12-2-2013

Sample Size/power Calculation for Stratified Case-cohort Design and Generalized Stratified Case-cohort Design

Wenrong Hu

SAMPLE SIZE/POWER CALCULATION FOR STRATIFIED CASE-COHORT DESIGN AND GENERALIZED STRATIFIED CASE-COHORT DESIGN

by

Wenrong Hu

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Mathematical Sciences

The University of Memphis

December 2013

ACKNOWLEDGEMENTS

I am extremely grateful to my advisor, Dr. Jianwen Cai (University of North Carolina Chapel Hill), whose inspiration, guidance, and commitment from the initial to the final stage facilitated a thorough development of the dissertation.

This thesis would not have been possible without Dr. E. Olusegun George's (University of Memphis) coordination and cooperation. Also I am grateful to Dr. Lih-Yuan Deng, Dr. Manohar L. Aggarwal (proposal committee), and Dr. Dale Bowman Armstrong (dissertation committee) (University of Memphis) for serving on the proposal and dissertation committee and providing their valuable comments and suggestions. My special thanks to Dr. Donglin Zeng (University of North Carolina Chapel Hill) for providing encouragement and guidance throughout this project.

I am indebted to many of my friends, colleagues, and all of those who supported me with the flexible working schedules and technical writing tips in any respect during the completion of the study.

I owe my deepest gratitude to my loving husband Sean J. Egan and family for their endless love, kindness, and support. I would also like to thank my parents for their loving support throughout my life.

ABSTRACT

Hu, Wenrong. PhD. The University of Memphis. December 2013. Sample Size/Power Calculation for Stratified Case-Cohort Design and Generalized Stratified Case-Cohort Design. Major Professor: Dr. E. Olusegun George.

Time to event is a commonly used endpoint in epidemiologic and disease prevention trials in order to study the relationship between risk factors and the endpoint. Case-cohort design that consists of a sub-cohort randomly sampled from full cohort, and all subjects with event is often applied in studies where the disease is rare and the cost of collecting the event information is high. With the non-rare event, a generalized case-cohort design is advocated in which a subset of events instead of all events is sampled. Cai and Zeng have proposed the general log-rank tests and the corresponding sample size/power formulas to compare the hazard rates between two groups under the case-cohort and the generalized case-cohort designs, respectively. However, in many practical situations, the population is not homogenous and stratification is considered. While stratification is increasingly commonly used in large cohorts, the stratified log-rank tests and the sample size and power estimation techniques have not been available even though these issues are critical to the study design. This dissertation is devoted to consider these issues and fulfill the availability. In addition to the development of the stratified general log-rank tests and the sample size/power formulae for both the stratified case-cohort design and the stratified generalized case-cohort design, simulation studies are to be conducted to examine the performance of the tests. Furthermore, optimal, proportional, and balanced sampling strategies are to be explored and recommendations are to be made. Two real epidemiological studies are to be presented to illustrate the sample size calculation under these sampling strategies.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

LIST OF TABLES

1 Introduction

Definition of Designs

Time to event is a commonly used endpoint for the risk factor assessment in epidemiologic studies and disease prevention trials. Case-cohort design (CC), originally proposed by Prentice (1986), has been often used in studying the time to event when the disease is rare and the cost of collecting the event information is expensive. A case-cohort sample consists of a sub-cohort, which is a random sample of the full cohort, and all the subjects with the event. Recently, Cai and Zeng (2007) have considered a generalized case-cohort design (GCC), in which a random sample is selected from the full cohort, and then a random sample is selected from the remaining events.

Stratified case-cohort design (SCC) based on large cohorts has been increasingly used in epidemiologic studies (Breslow et al., 2009). The full cohort can be stratified by a covariate which is available for all cohort members (Boice and Monson, 1977; Hrubec et al., 1989; Langholz and Jiao, 2006); the stratified case-cohort sample consists of the stratified sub-cohort, which is selected by stratified random sampling from the full cohort, and all the rest of the subjects with the event.

In this study we introduce a generalized stratified case-cohort design (GSCC) for a situation when the events are not rare. Given that the information on the stratification factors is known for all cohort members, a stratified random sample is generated from the full cohort and then a stratified random sample is generated from the remaining events in each stratum. These two stratified random samples compose a generalized stratified case-

cohort sample. The difference between GSCC and SCC is that only a fraction of events instead of all events is selected in a GSCC sample.

Examples

An example is provided to illustrate the stratified case-cohort design (Boice and Monson, 1977; Hrubec et al., 1989; Langholz and Jiao, 2006). A full cohort of 1,741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 was studied to investigate the breast cancer risk and the treatments, one of which required radiation exposure and the other did not. The full cohort of 1,741 female patients was stratified by age at first exposure to treatment. Seventy-five breast cancer cases were observed in 1,741 patients. A stratified sub-cohort with a total of 100 subjects was sampled without replacement, in which the number of sampled subjects was proportional to the number of breast cancer cases in each stratum (Langholz and Jiao, 2006). The stratified case-cohort sample contained the 100 subjects from the stratified sub-cohort and all 75 breast cancer cases.

