
             do until (exp_intensity<50);
            exp_intensity = exp(2.5 + .5*rannor(0));
                end;
         end;

    else if &exp_method = 2 then do;
        do until (exp_intensity<50);
            exp_intensity =  exp(.75 + 1*rannor(0));
                end;
         end;



    else exp_intensity = .;
    max_duration_exp = 15;
    max_follow_up    = &gamma_1 - ROUND(&gamma_2*ranexp(&seed),1);
    if max_follow_up < 1 then max_follow_up = 1;
    lag_risk_years   = 1*&lag_risk_years;
    output;
    end;
  keep cohort worker age_exp_begin age_risk_begin exp_intensity max_duration_exp max_follow_up lag_risk_years;
  run;

*** Assign age and cumulative exposure (under the true risk lag) at yearly follow-up intervals;
data cohort2;
  set cohort1;
  by cohort worker;
  retain temp_exp;
    temp_exp   = 0;
    if max_follow_up LE max_duration_exp then do;
      *** Scenario A1: max_follow_up LE max_duration_exp --> follow up ends at or before the end of exposure;
      scenario = 'A1';
      do follow_up_year = 1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
```

```
        output;
        end;
    end; *** End A1;
    else do; /*if max_follow_up GT max_duration_exp then*/
      *** Scenario A2: max_follow_up GT max_duration_exp --> follow up extends beyond the end of exposure;
      scenario = 'A2';
      do follow_up_year = 1 to max_duration_exp;
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
        output;
      end;
      temp_exp = temp_exp;
      do follow_up_year = max_duration_exp+1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        output;
      end;
    end; *** End A2;
run;

*** Assign hazards for risk of death (h) and censoring (c) and determine case/censor status for each follow-up year;
data cohort3;
  set cohort2;
  h = min(0.999,exp(&delta_0 + &delta_1*log(temp_age/&age_divisor) + &phi*temp_exp));
  c = min(0.999,exp(&neta_0 + &neta_1*log(temp_age/&age_divisor)));
  if h LE 0 then case = 0;
  else           case = ranbin(0,1,h);
  if c LE 0 then censor = 0;
  else           censor = ranbin(0,1,c);
  run;


*** Determine case and censor status by selecting the first observation with case=1 or censor=1
    if none then output the last observation;
*** Note that if the first observation with case=1 or censor=1 has both case=1 and censor=1,
    then case status is assigned automatically;
data cohort4;
  set cohort3;
  by cohort worker;
  retain stop;
  if first.cohort or first.worker then stop = 0;
  if stop = 0 then do;
    if case = 1 and censor = 1 then do;         case_status = 1; censor_status = 0; stop = 1; output; end;
    else if case = 1 then do;                   case_status = 1; censor_status = 0; stop = 1; output; end;
    else if censor = 1 then do;                 case_status = 0; censor_status = 1; stop = 1; output; end;
    else if last.cohort or last.worker then do; case_status = 0; censor_status = 0; stop = 1; output; end;
    end;
  run;

*** Compute age at risk end, actual cumulative exposure, time exposed, age at exposure end,
    actual duration of exposure and actual follow-up time;
data cohort5;
  set cohort4;
  age_risk_end = temp_age;
  if age_exp_begin + max_duration_exp < age_risk_end then do;
    *** Exposure ceased prior to risk end so truncation is not necessary;
    cumulative_exp = exp_intensity * max_duration_exp;
    age_exp_end = age_exp_begin + max_duration_exp;
    end;
  else do;
    *** Exposure extends beyond risk end so exposure is truncated at risk end;
    cumulative_exp = exp_intensity * (age_risk_end - age_risk_begin);
    age_exp_end = age_risk_end;
    end;
  time_exposed = age_exp_end - age_exp_begin;
  time_at_risk = age_risk_end - age_exp_begin;
  keep cohort worker age_exp_begin
       exp_intensity
       censor_status case_status cumulative_exp
       age_risk_begin age_risk_end age_exp_end
       max_duration_exp max_follow_up
```

```
      time_exposed time_at_risk;
  run;

*** Create final cohort to use in analyses;
data cohort;
  set cohort5;
run;

*** Get number of cases in cohort ***;
proc means data=cohort noprint;
 var case_status;
 output out=casesum n=n sum=cases;
run;

*** Clean up datasets;
ods exclude all;
proc datasets library=work;
  delete cohort1 cohort2 cohort3 cohort4 cohort5 cohort_summary_new;
run; quit; ods select all;
%mend cohort;

*****************************************************************************************************************;
*** RISKSETS macro definition:  create the risk sets for the cohort for use in Cox regression on the full cohort   ***;
***                             and nested case-control analyses                                                   ***;
***                             risk sets are defined based on attained age and attained age pus age at death or   ***;
***                             censor                                                                             ***;
*** input files:  cohort                                                                                          ***;
*** output files: risk_sets                                                                                       ***;
*****************************************************************************************************************;
%macro risksets(cohort);
%put CREATING RISK SETS FOR COHORT NUMBER: &i_cohort;

*** Identify the cases;
data cases;
  set cohort;
  if case_status = 1;
  case_age = age_risk_end;
  case_id = worker;
run;

*** Determine the number of cases and save as a macro variable;
proc means data=cases noprint;
  by cohort;
  var case_status;
  output out=n_cases sum=n_cases;
run;
data n_cases;
  set n_cases;
  call symput('n_cases',n_cases);
run;

*** For each case, identify members of the risk set for both matching on attained age (aacontrol) and ;
*** matching on attained age plus age at death or;
*** censor (dccontrol);
%do i_cases = 1 %to &n_cases;
  data case_n;
    set cases;
    if _n_=&i_cases;
    keep cohort case_age case_id;
    run;
  data risk_set_new;
    set cohort;
    if _n_ = 1 then set case_n;
    *** Select out eligible controls;
    if age_exp_begin LT case_age LE age_risk_end;
    *** Note - LT is important here because of the risk evaluation at yearly intervals;
    *** Identify the cases;
    case = (worker = case_id);
    *** Compute cumulative exposure truncated to the age of the case;
    new_exp1 = (exp_intensity)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
```

```
    label new_exp1 = 'TruncCumExp-unlagged';
run;

proc append base=risk_sets data=risk_set_new force;
run;

%END;

*** Prepare final dataset with all risk sets;
data risk_sets;
  set risk_sets;
  if cohort = . then delete;
  time = 2 - case;
run;

proc phreg data=risk_sets;
 by cohort;
 strata case_id;
 model time*case(0) = new_exp1;
 ods output parameterestimates=parameter_full;
run;

*** Clean up datasets;
ods exclude all;
proc datasets library=work;
  delete n_cases case_n risk_set_new;
  run; quit; ods select all;

proc means data=risk_sets noprint;
  var new_exp1;
  output out=rs_summ n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;

****************************** Separate Cases and Controls ********************************;
data cases; set risk_sets;
 if case = 1;
run;
data controls; set risk_sets;
 if case = 0;
run;

proc means data=cases noprint;
  by case;
  var new_exp1;
  output out=rs_summ_case n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;
proc means data=controls noprint;
  by case;
  var new_exp1;
  output out=rs_summ_cont n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;
%mend risksets;

%macro nestedcc_age;
%put NESTED CASE-CONTROL REGRESSION FOR COHORT NUMBER: &i_cohort REPS: &i_rep;