To construct a generalized stratified case-cohort sample, we assume that about 70 breast cancer patients were remaining after the stratified sub-cohort of 100 was sampled from the full cohort, and select a stratified random sample of 35 from the remaining 70 cases. The generalized stratified case-cohort sample then includes the stratified sub-cohort of 100 and the stratified random sample of 35 from the remaining cases.

Literature Review

In this section, we review the relevant literature on the statistical inference methodologies and design issues for CC, GCC, SCC, and GSCC designs.

Statistical Inference Methodologies in Case-Cohort Design

The CC design has been widely used in many studies (Schouten et al., 1993; Liao et al., 1997; Savitz et al., 2000; Beelen et al., 2008; McElrath et al., 2008; Szilagyi et al., 2008; Anderson et al., 2009; Sinner et al., 2010; Herder et al., 2011). The statistical inference methodologies of this design have been developed in many publications (Prentice, 1986; Self and Prentice, 1988; Lin and Ying, 1993; Barlow et al., 1994; Borgan et al., 1995; Chen and Lo, 1999; Chen, 2001; Chen, 2001a; Chen, 2001b). Prentice (1986) proposed a case-cohort design which involved covariate data only for cases experiencing the event and for members of a randomly selected sub-cohort. Relative risk estimation method was provided for case-cohort designed binary response and time to response data. Also proposed was a pseudo-likelihood approach where the risk set at each event consisted of the subjects who were at risk from the sub-cohort. A simulation study was presented to compare the case-cohort relative risk estimation methods to those from full cohort and case-control designs.

The case-cohort design involved the selection of a random sample of the entire cohort, and the assembly of covariate histories only for this random sub-cohort and for all cases. The sub-cohort constituted the comparison set of cases occurring at a range of event times. The sub-cohort also provided a basis for covariate monitoring during the course of cohort follow-up.

For case-cohort design in time to response data, estimation of the relative risk parameter used the pseudo-likelihood function. The only difference from the full cohort partial likelihood function was that the *i*th denominator factor was a sum over subjects at risk in the sub-cohort rather than over subjects at risk in the entire cohort. The maximum pseudo-likelihood estimate was obtained by solving the equation based on the differentiation of the log pseudo-likelihood function. Asymptotic properties of the estimate were derived from those of the score statistic. The variance function for the score statistic was also given. The conditions for convergence to a normal distribution with mean zero and a covariance matrix were further addressed by Self and Prentice (1988).

A small simulation study with several different cohort sizes was conducted to examine the performance of the estimation procedure above, comparing with that of full cohort and synthetic case-control estimation procedures. The standard case-control estimator, denoted case-control I, involved the selection of controls randomly from the entire risk set at each event time. The second estimator, denoted case-control II, excluded future cases from the control selection and allowed a given censored subject to serve as a control for at most one event time. Results were given at relative risk 1 (β = 0) and 2 (β = 0.693). Type I error rate was set at 0.05 and 0.10 for testing β = 0 and 0.693, and Wald test was used. Results from relative risk 1(β = 0) indicated that these designs, except for the case-control I, had similar estimated sample means and standard errors. Results from relative risk 2 (β = 0.693) indicated that the case-cohort sample standard errors were about midway between the full cohort and the corresponding case-control I or II sample standard errors, both at 100 and 300 expected subjects.

Prentice (1986) advocated that case-cohort design could be considered for epidemiologic cohort studies or disease prevention trials in which raw materials for covariate data ascertainment had been stored for all cohort members. Such raw materials might include blood serum samples, tissue specimens, or occupational exposure records. The case-cohort approach could reduce the cost of the costly analysis of raw materials in the assembly of covariate histories, while still allowing the monitoring of such histories on an ongoing basis by means of the sub-cohort. Also it could be considered as an alternative to a case-control design in the presence of a population-based disease registry since the relative risk estimation procedures did not require a cohort roster.

As indicated above, Prentice (1986) proposed a pseudo-likelihood approach to estimate the relative risk where the risk set at each event consisted of the subjects who were at risk in the sub-cohort. Self and Prentice (1988) modified this approach. They defined the risk set at each event time to consist of the subjects who were at risk in the sub-cohort and those who failed at that time. It involved a system for identifying the event times in the entire cohort, but did not require a cohort roster or even an enumeration of the entire cohort. Because the event rate was assumed to be absolutely continuous and at most one event would be from the cases at any time, an individual's contribution to the SP's estimator was negligible and thus SP's estimator $\tilde{\beta}$ was asymptotically equivalent to Prentice's estimator. They further developed the asymptotic distribution theories with sufficient conditions for these estimators. They established the asymptotic normality of the score statistic (3.1) and the relative risk estimator $\tilde{\beta}$ (3.2), and the weak convergence of the cumulative hazard function (3.3) was presented.

The asymptotic efficiency relative to full cohort estimators was computed to measure the efficiency loss as the offset of the cost saving in case-cohort design. Asymptotic relative efficiency (ARE) for the case-cohort pseudo-likelihood estimator $\tilde{\beta}$ relative to the partial likelihood estimator $\hat{\beta}$ for the full cohort was calculated for a special case only containing a single binary covariate and with the exponential relative risk function. It was observed that ARE increased slightly as the relative risk increased (i.e., ARE changed from 0.52 to 0.55 in the situation with the sub-cohort fraction = 0.053). However, it was also observed that ARE had a large increase as the sub-cohort fraction increased (i.e., ARE changed from 0.52 to 0.79 when the sub-cohort fraction increased from 0.053 to 0.158).

An important statement in Self and Prentice (1988) was that the asymptotic normality theory of the score function under CC design was allowed to be generalized to the stratified situation. According to this statement, the score statistic for the stratified case-cohort (SCC) sample can be obtained by summing the score functions in strata, and is asymptotically normally distributed; the asymptotic variance can be estimated by summing the variances at each stratum.

Lin and Ying (1993) proposed an estimating equation approach while considering the case-cohort design as a special case of the missing data problem under the Cox regression model. The estimating equation was constructed by approximating the conditional expectation in the score equation based on the partial likelihood using the data from the subjects who had the complete measurements on all covariate components at time *t*, in other words, from the subjects in the sub-cohort. This estimator was referred to as the approximate partial likelihood estimator (APLE). Also estimated was the corresponding

cumulative hazard function. The asymptotic normality theory was developed for the APLE.

The efficiency of the APLE relative to the maximum partial likelihood estimator (MPLE) using only subjects with full covariate measurements (CC) was investigated in the simulation studies. It was observed that the proposed APLE had a small bias which decreased when the sample size increased. The same was observed for the variance estimator. Furthermore, the APLE method yielded a smaller variance estimator and a higher power for the Wald test, especially when the missingness and censoring were heavy. As compared to the complete-case analysis, the APLE approach had the minimal efficiency loss in large-sample approximations and was adequate for practical use. It tended to be more efficient than CC especially for large cohorts with infrequent events. Two real examples taken from clinical and epidemiologic studies were analyzed.

Data missing completely at random (MCAR), a critical assumption to APLE and associated inference procedures, was satisfied in many situations where missing data was yielded by the design (e.g., case-cohort studies). In situations where data missing at random (MAR) was assumed (Rubin, 1976), the probability of missing on certain components of covariates depended on some completely observed variable. In this case, the authors suggested dividing the range of the completely observed variable into appropriate strata then the MCAR assumption might be reasonable within each stratum. The estimating function was allowed to be generalized for the stratified analysis.

Barlow (1994) developed a simple jackknife method to estimate the variance of the estimated parameter, a robust variance estimation for the case-cohort design. This jackknife estimator of variance was obtained by evaluating the influence of the individual

observation on the overall score for person *i* at time t_0 (Efron, 1982; Barlow and Prentice, 1988), which was shown to be equivalent to the robust variance estimator proposed by Lin and Wei (1989) for the standard Cox model.

A large simulation study was performed in order to investigate the properties of the proposed variance estimator for the case-cohort design with several sizes. The jackknife variance estimator was compared to the naive variance estimator using the partial likelihood method for the case-cohort design. The results showed that the naive standard deviation was underestimated, while the jackknife standard deviation and asymptotic standard deviation had approximately the correct test size, which was defined as the proportion of incorrect rejections to the null hypothesis. In addition, the results of the empirical power and relative efficiency suggested the jackknife method performed better than the naive method and was consistent with the corrected asymptotic estimates.

Chen and Lo (1999) improved Prentice's pseudo-likelihood estimator (Prentice, 1986) by using the information in all case samples completely rather than partially. According to Chen and Lo, the pseudo-likelihood estimating equation was considered as a function of the conditional joint distributions of covariate and observed times for cases or not cases, respectively. Compared to the Prentice's estimator, Chen and Lo's estimator resulted in more accurate estimation of the conditional joint distribution of cases by its empirical analogue for the full cohort than by its empirical analogue for the sub-cohort only. The asymptotic normality of the estimator was developed and the associated asymptotic variance was compared to that of the Prentice's estimator. While both estimators counted the randomness of observations in the full cohort and the randomness of the sub-cohort sampling, Chen and Lo's estimated variance was smaller than

Prentice's because the former did not include the randomness of the cases in the sub-cohort.