  *** Select out the appropriate number of controls for each case;
data cases; set risk_sets;
 if case = 1;
run;
data controls; set risk_sets;
 if case = 0;
run;

proc surveyselect data=controls out=out_1 method=srs sampsize=1 SELECTALL noprint;
 strata case_id;
run;
data ncc_1; set cases out_1;
proc sort data=ncc_1; by case_id case;
```

```
run;
proc phreg data=ncc_1 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_1;
run;


proc surveyselect data=controls out=out_3 method=srs sampsize=3 SELECTALL noprint;
 strata case_id;
run;
data ncc_3; set cases out_3;
proc sort data=ncc_3; by case_id case;
run;
proc phreg data=ncc_3 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_3;
run;


proc surveyselect data=controls out=out_5 method=srs sampsize=5 SELECTALL noprint;
 strata case_id;
run;
data ncc_5; set cases out_5;
proc sort data=ncc_5; by case_id case;
run;
proc phreg data=ncc_5 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_5;
run;


proc surveyselect data=controls out=out_10 method=srs sampsize=10 SELECTALL noprint;
 strata case_id;
run;
data ncc_10; set cases out_10;
proc sort data=ncc_10; by case_id case;
run;
proc phreg data=ncc_10 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_10;
run;


proc surveyselect data=controls out=out_15 method=srs sampsize=15 SELECTALL noprint;
 strata case_id;
run;
data ncc_15; set cases out_15;
proc sort data=ncc_15; by case_id case;
run;
proc phreg data=ncc_15 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_15;
run;


proc surveyselect data=controls out=out_20 method=srs sampsize=20 SELECTALL noprint;
 strata case_id;
run;
data ncc_20; set cases out_20;
proc sort data=ncc_20; by case_id case;
```

```
run;
proc phreg data=ncc_20 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_20;
run;


proc datasets library=work;
  delete ncc_1 ncc_3 ncc_5 ncc_10 ncc_15 ncc_20 ;
run;


%mend;

*************************************************************************************************************;
*** ITERATE macro definition:  iterate through {create cohort, summarize, define risk sets, Cox regression,      ***;
***                            NCC sampling, and Cox regression                                                  ***;
*** input files:  none                                                                                          ***;
*** output files: Lots!                                                                                         ***;
*************************************************************************************************************;


%macro iterate(/*************************************************************************************************/
               /*                                                                                               */
               /*** input parameters that might vary within a set of simulations                                */
               n_cohorts      = ,             /* number of cohorts                                             */
               n_workers      = ,             /* number of workers per cohort                                  */
               n_rep          = ,             /* number of repetitions per cohort                              */
               exp_method     = ,             /* exp_method in 1=uniform, 2=lognormal, 3=exponential           */
               phi            = ,             /* risk parameter                                                */
               /***                                                                                           */
               /*** input parameters that should remain constant within a set of simulations                    */
               seed           = 0,            /* initial seed                                                  */
               beta_1         = 18,           /* age at exposure begin parameter                               */
               beta_2         = 10,           /* age at exposure begin parameter                               */
               gamma_1        = 40,           /* max follow-up parameter                                       */
               gamma_2        = 5,            /* max follow-up parameter                                       */
               zeta_1         = 25,           /* max duration of exposure parameter                            */
                             delta_0        = ,
               delta_1        = 1.5,            /* risk parameter                                              */
               age_divisor    = 55,           /* divisor for age in risk and censoring models                  */
               neta_0         = -5.0,         /* censoring parameter                                           */
               neta_1         = 5.0);          /* censoring parameter                                          */
/*************************************************************************************************************/

title1 "Simulating &n_cohorts cohorts of size &n_workers workers using a seed of &seed.";
title2 "Exposure based on method &exp_method and risk lag = &lag_risk_years years.";
title3 "Risk parameters include delta_0=&delta_0, delta_1=&delta_1 and phi=&phi.";
title4 "Censoring parameters include neta_0=&neta_0 and neta_1=&neta_1.";

*** Assign Library for data and log to be saved ***;
data _null_;
  rr = exp(&phi);
  call symput('rr', trim(left(rr)));
run;
libname steve "&lib\&exp_method\&rr";

filename mylist "&lib\&exp_method\&rr\listing.lst";
filename mylog  "&lib\&exp_method\&rr\log.log";
proc printto log=mylog print=mylist;
run;

data param_full; run;
data param_1; run;
data param_3; run;
data param_5; run;
data param_10; run;
data param_15; run;
```

```
data param_20; run;


data casesummary; run;
data rs_summary; run;
data rs_summary_cont; run;
data rs_summary_case; run;

*** Iterate by cohort ***;
%do i_cohort = 1 %to &n_cohorts;
    %cohort(&i_cohort);
    data casesummary; set casesummary casesum; run;


    %risksets(&i_cohort);
    data param_full; set param_full parameter_full; keep cohort estimate stderr probchisq; run;

         data rs_summary; set rs_summary rs_summ; run;
         data rs_summary_cont; set rs_summary_cont rs_summ_cont; run;
         data rs_summary_case; set rs_summary_case rs_summ_case; run;
  %do i_rep = 1 %to &n_rep;

    %nestedcc_age;
         data param_1; set param_1 parameter_1; keep cohort estimate stderr probchisq; run;
         data param_3; set param_3 parameter_3; keep cohort estimate stderr probchisq; run;
    data param_5; set param_5 parameter_5; keep cohort estimate stderr probchisq; run;
    data param_10; set param_10 parameter_10; keep cohort estimate stderr probchisq; run;
         data param_15; set param_15 parameter_15; keep cohort estimate stderr probchisq; run;
    data param_20; set param_20 parameter_20; keep cohort estimate stderr probchisq; run;

  %end;



    proc datasets library=work;
          delete cohort risk_sets;
         run;


%end;
*** Clear out titles;
title1;title2;title3;title4;

*** Save data ***;
data steve.param_full; set param_full; run;
data steve.param_1; set param_1; run;
data steve.param_3; set param_3; run;
data steve.param_5; set param_5; run;
data steve.param_10; set param_10; run;
data steve.param_15; set param_15; run;
data steve.param_20; set param_20; run;

data steve.summary; set casesummary; run;

data steve.rs_summary; set rs_summary; run;
data steve.rs_summary_cont; set rs_summary_cont; run;
data steve.rs_summary_case; set rs_summary_case; run;

************* Print the Log and Output in their respective windows *********************;
proc printto;
run;

%mend iterate;


%let lib =;


%iterate(seed       = ,
         n_cohorts  = ,
```

```
n_rep      = ,
 n_workers  = ,
delta_0    = ,
 exp_method = ,
 lag_risk_years = ,
 phi        = );
```

Appendix C:

Performing Chen's Regression for a Given Cohort

```
%macro b;

*** Determine the number of CASES AND TOTAL CONTROLS********************;
proc means data=_cases noprint;
  var &strata;
  output out=n_rs n=n_rs;
  run;
data _NULL_;
  set n_rs;
  call symput('n_rs',n_rs);
run;

proc means data=_controls noprint;
  var &STRATA;
  output out=n_WKRS n=n_WKRS;
  run;
data _NULL_;
  set n_WKRS;
  call symput('n_WKRS',n_WKRS);
run;
*****************************************************************;