Furthermore, the asymptotic relative efficiencies of the Chen and Lo's proposed estimator relative to the Prentice's estimator were evaluated. It was observed that Chen and Lo's estimator was more efficient than Prentice's estimator.

Chen (2001) proposed an optimal sample reuse method via local averaging approach towards efficient estimation and inference for the Cox's regression model for a class of sampling schemes including case-cohort design as the special case. The main idea of sample reuse in constructing the key estimating function was to use a local average of observed covariates to estimate each missing covariate. Specifically, suppose the covariate of the j th individual was not observed, and the event time y_j was in the specified intervals, the partial likelihood estimator for $t \le y_j$ could be obtained by using the observed covariates information from the individuals whose event times were within the specified intervals. Asymptotic normality of the local average estimator was presented. The asymptotic variance function of the local average estimator was developed.

A large simulation study was conducted to compare the local average method with the Prentice (1986) and Chen and Lo (1999) approaches using the case-cohort data. The average of the parameter estimates, the empirical and theoretical standard deviations, and the relative efficiencies were calculated for the case-cohort samples with various sub-cohort sizes and hazard ratios. Relative efficiencies were with respect to Cox's estimator based on the full cohort data. These simulation results demonstrated that the proposed local average method was superior to the estimator of Prentice and that of Chen and Lo,

while Chen and Lo's was observed to be more efficient than Prentice's. In general, the improvement in efficiency was significant in the local average approach over other estimation methods.

Sample Size and Power Estimation in Case-Cohort Design

The issues of sample size and power estimation, considered the key elements at the design stage of a study, were not addressed as much as the coefficient estimation and inferences based on our broad literature review. Cai and Zeng (2004) proposed a log-rank type of test statistics for the case-cohort study with rare events, where one of the test statistic was equivalent to the score test based on the pseudo-partial likelihood function in Self and Prentice (1988). Explicit form for sample size and power calculation were derived based on the proposed tests.

According to Cai and Zeng (2004), the proposed log-rank type test statistic used inverse sampling proportion weighting to approximate the corresponding quantities in the full cohort. The asymptotic variance included the variance in full cohort and variance resulting from the sub-cohort sampling. The variance in full cohort was approximated by using the CC data. Given their proposed test statistic SP_n and the variance $\hat{\sigma}^2_{sp_n}$, the null hypothesis that two groups have the equal cumulative hazard function would be rejected if $n^{-1/2} SP_n \Big/ \sqrt{\hat{\sigma}^2_{sp_n}} < z_\alpha$, where z_α is the critical value of standard normal distribution at the significant level of α. The power function is developed and the assumptions to derive the power function were addressed in the paper (Cai and Zeng, 2004).

A number of numeric studies were conducted to evaluate the log-rank test among the CC, full cohort, and sub-cohort. The empirical type I error and test power were calculated

and compared among these designs. It was observed that the relative efficiency from CC design relative to full cohort was close to 1 when the event rate was low. The sample sizes calculated for the simple random sample and CC sample under the pre-specified significance level α and power β suggested that the CC design was cost-effective.

Statistical Inference Methodologies in Stratified Case-Cohort Design

For SCC studies, most of previous research work was focused on statistical inference for the Cox regression model. Borgan et al. (2000) presented several estimation methods for the analysis of such SCC samples based on the pseudo-likelihood provided by Prentice (1986). The bias and efficiency among these estimation methods were compared to each other and to the randomly sampled CC design. The simulation study results suggested that these estimations performed reasonably well and efficiently.

Borgan et al. (2000) proposed a variant of the stratified case-cohort design (SCC) for a situation that a correlate of the exposure (or prognostic factor) of interest was available for all cohort members, and exposure information was to be collected for a case-cohort sample. In the SCC the cohort was stratified according to the correlate, and the sub-cohort was selected by stratified random sampling. Several analysis methods for the stratified case-cohort samples were presented and the bias and efficiency of these methods were compared to each other and to the randomly sampled case-cohort design. Events in the cohort were assumed to occur according to Cox's (1972) proportional hazards model, while the regression coefficients β_0 measured the effect of the covariates. Three pseudo-likelihood estimators of β_0 were presented and compared in this paper and are given below:

(1) Estimator I was Prentice's estimator naturally generalized to the stratified sampling (Prentice, 1986; Self and Prentice, 1988). The sampled risk set included the sub-cohort only, with the weight as the ratio of the total number of cohort members over the total number of sub-cohorts within the given stratum.

(2) Estimator II was presented by Kalbfleisch and Lawless (1988). The sampled risk set included the sub-cohort members and all cases. For the non-events, the weight was considered as the ratio of the total number of non-events in cohort over the total number of non-events in sub-cohort at each stratum; for the cases (events), the weight was considered as 1.