********************** Calculate b and w **********************;
PROC IML;

USE BETA;
READ VAR {&beta_var} INTO BETA;

do time=1 to &n_rs;

        top=0;
        bottom=0;

    use _cases;
    READ point time VAR {case_age} INTO t;
        READ point time VAR {&strata} INTO strata;

    do worker=1 to &n_WKRS;
       USE _controls;
            read point worker VAR {case_age} into case_age;
       read point worker VAR {&VAR} into Z;
            phi=&f_phi;

            top = TOP + (1/&n)*exp(beta*z`)*phi;
         end;
         do rs=1 to &n_rs;
       use _cases;
            read point rs var {case_age} into case_age;
       phi=&f_phi;

            bottom = bottom + (&m./&n.)*phi;
         end;

         b=top/bottom;
         next=strata||b;
         beta_hat = beta_hat//next;
end;

 CREATE bhat_all FROM beta_hat[colname={"&strata" 'bhat'}];
 APPEND FROM beta_hat;

RUN;
```

```
DATA _controls; MERGE _controls BHAT_ALL; BY &strata;
  W = ((&m.*bhat)/(exp(&bx) + &m.*bhat));
DATA _cases; MERGE _cases BHAT_ALL; BY &strata;
  W = ((&m.*bhat)/(exp(&bx) + &m.*bhat));
RUN;
%mend;

%macro s;

********************** Calculate S0 and S1, S2 ********************;
L=SHAPE(0,1,1);
F=SHAPE(0,1,&V);
FP=SHAPE(0,&V,&V);

do time=1 to &n_rs;
    use _cases;
    READ point time VAR {case_age} INTO t;

    S0=SHAPE(0,1,1);
    S1=SHAPE(0,1,&V);
    S2=SHAPE(0,&V,&V);

    USE _controls;
    do worker=1 to &n_WKRS;
       read point worker VAR {w} into w;
            read point worker VAR {case_age} into case_age;
       read VAR {&VAR} into Z;

            phi=&f_phi;

            mult=(1/&N)*w*exp(beta*z`)*phi;
       S0=S0 + MULT;
       S1=S1 + MULT*Z;
       S2=S2 + MULT*(Z`*Z);
    end;

    USE _cases;
    READ POINT TIME VAR {W} INTO W;
    READ POINT TIME VAR {&VAR} INTO Z;

         L = L + W*(beta*z` - LOG(S0));
    F = F + W*(Z - (1/S0)*S1);
    FP = FP - W*(((1/S0)*S2) - (((1/S0)*S1)`*((1/S0)*S1)));

end;

%MEND;



%macro newton;

********************** Perform Newton Algorithm **********************;

PROC IML;
 USE BETA;
 READ VAR {&beta_var} INTO BETA;

 do r=1 to 10 until (s<.0000001);
  %s;
  BETA=(BETA` - INV(FP)*F`)`;
  S=F*F`;
 end;

 VARIABLE={&VARQ};
 VARIABLE=VARIABLE`;
 ESTIMATE=BETA`;
 COV=-INV(FP);
 VARIANCE=VECDIAG(COV);
```

```
 GRADIENT=F`;

 OUTPUT=ESTIMATE||VARIANCE||GRADIENT;
 CREATE OUTPUT FROM OUTPUT[COLNAME={'ESTIMATE' 'VAR' 'GRADIENT'}] ;
 APPEND FROM OUTPUT;

 CREATE PARAMETER FROM VARIABLE[COLNAME={'PARAMETER'}] ;
 APPEND FROM VARIABLE;

 CREATE L FROM L[COLNAME={'L'}] ;
 APPEND FROM L;
RUN;
DATA OUTPUT; MERGE PARAMETER OUTPUT;
PROC PRINT DATA=OUTPUT;
RUN;

%mend;

%macro analyze(data=,n=,m=,strata=,var=,supp=);
DATA N;
 M=count("&VAR", ' ');
 v=M+1;
 CALL SYMPUT('v',v);
RUN;
%let beta_var=;
%LET VARQ=;

%DO I=1 %TO &v;
  %LET VARRI=%SCAN(&VAR,&I, %STR( ));
  DATA _NULL_;
         CALL SYMPUT('beta_var', SYMGET('beta_var')|| ' '||"bp_&VARRI. ");
         CALL SYMPUT('varq', SYMGET('varq')|| ' '||"'&VARRI.' ");
  RUN;
%END;

%LET BX =;
%DO I=1 %TO &v;
  %LET VARRI=%SCAN(&VAR,&I, %STR( ));
  DATA _NULL_;
    IF &I NE &v THEN DO;
           call symput('BX', symget('BX') ||' '||"&VARRI.*BP_&VARRI. +");
         END;
         ELSE DO;
           call symput('BX', symget('BX') ||' '||"&VARRI.*BP_&VARRI.");
         END;
  RUN;
%END;

   ********** Define the function phi ************************************;
   ********** Macro variable supp defines the support of phi **************;
   %let f_phi=(abs(t - case_age) <= &supp)*((&supp)**2 - (t - case_age)**2);

proc phreg data=&data;
 strata &strata;
 model time*case(0) = &var;
 ods output parameterestimates=ncc_p;
run;

*** Create dataset containing estimates ***************;
proc transpose data=ncc_p out=beta prefix=bp_; id parameter;
 var estimate;
run;
data betap beta; set beta;
 drop _name_ _label_;
run;
DATA NCC; IF _N_=1 THEN SET BETAP; SET &DATA; RUN;
*********************************************************;

DATA _controls; SET NCC;
   IF CASE=0;
```

```
DATA _cases; SET NCC;
   IF CASE=1;
RUN;

%b;
%newton;
proc print data=OUTPUT;
run;
%mend;

%macro cohort(cohort);
%put; %put; %put CREATING COHORT NUMBER: &i_cohort;

*** Randomly assign age at exposure begin (integer), exposure intensity, maximum exposure duration (integer), and
     maximum follow-up (integer);
data cohort1;
  cohort = 1*&cohort;
  do worker = 1 to &n_workers;
    age_exp_begin    = &beta_1 + ROUND(&beta_2*ranexp(&seed),1);
    age_risk_begin   = age_exp_begin;

    if      &exp_method = 0 then do;
        do until (0<exp_intensity<50);
            exp_intensity = 25 + 8*rannor(0);
                 end;
         end;

    else if &exp_method = 1 then do;
             do until (exp_intensity<50);
            exp_intensity = exp(2.5 + .5*rannor(0));
                 end;
         end;

    else if &exp_method = 2 then do;
        do until (exp_intensity<50);
            exp_intensity =  exp(.75 + 1*rannor(0));
                 end;
         end;



    else exp_intensity = .;
    max_duration_exp = 15;
    max_follow_up    = &gamma_1 - ROUND(&gamma_2*ranexp(&seed),1);
    if max_follow_up < 1 then max_follow_up = 1;
    lag_risk_years   = 1*&lag_risk_years;
    output;
    end;
  keep cohort worker age_exp_begin age_risk_begin exp_intensity max_duration_exp max_follow_up lag_risk_years;
  run;

*** Assign age and cumulative exposure (under the true risk lag) at yearly follow-up intervals;
data cohort2;
  set cohort1;
  by cohort worker;
  retain temp_exp;
    temp_exp   = 0;
    if max_follow_up LE max_duration_exp then do;
      *** Scenario A1: max_follow_up LE max_duration_exp --> follow up ends at or before the end of exposure;
      scenario = 'A1';
      do follow_up_year = 1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
        output;
        end;
    end; *** End A1;
    else do; /*if max_follow_up GT max_duration_exp then*/
      *** Scenario A2: max_follow_up GT max_duration_exp --> follow up extends beyond the end of exposure;
      scenario = 'A2';
      do follow_up_year = 1 to max_duration_exp;
```