(3) Estimator III was proposed by the author based on the estimator I. If the case occurred inside the sub-cohort, the risk set and weight were the same as the estimator I; If the cases occurred outside the sub-cohort, the risk set included the sub-cohorts plus the case minus a randomly selected subject from the sub-cohort within the stratum, while the weight was the same regardless of whether the case was inside or outside the sub-cohort.

The score-unbiasedness was investigated in the sense that the expectation of the pseudo-score was exactly equal to zero at the true parameter value. Estimator I and III were score-unbiased while Estimator II was score-biased. The asymptotic distributions for Estimators I and III were derived while that for Estimator II was not. The asymptotic covariance matrix can be estimated using the observed pseudo-information for Estimator I and the empirical covariance matrix based on the sample from stratum l.

Average estimates of β with its empirical standard deviation were obtained based on repeated sampling of 1,000 cohorts each with 1,000 individuals for each of the estimators I, II, and III. Furthermore, estimators were provided for the full cohort, unstratified case-

cohorts corresponding to the analysis methods I, II, and III, and nested case-control, and counter-matched case-control samplings (Borgan et al., 1995). It was observed that all three estimators I, II, and III gave similar results with the difference of little practical importance. All three estimates from the stratified case-cohort had the best performance among all the other cohorts mentioned above.

The simulation results above indicated that if a correlate of exposure was available for all cohort members, it could be advantageous to stratify the sampling of the sub-cohort to over-represent more highly exposed subjects. While the natural generalization of Prentice's (1986) pseudo-likelihood for simple random sampling was clearly inefficient for estimation of rate ratio parameters, Estimator III solved this problem while retaining score-unbiasedness. Another important note was that the data requirements were not the same for all three stratified estimators. Estimator II required the full covariate histories for the cases, while Estimators I and III only needed the cases' covariate values at their event times.

It was concluded that all these analysis methods for the stratified case-cohort samples performed well and were more cost-efficient than the randomly selected sub-cohort. Therefore these methods were recommended for the clinical trials in which subjects entered the study at time zero (at diagnosis or treatment) and a correlate of a prognostic factor was collected for all study subjects at the time of entry of the study.

In stratified case-cohort data analyses, it is popular to use the “robust” approach proposed by Barlow et al. (1994, 1999), in which the case and control observations are weighted by the inverse sampling probabilities for estimating the Cox regression model (Horvitz et al., 1952). However, this approach mainly focuses on the members in the

case-cohort and ignores those not sampled in the case-cohort but available in the full cohort. In order to use the whole cohort in the analysis of case-cohort data, Breslow et al. (2009) proposed a method through adjustment of the sampling weight via calibration or estimation to improve the precision (Deville et al., 1992; Sarndal et al., 1992). The Atherosclerosis Risk in Communities (ARIC) study was used as an example to illustrate Breslow's approach (ARIC Investigators, 1989). The study objective was to estimate the hazard ratio of coronary heart disease (CHD) in relation with the levels of lipoprotein-associated phospholipase A2 (Lp-PLA2) and C-reactive protein (Ballantyne, 2004). A case-cohort sample including 608 cases and 740 non-cases were stratified into 8 strata based on age, sex, and ethnicity (Barlow; 1994; Borgan et al., 2000). The full cohort included a total of 12,345 subjects. In comparison with the robust approach, Breslow re-analyzed the data by the following four steps: 1) applied the case-cohort data to predicted Lp-PLA2 by using a linear regression on race, sex, low density lipoprotein cholesterol, high density lipoprotein cholesterol, systolic and diastolic blood pressures, and the sex X race interaction; 2) used the prediction equation to impute Lp-PLA2 for all cohorts; 3) used the full cohort data containing the imputed Lp-PLA2 variable and other known variables to fit the Cox model, obtained the imputed delta-beta, the estimated influence function contribution for each subject; and 4) used the imputed delta-beta as an auxiliary variable along with the case-cohort data to calibrate or estimate the Cox regression coefficients. The results suggested a dramatic reduction in standard error of estimated coefficients comparing with the Barlow's robust approach.

Sample Size and Power Estimation in Generalized Case-Cohort Design

For the situations that the incidence of disease event is not low, it may not be necessary to include all events in the case-cohort. Cai and Zeng (2007) advocated a GCC design for this situation. The difference from CC design was that only a fraction of events are sampled instead of including all events in the study. A GCC sample contained a sub-cohort sampled at random from full cohort without replacement, and a sample selected at random from the remaining events without replacement. Similar to CC design, the log-rank test statistic for GCC was derived in that the sampled non-events in GCC represented the non-events in full cohort and the events sampled from the remaining events represented the events in full cohort.

Cai and Zeng (2007) showed that the general log-rank test statistic for GCC has asymptotic normality. The asymptotic variance included the variance from the full cohort and those from samplings. The estimated formula was provided in the paper. Given W_n the test statistic and $\hat{\sigma}_w^2$ the estimated asymptotic variance for the test statistic, and assuming that the null hypothesis of two groups had equal cumulative hazard rates in the GCC sample, if $n^{-1/2}W_n \Big/ \sqrt{\hat{\sigma}_w^2} < z_\alpha$, H_0 would be rejected. z_α was the critical value of the standard normal distribution at the significant level of α.

The power function based on the alternative hypothesis was also developed and the lower bound and upper bound were presented when the censoring distribution was unknown. The simulation studies were conducted to compare the testing powers between the GCC, CC, and full cohort designs. In addition, the theoretical power bounds from GCC were compared to the powers estimated from the CC and full cohort designs. It was observed that the GCC method (sampling fraction of events $q = 0.5$) achieved similar

power as the CC method (all events $q = 1$) when the disease event rate is high. The simulation studies also showed that the efficiency loss due to sampling only part of the events under the GCC design was very low when the incidence of disease was not rare.

Weighted Kaplan-Meier and Renyi-Type Permutation Tests

As indicated in the previous sections, we used the log-rank test to detect the difference of hazard function between two groups under the stratified case-cohort design. However, if the alternatives are the stochastically ordered survival, the log-rank test may not necessarily be sensitive (Pepe and Fleming, 1989; Fleming and Harrington, 1991; Cai and Shen, 2000). In this case, we have to look for the appropriate tests that may be sensitive to the alternatives of stochastic ordering.

Pepe and Fleming (1989) introduced a class of statistics based on the integrated weighted difference in Kaplan-Meier estimators for the two-sample censored data problem. Because the Kaplan-Meier statistics are a natural measure of the difference in survival between the two groups, they are intuitive for and sensitive against the alternative of stochastically ordered survival, particularly if the hazard functions cross.

However, because the Kaplan-Meier estimator can be unstable in the presence of heavy censoring, Pepe and Fleming (1989) proposed a weighted Kaplan-Meier or WKM to solve this problem while the weighted functions were constrained to be a function of the censoring survival functions estimators to ensure the stability of the WKM statistics. In practice, the geometric average of the two censoring survival function estimators in sample 1 and 2 or the square root of geometric average are often chosen as the weighted functions.

Furthermore, Pepe and Fleming (1989) compared the WKM against the log-rank statistics by conducting the small-sample simulation studies under the various stochastic ordering configurations including the Weibull proportional hazard alternatives, early hazard difference, late hazard difference, and crossing hazards alternatives. The simulation results suggested that the WKM statistics were more favorable than the log-rank statistics not only under the crossing hazards alternatives but also under the proportional hazards alternatives in the specified simulated samples.

Cai and Shen (2000) proposed a class of two-sample non-parametric permutation tests for comparing marginal survival functions with clustered failure time data. While the individual subjects are independent within the clusters and the two-sample generalized log-rank test statistic and its asymptotic variance were estimable, however, the log-rank test cannot be directly applied to the clustered survival data as it often led to an inflated type I error rate (Cai and Shen, 2000). Instead, Cai and Shen proposed the permutation test by using the log-rank test statistics.

According to the permutation principle, assume N permutation samples were generated based on the observed data, and the log-rank statistics were calculated for each sample and for the observed data. The exact permutation *p*-value for log-rank statistics was obtained as the probability of log-rank statistics of each sample equal to or greater than that of the observed data.

Meanwhile, Cai and Shen (2000) derived Renyi-type test based on the log-rank test statistics. Renyi-type test was defined as the supremum version of the log-rank test statistics at any time t ($t \geqq 0$). Because it presented the maximal deviation at each time t,

Renyi-type method was considered more powerful against a wide range of non-proportional hazards alternatives (Gill, 1980; Fleming et al., 1987;Cai and Shen, 2000).

Renyi-type test statistic was calculated by taking the maximum value among all log-rank statistics at any time *t* in the observed data. For each of N permutation samples that were generated based on the observed data, the Renyi-type test statistic were calculated similarly. The exact permutation *p*-value for Renyi-type statistics was obtained as the probability of Renyi-type statistics of each sample equal to or greater than that of the observed data.

A series of simulation studies were conducted to assess the size and power of the proposed tests above with the configurations of alternative hypotheses including proportional, early, late, and middle differences of the two groups, respectively. Weight functions were considered and assigned for log-rank statistics and Renyi-type statistics appropriately. The simulation results suggested that the proposed permutation tests were valid and desirable comparing with the ordinary log-rank test.

Furthermore, the permutation tests were applied to analyze the trial data from the Hypertension Detection and Follow-up Program (HDFP), the primary objective of which was to compare the survival time for the stepped care versus referred care groups. A total of 10,474 households were enrolled and assigned to either stepped care or referred care group at random. The hypertension patients within a household were considered from the same cluster. The permutation test was compared with the ordinary log-rank test. The *p*-value was increased in the permutation method, which indicated that the correlation of failure times with the households (cluster) was adjusted appropriately.