```
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
        output;
      end;
      temp_exp = temp_exp;
      do follow_up_year = max_duration_exp+1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        output;
      end;
    end; *** End A2;
run;

*** Assign hazards for risk of death (h) and censoring (c) and determine case/censor status for each follow-up year;
data cohort3;
  set cohort2;
  h = min(0.999,exp(&delta_0 + &delta_1*log(temp_age/&age_divisor) + &phi*temp_exp));
  c = min(0.999,exp(&neta_0 + &neta_1*log(temp_age/&age_divisor)));
  if h LE 0 then case = 0;
  else           case = ranbin(0,1,h);
  if c LE 0 then censor = 0;
  else           censor = ranbin(0,1,c);
  run;


*** Determine case and censor status by selecting the first observation with case=1 or censor=1
    if none then output the last observation;
*** Note that if the first observation with case=1 or censor=1 has both case=1 and censor=1,
    then case status is assigned automatically;
data cohort4;
  set cohort3;
  by cohort worker;
  retain stop;
  if first.cohort or first.worker then stop = 0;
  if stop = 0 then do;
    if case = 1 and censor = 1 then do;         case_status = 1; censor_status = 0; stop = 1; output; end;
    else if case = 1 then do;                   case_status = 1; censor_status = 0; stop = 1; output; end;
    else if censor = 1 then do;                 case_status = 0; censor_status = 1; stop = 1; output; end;
    else if last.cohort or last.worker then do; case_status = 0; censor_status = 0; stop = 1; output; end;
    end;
  run;

*** Compute age at risk end, actual cumulative exposure, time exposed, age at exposure end,
    actual duration of exposure and actual follow-up time;
data cohort5;
  set cohort4;
  age_risk_end = temp_age;
  if age_exp_begin + max_duration_exp < age_risk_end then do;
    *** Exposure ceased prior to risk end so truncation is not necessary;
    cumulative_exp = exp_intensity * max_duration_exp;
    age_exp_end = age_exp_begin + max_duration_exp;
    end;
  else do;
    *** Exposure extends beyond risk end so exposure is truncated at risk end;
    cumulative_exp = exp_intensity * (age_risk_end - age_risk_begin);
    age_exp_end = age_risk_end;
    end;
  time_exposed = age_exp_end - age_exp_begin;
  time_at_risk = age_risk_end - age_exp_begin;
  keep cohort worker age_exp_begin
       exp_intensity
       censor_status case_status cumulative_exp
       age_risk_begin age_risk_end age_exp_end
       max_duration_exp max_follow_up
       time_exposed time_at_risk;
  run;

*** Create final cohort to use in analyses;
data cohort;
  set cohort5;
run;
```

```
*** Get number of cases in cohort ***;
proc means data=cohort noprint;
 var case_status;
 output out=casesum n=n sum=cases;
run;

*** Clean up datasets;
ods exclude all;
proc datasets library=work;
  delete cohort1 cohort2 cohort3 cohort4 cohort5 cohort_summary_new;
run; quit; ods select all;
%mend cohort;
*****************************************************************************************************************;
*** RISKSETS macro definition:  create the risk sets for the cohort for use in Cox regression on the full cohort   ***;
***                             and nested case-control analyses                                                   ***;
***                             risk sets are defined based on attained age and attained age pus age at death or   ***;
***                             censor                                                                             ***;
*** input files:  cohort                                                                                          ***;
*** output files: risk_sets                                                                                       ***;
*****************************************************************************************************************;
%macro risksets(cohort);
%put CREATING RISK SETS FOR COHORT NUMBER: &i_cohort;

*** Identify the cases;
data cases;
  set cohort;
  if case_status = 1;
  case_age = age_risk_end;
  case_id = worker;
run;

*** Determine the number of cases and save as a macro variable;
proc means data=cases noprint;
  by cohort;
  var case_status;
  output out=n_cases sum=n_cases;
run;
data n_cases;
  set n_cases;
  call symput('n_cases',n_cases);
run;

*** For each case, identify members of the risk set for both matching on attained age (aacontrol) and ;
*** matching on attained age plus age at death or;
*** censor (dccontrol);
%do i_cases = 1 %to &n_cases;
  data case_n;
    set cases;
    if _n_=&i_cases;
    keep cohort case_age case_id;
    run;
  data risk_set_new;
    set cohort;
    if _n_ = 1 then set case_n;
    *** Select out eligible controls;
    if age_exp_begin LT case_age LE age_risk_end;
    *** Note - LT is important here because of the risk evaluation at yearly intervals;
    *** Identify the cases;
    case = (worker = case_id);
    *** Compute cumulative exposure truncated to the age of the case;
    new_exp1 = (exp_intensity)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
    label new_exp1 = 'TruncCumExp-unlagged';
run;

proc append base=risk_sets data=risk_set_new force;
run;

%END;
```

```
*** Prepare final dataset with all risk sets;
data risk_sets;
  set risk_sets;
  if cohort = . then delete;
  time = 2 - case;
run;

proc phreg data=risk_sets;
 by cohort;
 strata case_id;
 model time*case(0) = new_exp1;
 ods output parameterestimates=parameter_full;
run;

*** Clean up datasets;
ods exclude all;
proc datasets library=work;
  delete n_cases case_n risk_set_new;
  run; quit; ods select all;

proc means data=risk_sets noprint;
  var new_exp1;
  output out=rs_summ n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;

****************************** Separate Cases and Controls *********************************;
data cases; set risk_sets;
 if case = 1;
run;
data controls; set risk_sets;
 if case = 0;
run;

proc means data=cases noprint;
  by case;
  var new_exp1;
  output out=rs_summ_case n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;
proc means data=controls noprint;
  by case;
  var new_exp1;
  output out=rs_summ_cont n=n sum=sum mean=mean var=var skew=skew min=min max=max;
run;
%mend risksets;

%macro nestedcc_age;
%put NESTED CASE-CONTROL REGRESSION FOR COHORT NUMBER: &i_cohort REPS: &i_rep;


  *** Select out the appropriate number of controls for each case;
data cases; set risk_sets;
 if case = 1;
run;
data controls; set risk_sets;
 if case = 0;
run;

proc surveyselect data=controls out=out_5 method=srs sampsize=5 SELECTALL noprint;
 strata case_id;
run;
data ncc_5; set cases out_5;
proc sort data=ncc_5; by case_id case;
run;
proc phreg data=ncc_5 nosummary;
 by cohort;
 model time*case(0)=new_exp1;
 strata case_id;
 ods output parameterestimates=parameter_5;
run;
%analyze(data=ncc_5,n=5000,m=5,strata=case_id,var=new_exp1,supp=1);
```

```
data phi_1_parameter_5; set output;
%analyze(data=ncc_5,n=5000,m=5,strata=case_id,var=new_exp1,supp=2);
data phi_2_parameter_5; set output;
%analyze(data=ncc_5,n=5000,m=5,strata=case_id,var=new_exp1,supp=5);
data phi_3_parameter_5; set output;


proc datasets library=work;
  delete ncc_5 ncc_20 ;
run;


%mend;

*************************************************************************************************************;
*** ITERATE macro definition:  iterate through {create cohort, summarize, define risk sets, Cox regression,      ***;
***                            NCC sampling, and Cox regression                                                  ***;
*** input files:  none                                                                                          ***;
*** output files: Lots!                                                                                         ***;
*************************************************************************************************************;