Study Objectives

For the stratified version of CC and GCC designs, the sample size and power calculation methods based on the log-rank type test have not been addressed before. Without knowing the sample size and power information, we may not be able to plan the SCC and GSCC designs for epidemiologic studies and disease prevention trials. The objective of this PhD study is to provide solutions to these critical issues. In this study, we propose the stratified log-rank statistic, derive the formula for the sample size and power calculation, and evaluate the performance of these formulas for the SCC and GSCC designs. We also address the optimal, proportional, and balanced sampling strategies and provide some practical guidelines. A large number of simulation studies are conducted to evaluate the proposed tests and the computational methods. The real epidemiological studies are presented to illustrate the sample size calculation under the optimal, proportional, and balanced sampling strategies for the SCC and GSCC designs.

2 Sample Size/Power Calculation for Stratified Case-Cohort Design (SCC)

Introduction

Time to event is a commonly used endpoint for risk factor assessment in epidemiologic studies or disease prevention trials (Kalbfleisch and Lawless, 1988; ARIC Investigators, 1989; Schouten et al., 1993; Liao et al., 1997; Savitz et al., 2000; Ballantyne, 2004). Case-cohort design (CC), originally proposed by Prentice (1986), has been often used in studying the time to event when the disease is rare and the cost of collecting the event information is expensive. A case-cohort sample consists of a sub-cohort, which is a random sample of the full cohort, and all the subjects with the event (cases). Statistical analysis methods for data from CC design have been described in many publications (Prentice, 1986; Barlow and Prentice, 1988; Self and Prentice, 1988; Lin and Ying, 1993; Barlow et al., 1994; Borgan et al., 1995; Barlow et al., 1999; Chen and Lo, 1999; Chen, 2001; Chen, 2001a; Chen, 2001b; Kang and Cai, 2009). For the case-cohort study with rare events, Cai and Zeng (2004) proposed a log-rank type of test statistic, in which the test statistic is equivalent to the score test based on the pseudo-partial likelihood function, as described in Self and Prentice (1988). Cai and Zeng (2004) gave an explicit procedure for sample size and power calculation based on the proposed tests.

In most studies, study populations are not homogenous and a stratified case-cohort design (SCC) may be more appropriate (Boice and Monson, 1977; Hrubec et al., 1989; Langholz and Jiao, 2006). The stratified case-cohort sample consists of stratified sub-cohorts selected by stratified random sampling from the full cohort and all the cases. In

an example provided by Langholz and Jiao (2006), a full cohort of 1,741 female patients who were discharged from two tuberculosis sanatoria in Massachusetts between 1930 and 1956 was studied to investigate breast cancer risk and to compare treatments between treatment regimens including and not including radiation exposure. The full cohort of 1,741 female patients was stratified by age at first exposure to treatment. Seventy-five breast cancer cases were observed in 1,741 patients. A stratified sub-cohort with a total of 100 subjects was sampled by age at the first exposure strata. The stratified case-cohort sample thus contained the 100 subjects from the stratified sub-cohort and all 75 breast cancer cases (2006).

Stratified methods for analyzing data from SCC design have been studied extensively (Borgan et al., 2000; Breslow et al., 2009). However, the sample size and power calculations of the SCC design have not been previously addressed. Without the sample size and power information, proper planning of clinical and other studies using a SCC design is impossible. This paper provides solutions to these critical issues. Specifically, we propose a stratified log-rank statistic, derive expressions for sample size and power calculations, and evaluate the performance of proposed statistics. We also address different sampling strategies and provide some practical guidelines. We conduct several simulation studies to evaluate the proposed tests and the computational methods. In addition, we investigate the relative efficiency and cost efficiency of the SCC design against the full cohort and stratified sub-cohort.

Stratified Log-rank Test

Notation

Assume that there are n subjects and L strata in a stratified full cohort, and n_l subjects in stratum l ($l = 1, \ldots, L$). Assume that there are two treatment groups and n_{lj} subjects in group j ($j = 1, 2$) of stratum l. Let T_{lij} represent the event time and C_{lij} the censoring time for subject i in group j and stratum l ($i = 1, \ldots, n_{lj}$; $j = 1, 2$; $l = 1, \ldots, L$), it is reasonable to assume the $T_{lij}s$ are independent of each other. Let J_{lij} be the dichotomous variable indicating the group status, $X_{lij} = T_{lij} \wedge C_{lij}$ be the observed time, where $a \wedge b$ denotes the minimum of a and b, and $\Delta_{lij} = I(T_{lij} \leq C_{lij})$ be the failure indicator, in which $\Delta_{lij} = 1$ denotes observed failure and $\Delta_{lij} = 0$ denotes censoring.