%macro iterate(/*************************************************************************************************************/
               /*                                                                                                           */
               /*** input parameters that might vary within a set of simulations                                           */
               n_cohorts      = ,             /* number of cohorts                                                      */
               n_workers      = ,             /* number of workers per cohort                                           */
               n_rep          = ,             /* number of repetitions per cohort                                       */
               exp_method     = ,             /* exp_method in 1=uniform, 2=lognormal, 3=exponential                    */
               phi            = ,             /* risk parameter                                                         */
               lag_risk_years = ,             /* lag employed when determining case status - must be an integer          */
               /***                                                                                                       */
               /*** input parameters that should remain constant within a set of simulations                             */
               seed           = 0,            /* initial seed                                                           */
               beta_1         = 18,           /* age at exposure begin parameter                                        */
               beta_2         = 10,           /* age at exposure begin parameter                                        */
               gamma_1        = 40,           /* max follow-up parameter                                                */
               gamma_2        = 5,            /* max follow-up parameter                                                */
               zeta_1         = 25,           /* max duration of exposure parameter                                     */
                             delta_0        = ,
               delta_1        = 1.5,            /* risk parameter                                                      */
               age_divisor    = 55,           /* divisor for age in risk and censoring models                           */
               neta_0         = -5.0,         /* censoring parameter                                                    */
               neta_1         = 5.0);          /* censoring parameter                                                   */
/*************************************************************************************************************/

title1 "Simulating &n_cohorts cohorts of size &n_workers workers using a seed of &seed.";
title2 "Exposure based on method &exp_method and risk lag = &lag_risk_years years.";
title3 "Risk parameters include delta_0=&delta_0, delta_1=&delta_1 and phi=&phi.";
title4 "Censoring parameters include neta_0=&neta_0 and neta_1=&neta_1.";

*** Assign Library for data and log to be saved ***;
data _null_;
  rr = exp(&phi);
  call symput('rr', trim(left(rr)));
run;
libname steve "&lib\&exp_method\&rr";

filename mylist "C:\Users\Steve\Desktop\New Method\30 Cases\log-output\listing.lst";
filename mylog  "C:\Users\Steve\Desktop\New Method\30 Cases\log-output\log.log";
proc printto log=mylog print=mylist;
run;

data param_full; run;
data param_5; run;
data param_20; run;

data phi_1_param_5; run;
data phi_2_param_5; run;
data phi_3_param_5; run;
data phi_1_param_20; run;
```

```
data phi_2_param_20; run;
data phi_3_param_20; run;

data casesummary; run;
data rs_summary; run;
data rs_summary_cont; run;
data rs_summary_case; run;

data all_time;

*** Iterate by cohort ***;
%do i_cohort = 1 %to &n_cohorts;
    %cohort(&i_cohort);
    data casesummary; set casesummary casesum; run;

        data time; set cohort; if case_status=1; keep cohort worker age_risk_end;
        proc sort data=time; by age_risk_end;
        data all_time; set all_time time;


    %risksets(&i_cohort);
    data param_full; set param_full parameter_full; keep cohort estimate stderr probchisq; run;

        data rs_summary; set rs_summary rs_summ; run;
        data rs_summary_cont; set rs_summary_cont rs_summ_cont; run;
        data rs_summary_case; set rs_summary_case rs_summ_case; run;
  %do i_rep = 1 %to &n_rep;

    %nestedcc_age;
    data param_5; set param_5 parameter_5; keep cohort estimate stderr probchisq; run;
    data param_20; set param_20 parameter_20; keep cohort estimate stderr probchisq; run;

        data phi_1_param_5; set  phi_1_param_5 phi_1_parameter_5; run;
        data phi_2_param_5; set  phi_2_param_5 phi_2_parameter_5; run;
        data phi_3_param_5; set  phi_3_param_5 phi_3_parameter_5; run;
        data phi_1_param_20; set phi_1_param_20 phi_1_parameter_20; run;
        data phi_2_param_20; set  phi_2_param_20 phi_2_parameter_20; run;
        data phi_3_param_20; set phi_3_param_20 phi_3_parameter_20;  run;

  %end;



    proc datasets library=work;
         delete cohort risk_sets;
        run;


%end;
*** Clear out titles;
title1;title2;title3;title4;

*** Save data ***;

data steve.param_full; set param_full; run;
data steve.param_5; set param_5; run;
data steve.param_20; set param_20; run;

        data steve.phi_1_param_5; set  phi_1_param_5 ; run;
        data steve.phi_2_param_5; set  phi_2_param_5 ; run;
        data steve.phi_3_param_5; set  phi_3_param_5 ; run;
        data steve.phi_1_param_20; set phi_1_param_20 ; run;
        data steve.phi_2_param_20; set  phi_2_param_20 ; run;
        data steve.phi_3_param_20; set phi_3_param_20 ;  run;

data steve.summary; set casesummary; run;

data steve.rs_summary; set rs_summary; run;
data steve.rs_summary_cont; set rs_summary_cont; run;
data steve.rs_summary_case; set rs_summary_case; run;
```

```
data steve.all_time; set all_time; run;

************* Print the Log and Output in their respective windows ********************;
proc printto;
run;

%mend iterate;



%let lib =;


%iterate(seed       = 0,
         n_cohorts  =,
         n_rep      =,
         n_workers  =,
         delta_0    =,
         exp_method =,
         lag_risk_years =,
         phi        =);
```

Appendix D:

Chapter 4 Simulations: Creating and Analyzing Realistic Occupational Cohorts with Different True Hazard Function Models

```
/*************************************************************************************************************/
/***                                                                                                       ***/
/*** author:  Stephen Bertke/Misty Hein                                                                    ***/
/*** purpose: To generate occupational cohorts under various scenarios for risk. Analyze full cohorts using    ***/
/***          using various models for the underlying risk function.                                       ***/
/*************************************************************************************************************/

%MACRO COX(DATA=ANALYTIC,LOGLIN=,LIN=,INITIAL=,MAX=,TECH=,outparameter=,outaic=);
*** Perform Cox Regression, allowing for a linear model ***************************;

DATA _NULL_;
 LOG_N = 0;
 LIN_N = 0;
 N = 0;

IF "&LOGLIN" NE '' THEN DO;
 A=count("&LOGLIN", ' ');
 LOG_N=A+1;
END;
IF "&LIN" NE '' THEN DO;
 B=count("&LIN", ' ');
 LIN_N=B+1;
END;

%PUT 1 LIN=&LIN;

 N=LOG_N+LIN_N;

 IF "&INITIAL" = '' THEN DO;
   CALL SYMPUT('INITIAL', '0');
   DO I=2 TO N;
     CALL SYMPUT('INITIAL', SYMGET('INITIAL') ||' 0');
   END;
 END;

 call symput('VARIABLES', symget('LOGLIN') ||' '||symget('LIN'));
 CALL SYMPUT('LOG_N',LOG_N);
 CALL SYMPUT('LIN_N',LIN_N);
 CALL SYMPUT('N',N);
RUN;

%LET ARRAY =;
%LET START =;
%LET LOG_VAR =0;
%LET LIN_VAR =0;
%LET LOG_VAR_ONE =0;
%LET LIN_VAR_ONE =0;

%DO I=1 %TO &N;
  %LET VAR=%SCAN(&VARIABLES,&I, %STR( ));
  %LET IN =%SCAN(&INITIAL,&I, %STR( ));
  DATA _NULL_;
        call symput('ARRAY', symget('ARRAY') ||' '||"ARRAY &VAR._AR{&MAX.};");
        call symput('START', symget('START') ||' '||"&VAR.=&IN.");
  RUN;
%END;

%PUT 2 LIN=&LIN;
```

```
%IF &LOG_N NE 0 %THEN %DO;
DATA _NULL_;
  LOG_VAR ='';
  LOG_VAR_ONE ='';
  call symput('LOG_VAR', LOG_VAR);
  call symput('LOG_VAR_ONE', LOG_VAR_ONE);
RUN;
%LET LOG_VAR =;
%LET LOG_VAR_ONE =;
%DO I=1 %TO &LOG_N;
  %LET VAR=%SCAN(&LOGLIN,&I, %STR( ));
  DATA _NULL_;
    IF &I NE &LOG_N THEN DO;
           call symput('LOG_VAR', symget('LOG_VAR') ||' '||"&VAR._AR{i}*&VAR. +");
           call symput('LOG_VAR_ONE', symget('LOG_VAR_ONE') ||' '||"&VAR._AR{1}*&VAR. +");
        END;
        ELSE DO;
           call symput('LOG_VAR', symget('LOG_VAR') ||' '||"&VAR._AR{i}*&VAR.");
           call symput('LOG_VAR_ONE', symget('LOG_VAR_ONE') ||' '||"&VAR._AR{1}*&VAR.");
        END;
  RUN;
%END;
%END;

%PUT 3 LIN=&LIN,;

%IF &LIN_N NE 0 %THEN %DO;
DATA _NULL_;
  LIN_VAR ='';
  LIN_VAR_ONE ='';
  call symput('LIN_VAR', LIN_VAR);
  call symput('LIN_VAR_ONE', LIN_VAR_ONE);
RUN;
%DO I=1 %TO &LIN_N;
  %LET VAR=%SCAN(&LIN,&I, %STR( ));
  DATA _NULL_;
    IF &I NE &LIN_N THEN DO;
           call symput('LIN_VAR', symget('LIN_VAR') ||' '||"&VAR._AR{i}*&VAR. +");
           call symput('LIN_VAR_ONE', symget('LIN_VAR_ONE') ||' '||"&VAR._AR{1}*&VAR. +");
        END;
        ELSE DO;
           call symput('LIN_VAR', symget('LIN_VAR') ||' '||"&VAR._AR{i}*&VAR.");
           call symput('LIN_VAR_ONE', symget('LIN_VAR_ONE') ||' '||"&VAR._AR{1}*&VAR.");
        END;
  RUN;
%END;
%END;


proc nlmixed TECH=&TECH data= &DATA;
   parms &START;
   sum=0;
   &ARRAY;
   array c{&MAX};
   do i = 2 to &MAX; sum=sum + (exp(&LOG_VAR)*(1 + &LIN_VAR))*c{i}; end;
   eta = sum / (exp(&LOG_VAR_ONE)*(1 + &LIN_VAR_ONE));
   p = eta/(1+ eta);
   model time ~ binary(p);
   ods output ParameterEstimates=&outparameter FitStatistics=&outaic;
run;
%PUT VARIABLES = &VARIABLES LOG_N = &LOG_N LIN_N = &LIN_N N = &N;
%MEND;


%MACRO RISKSET2(IN=DATA,VARIABLES=,MAX=,OUT=ANALYTIC);
*** Reformat a data-set so that all observation in a risk-set appear in a single row ***************************;

DATA N;
 M=count("&VARIABLES", ' ');
 N=M+1;
```

```
 CALL SYMPUT('N',N);
RUN;
%LET KEEP =;
%LET ARRAY =;
%LET INITIAL =;
%LET ITERATE =;
%LET caseITERATE =;

%DO I=1 %TO &N;
  %LET VAR=%SCAN(&VARIABLES,&I, %STR( ));
  DATA _NULL_;
    call symput('KEEP', symget('KEEP') ||' '||"&VAR._AR1-&VAR._AR&MAX");
         call symput('ARRAY', symget('ARRAY') ||' '||"ARRAY &VAR._AR{&MAX.};");
         call symput('INITIAL', symget('INITIAL') ||' '||"&VAR._AR(T)=0;");
         call symput('ITERATE', symget('ITERATE') ||' '||"&VAR._AR{i}=&VAR.;");
         call symput('caseITERATE', symget('caseITERATE') ||' '||"&VAR._AR{1}=&VAR.;");
  RUN;
%END;

data &out (keep= case_id time c1-c&MAX &KEEP ); set &IN;
   by case_id;
   &ARRAY;
   array c{&MAX};
   retain i c1-c&MAX &KEEP;

   if first.case_id then do;
      i = 1;
      do t=1 to &MAX;
         &INITIAL; c(t)=0;
      end;
   end;

if case ne 1 then do;
   i = i + 1;
   &ITERATE;

   c{i}=1;
end;
if case = 1 then do;
   i = i + 1;
   &caseITERATE;

   c{1}=1;
end;


   if last.case_id then do;
      time = 0;
      output;
   end;
run;


%MEND;
%macro cohort(cohort);
%put; %put; %put SIM CREATING COHORT NUMBER: &cohort;

*** Randomly assign age at exposure begin (integer), exposure intensity, maximum exposure duration (integer), and
     maximum follow-up (integer);
data cohort1;
  cohort = 1*&cohort;
  do worker = 1 to &n_workers;
    age_exp_begin    = &beta_1 + ROUND(&beta_2*ranexp(&seed),1);
    age_risk_begin   = age_exp_begin;

    if      &exp_method = 1 then do;
        do until (exp_intensity<50);
            exp_intensity =  exp(2.5 + .5*rannor(0));
                 end;
         end;
```

```
    else if &exp_method = 2 then do;
        do until (exp_intensity<50);
            exp_intensity =  exp(.75 + 1*rannor(0));
                 end;
         end;


    else exp_intensity = .;

    %do er = 1 %to 5;
          %if &er=1 or &er=3 or &er=5 %then %do;
       fake_intensity&er. = exp_intensity*exp((&er/10)*rannor(0));
         %end;
         %end;




    max_duration_exp = 15;
    max_follow_up    = &gamma_1 - ROUND(&gamma_2*ranexp(&seed),1);
    if max_follow_up < 1 then max_follow_up = 1;
    output;
    end;
  keep cohort worker age_exp_begin age_risk_begin exp_intensity max_duration_exp max_follow_up fake_intensity1 fake_intensity3
fake_intensity5;
  run;


*** Assign age and cumulative exposure (under the true risk lag) at yearly follow-up intervals;
data cohort2;
  set cohort1;
  by cohort worker;
  retain temp_exp;
  length scenario $3;
    temp_exp   = 0;
    if max_follow_up LE max_duration_exp then do;
      *** Scenario A1: max_follow_up LE max_duration_exp --> follow up ends at or before the end of exposure;
      scenario = 'A1';
      do follow_up_year = 1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
        output;
        end;
      end; *** End A1;
    else /*if max_follow_up GT max_duration_exp then*/ do;
      *** Scenario A2: max_follow_up GT max_duration_exp --> follow up extends beyond the end of exposure;
      scenario = 'A2';
      do follow_up_year = 1 to max_duration_exp;
        temp_age = age_risk_begin + follow_up_year;
        temp_exp = temp_exp + exp_intensity;
        output;
        end;
      temp_exp = temp_exp;
      do follow_up_year = max_duration_exp+1 to max_follow_up;
        temp_age = age_risk_begin + follow_up_year;
        output;
        end;
      end; *** End A2;
  run;

*** Assign hazards for risk of death (h) and censoring (c) and determine case/censor status for each follow-up year;
data cohort3;
  set cohort2;

if &true=1 then do;
  h = min(0.999,exp(&delta_0 + &delta_1*log(temp_age/&age_divisor) + &phi*temp_exp));
end;
if &true=2 then do;
  h = min(0.999,exp(&delta_0 + &delta_1*log(temp_age/&age_divisor))*(1 + &phi*temp_exp));
```

```
end;
if &true=3 then do;
  h = min(0.999,exp(&delta_0 + &delta_1*log(temp_age/&age_divisor) + &phi*log(temp_exp)));
end;

  c = min(0.999,exp(&neta_0 + &neta_1*log(temp_age/&age_divisor)));
  if h LE 0 then case = 0;
  else           case = ranbin(0,1,h);
  if c LE 0 then censor = 0;
  else           censor = ranbin(0,1,c);
  run;

*** Determine case and censor status by selecting the first observation with case=1 or censor=1
    if none then output the last observation;
*** Note that if the first observation with case=1 or censor=1 has both case=1 and censor=1,
    then case status is assigned automatically;
data cohort4;
  set cohort3;
  by cohort worker;
  retain stop;
  if first.cohort or first.worker then stop = 0;
  if stop = 0 then do;
    if case = 1 and censor = 1 then do;         case_status = 1; censor_status = 0; stop = 1; output; end;
    else if case = 1 then do;                   case_status = 1; censor_status = 0; stop = 1; output; end;
    else if censor = 1 then do;                 case_status = 0; censor_status = 1; stop = 1; output; end;
    else if last.cohort or last.worker then do; case_status = 0; censor_status = 0; stop = 1; output; end;
    end;
  run;


*** Compute age at risk end, actual cumulative exposure, time exposed, age at exposure end,
    actual duration of exposure and actual follow-up time;
data cohort5;
  set cohort4;
  age_risk_end = temp_age;
  if age_exp_begin + max_duration_exp < age_risk_end then do;
    *** Exposure ceased prior to risk end so truncation is not necessary;
    cumulative_exp = exp_intensity * max_duration_exp;
    age_exp_end = age_exp_begin + max_duration_exp;
    end;
  else do;
    *** Exposure extends beyond risk end so exposure is truncated at risk end;
    cumulative_exp = exp_intensity * (age_risk_end - age_risk_begin);
    age_exp_end = age_risk_end;
    end;
  time_exposed = age_exp_end - age_exp_begin;
  time_at_risk = age_risk_end - age_exp_begin;
  keep cohort worker age_exp_begin fake_intensity1 fake_intensity3 fake_intensity5
       exp_intensity  censor_status case_status cumulative_exp
       age_risk_begin age_risk_end age_exp_end
       max_duration_exp max_follow_up
       time_exposed time_at_risk;
  run;


*** Create final cohort to use in analyses;
data cohort;
  set cohort5;
  run;

*** Get number of cases in cohort ***;
proc means data=cohort noprint;
 var case_status;
 output out=casesum n=n sum=cases;
run;

*** Clean up datasets;
ods exclude all;
proc datasets library=work;
  delete cohort1 cohort2 cohort3 cohort4 cohort5 cohort_summary_new;
```

```
run; quit; ods select all;
%mend cohort;
*****************************************************************************************************************;
*** RISKSETS macro definition:  create the risk sets for the cohort for use in Cox regression on the full cohort   ***;
***                             and nested case-control analyses                                                    ***;
***                             risk sets are defined based on attained age and attained age pus age at death or   ***;
***                             censor                                                                              ***;
*** input files:  cohort                                                                                           ***;
*** output files: risk_sets                                                                                        ***;
*****************************************************************************************************************;
%macro risksets(cohort);
%put CREATING RISK SETS FOR COHORT NUMBER: &i_cohort;

*** Identify the cases;
data cases;
  set cohort;
  if case_status = 1;
  case_age = age_risk_end;
  case_id = worker;
run;

*** Determine the number of cases and save as a macro variable;
proc means data=cases noprint;
  by cohort;
  var case_status;
  output out=n_cases sum=n_cases;
run;
data n_cases;
  set n_cases;
  call symput('n_cases',n_cases);
run;

*** For each case, identify members of the risk set for both matching on attained age (aacontrol) and ;
*** matching on attained age plus age at death or;
*** censor (dccontrol);
%do i_cases = 1 %to &n_cases;
  data case_n;
    set cases;
    if _n_=&i_cases;
    keep cohort case_age case_id;
    run;
  data risk_set_new;
    set cohort;
    if _n_ = 1 then set case_n;
    *** Select out eligible controls;
    if age_exp_begin LT case_age LE age_risk_end;
    *** Note - LT is important here because of the risk evaluation at yearly intervals;
    *** Identify the cases;
    case = (worker = case_id);
    *** Compute cumulative exposure truncated to the age of the case;
    exp = (exp_intensity)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
         fake1 = (fake_intensity1)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
         fake3 = (fake_intensity3)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
         fake5 = (fake_intensity5)*min((case_age-age_exp_begin),(age_exp_end-age_exp_begin));
run;

proc append base=risk_sets data=risk_set_new force;
run;

%END;

%RISKSET2(IN=risk_sets,VARIABLES=exp fake1 fake3 fake5,MAX=&n_workers,OUT=ANALYTIC);
%COX(DATA=ANALYTIC,LOGLIN=,LIN=exp,INITIAL=&phi, MAX=&n_workers,TECH=trureg,outparameter=parameter_correct_lin,
outaic=fit_correct_lin);
proc phreg data=cohort;
  by cohort;
  model (age_exp_begin, age_risk_end) * case_status(0) = log_exp / ties=breslow rl;
  new_exp1 = (exp_intensity)*min((age_risk_end-age_exp_begin),(age_exp_end-age_exp_begin));
  ods output parameterestimates=parameter_correct_log FitStatistics=fit_correct_log;
  log_exp = log(new_exp1);
```

```
run;
proc phreg data=cohort;
  by cohort;
  model (age_exp_begin, age_risk_end) * case_status(0) = new_exp1 / ties=breslow rl;
  new_exp1 = (exp_intensity)*min((age_risk_end-age_exp_begin),(age_exp_end-age_exp_begin));
  ods output parameterestimates=parameter_correct_exp FitStatistics=fit_correct_exp;
run;

%do er = 1 %to 5;
%if &er=1 or &er=3 or &er=5 %then %do;

%COX(DATA=ANALYTIC,LOGLIN=,LIN=fake&er,INITIAL=0, MAX=&n_workers,TECH=trureg,outparameter=parameter_linfake&er,
outaic=fit_linfake&er);
proc phreg data=cohort;
  by cohort;
  model (age_exp_begin, age_risk_end) * case_status(0) = log_fake / ties=breslow rl;
  fake_new_exp1 = (fake_intensity&er)*min((age_risk_end-age_exp_begin),(age_exp_end-age_exp_begin));
  ods output parameterestimates=parameter_logfake&er FitStatistics=fit_logfake&er;
  log_fake = log(fake_new_exp1);
run;
proc phreg data=cohort;
  by cohort;
  model (age_exp_begin, age_risk_end) * case_status(0) = fake_new_exp1 / ties=breslow rl;
  fake_new_exp1 = (fake_intensity&er)*min((age_risk_end-age_exp_begin),(age_exp_end-age_exp_begin));
  ods output parameterestimates=parameter_expfake&er FitStatistics=fit_expfake&er;
run;

%end;
%end;

%mend risksets;


*****************************************************************************************************************;
*** ITERATE macro definition:  iterate through {create cohort, summarize, define risk sets, Cox regression,      ***;
***                            NCC sampling, and Cox regression                                                   ***;
*** input files:  none                                                                                           ***;
*** output files: Lots!                                                                                          ***;
*****************************************************************************************************************;


%macro iterate(/*************************************************************************************************/
               /*** input parameters that might vary within a set of simulations                                 */
               n_cohorts      = ,            /* number of cohorts                                              */
               n_workers      = ,            /* number of workers per cohort                                   */
               exp_method     = 2 ,          /* exp_method in 1=uniform, 2=lognormal, 3=exponential            */
               phi            = ,            /* risk parameter                                                 */
               true           = ,            /* true hazard function. 1=Log-Lin, 2=Lin, 3=Power,               */
               /***                                                                                            */
               /*** input parameters that should remain constant within a set of simulations                     */
               seed           = 0,           /* initial seed                                                   */
               beta_1         = 18,          /* age at exposure begin parameter                                */
               beta_2         = 10,          /* age at exposure begin parameter                                */
               gamma_1        = 40,          /* max follow-up parameter                                        */
               gamma_2        = 5,           /* max follow-up parameter                                        */
               zeta_1         = 25,          /* max duration of exposure parameter                             */
               delta_0        = ,            /* risk parameter                                                 */
               delta_1        = 1.5,           /* risk parameter                                                 */
               age_divisor    = 55,          /* divisor for age in risk and censoring models                   */
               neta_0         = -5.0,        /* censoring parameter                                            */
               neta_1         = 5.0,         /* censoring parameter                                            */
               ncc_rep        = 1);          /* number of repetitions for nested case-control sampling         */
/*****************************************************************************************************************/

title1 "SIM  Simulating &n_cohorts cohorts of size &n_workers workers using a seed of &seed.";
title3 "Risk parameters include delta_0=&delta_0, delta_1=&delta_1 and phi=&phi.";
title4 "Censoring parameters include neta_0=&neta_0 and neta_1=&neta_1.";

*** Assign Library for data and log to be saved ***;
data _null_;
```

```
  rr = EXP(&phi);
  call symput('rr', trim(left(rr)));
run;
libname steve "&lib\data\&rr";
%put &lib\data\&rr;


filename mylist "&lib\log-output\listing.lst";
filename mylog  "&lib\log-output\log.log";
proc printto log=mylog print=mylist;
run;


data P_correct_lin; run;
data f_correct_lin; run;
data P_correct_log; run;
data f_correct_log; run;
data P_correct_exp; run;
data f_correct_exp; run;

%do er = 1 %to 5;
%if &er=1 or &er=3 or &er=5 %then %do;
data P_linfake&er; run;
data P_expfake&er; run;
data P_logfake&er; run;

data f_expfake&er; run;
data f_linfake&er; run;
data f_logfake&er; run;
%end;
%end;



%do c=0 %to 10;
  data all_cohorts_&c; run;
%end;
data casesummary; run;


*** Iterate by cohort ***;
%do i_cohort = 1 %to &n_cohorts;
    %cohort(&i_cohort);

        data _null_;
         c = floor(&i_cohort/100);
         call symput('c',trim(left(c)));
        run;

        data all_cohorts_&c; set all_cohorts_&c cohort;
          keep cohort worker age_exp_begin age_exp_end age_risk_end case_status
               exp_intensity fake_intensity1 fake_intensity3 fake_intensity5;
        run;
    data casesummary; set casesummary casesum; run;



    %risksets(&i_cohort);

    data P_correct_lin; set P_correct_lin parameter_correct_lin;
                        keep cohort estimate standarderror probt; cohort=&i_cohort; run;
    data f_correct_lin; set f_correct_lin fit_correct_lin; if descr='AIC (smaller is better)'; cohort=&i_cohort; run;
    data P_correct_log; set P_correct_log parameter_correct_log;
                        keep cohort estimate stderr probchisq; cohort=&i_cohort; cohort=&i_cohort; run;
    data f_correct_log; set f_correct_log fit_correct_log; keep cohort criterion withcovariates; if criterion='AIC'; run;
    data P_correct_exp; set P_correct_exp parameter_correct_exp;
                        keep cohort estimate stderr probchisq; cohort=&i_cohort; cohort=&i_cohort; run;
    data f_correct_exp; set f_correct_exp fit_correct_exp; keep cohort criterion withcovariates; if criterion='AIC'; run;

%do er = 1 %to 5;
```

```
%if &er=1 or &er=3 or &er=5 %then %do;
        data P_linfake&er; set P_linfake&er parameter_linfake&er;
                           keep cohort estimate  standarderror probt; cohort=&i_cohort; run;
        data P_logfake&er; set P_logfake&er parameter_logfake&er;
                           keep cohort estimate stderr probchisq; cohort=&i_cohort; run;
        data P_expfake&er; set P_expfake&er parameter_expfake&er;
                           keep cohort estimate stderr probchisq; cohort=&i_cohort; run;

        data f_linfake&er; set f_linfake&er fit_linfake&er; if descr='AIC (smaller is better)'; cohort=&i_cohort; run;
        data f_logfake&er; set f_logfake&er fit_logfake&er;
                           keep cohort criterion withcovariates; if criterion='AIC'; cohort=&i_cohort; run;
        data f_expfake&er; set f_expfake&er fit_expfake&er;
                           keep cohort criterion withcovariates; if criterion='AIC'; cohort=&i_cohort; run;
%end;
%end;

proc datasets; delete risk_sets;
run;

%end;

*** Clear out titles;
title1;title2;title3;title4;

*** Save data ***;
%do er = 1 %to 5;

    data steve.P_correct_lin; set P_correct_lin ; run;
    data steve.f_correct_lin; set f_correct_lin ; run;
         data steve.P_correct_log; set P_correct_log ; run;
    data steve.f_correct_log; set f_correct_log ; run;
         data steve.P_correct_exp; set P_correct_exp ; run;
    data steve.f_correct_exp; set f_correct_exp ; run;

%if &er=1 or &er=3 or &er=5 %then %do;
         data steve.P_linfake&er; set P_linfake&er ;
         data steve.P_logfake&er; set P_logfake&er ;
         data steve.P_expfake&er; set P_expfake&er ;

         data steve.f_linfake&er; set f_linfake&er ;
         data steve.f_logfake&er; set f_logfake&er ;
         data steve.f_expfake&er; set f_expfake&er ;
%end;
%end;




%do c=0 %to 10;
  data steve.all_cohorts_&c; set all_cohorts_&c; run;
%end;
data steve.summary; set casesummary; run;

************* Print the Log and Output in their respective windows *********************;
proc printto;
run;
*** Clear out titles;
title1;


%mend iterate;

%let lib=;



%iterate(seed       = ,
         n_cohorts  = ,
         n_workers  = ,
         true       = ,
```

```
delta_0     = ,
 exp_method = ,
 f_std      = ,
 phi        = );
```