In the SCC design, group labels are measured for all the cases and a stratified sub-cohort sample. Specifically, we assume that $\tilde{n}_l$ subjects are randomly sampled into a sub-cohort from n_l subjects in stratum l, and sub-cohort size is $\tilde{n} = \sum_{l=1}^{L} \tilde{n}_l$. Let $\xi_{lij} = 1$ denote that subject i in group j and stratum l is selected into the sub-cohort and $\xi_{lij} = 0$ otherwise. Let γ_l be the proportion of subjects in group 1 and $(1-\gamma_l)$ the proportion of subjects in group 2 in stratum l. All subjects in the sub-cohort and all events in the L strata make up the stratified case-cohort sample.

Test Statistic

We consider a log-rank type of test to compare the hazard rates between the two groups in SCC. The null hypothesis is H_0: $\Lambda_{l1}(t) = \Lambda_{l2}(t)$, $l = 1, \ldots L$, $t \in [0, \Gamma]$, where Γ

is the length of study period and $\Lambda_{lj}(t)$ the cumulative hazard function of the event time in group j (j = 1, 2) in stratum l. The weighted stratified log-rank test statistic for the full cohort (Self and Prentice, 1988) may be expressed as

$W_n^* = \sum_{l=1}^{L} \int_0^{\tau} \frac{\omega(t)\overline{Y}_{l1}(t)\overline{Y}_{l2}(t)}{\overline{Y}_{l1}(t)+\overline{Y}_{l2}(t)} \left\{ \frac{d\overline{N}_{l1}(t)}{\overline{Y}_{l1}(t)} - \frac{d\overline{N}_{l2}(t)}{\overline{Y}_{l2}(t)} \right\}$, where $\overline{Y}_{lj}(t)$ is the number of subjects at risk and $\overline{N}_{lj}(t)(j=1,2)$ is a counting process representing the number of events at time t in group j and stratum l, and $\omega(t)$ is a weight function. The formula above can also be expressed as

$$W_n^* = \sum_{l=1}^{L} \sum_{i=1}^{n_{l1}} \frac{\Delta_{li1}\omega(X_{li1})\overline{Y}_{l2}(X_{li1})}{\overline{Y}_{l1}(X_{li1})+\overline{Y}_{l2}(X_{li1})} - \sum_{l=1}^{L} \sum_{i=1}^{n_{l2}} \frac{\Delta_{li2}\omega(X_{li2})\overline{Y}_{l1}(X_{li2})}{\overline{Y}_{l1}(X_{li2})+\overline{Y}_{l2}(X_{li2})}. \quad (1)$$

For the full cohort, the log-rank test statistic is known to be the same as the score function of the Cox partial likelihood function (Cai and Zeng, 2004; Self and Prentice, 1988). In the stratified case-cohort sample, covariate information is only available for subjects in the sub-cohort and the cases. Under this situation, we can use the sub-cohort data to approximate $\overline{Y}_{lj}(t)$ by $\widetilde{Y}_{lj}(t)/p_l$, where $\widetilde{Y}_{lj}(t)$ is the number of subjects at risk for group j and stratum l in the sub-cohort, and p_l is the sampling fraction of the sub-cohort in stratum l. Hence, we use these approximations and obtain the stratified case-cohort test statistic

$$W_n \equiv \sum_{l=1}^{L} W_{nl} \equiv \sum_{l=1}^{L} \sum_{i=1}^{n_{l1}} \frac{\Delta_{li1}\omega(X_{li1})\widetilde{Y}_{l2}(X_{li1})}{\widetilde{Y}_{l1}(X_{li1})+\widetilde{Y}_{l2}(X_{li1})} - \sum_{l=1}^{L} \sum_{i=1}^{n_{l2}} \frac{\Delta_{li2}\omega(X_{li2})\widetilde{Y}_{l1}(X_{li2})}{\widetilde{Y}_{l1}(X_{li2})+\widetilde{Y}_{l2}(X_{li2})}, \quad (2)$$

where $\widetilde{Y}_{lj}(t) = \sum_{i=1}^{\widetilde{n}_{lj}} I(X_{lij} \geq t)$, and $\widetilde{n}_{lj}$ is the number of subjects in group j and stratum l in sub-cohort. Since all the quantities in the summation only contribute to the summation if $\Delta_{li1} = 1$ or $\Delta_{li2} = 1$, the above referenced W_n can be obtained based on the observed data.

This test statistic is the score function of the stratified version of the pseudo partial likelihood function and formula (2) can be shown to have an asymptotic normal distribution (Self and Prentice, 1988).

Asymptotic Variance

The asymptotic variance of W_n is the summation of the asymptotic variance of W_{nl} from all the strata. The traditional case-cohort design is considered as a special case of SCC with the number of strata $L = 1$ (Borgan et al., 1995 and 2000; Self and Prentice, 1988). Given the proportion of subjects in group 1 $\gamma_l = n_{l1}/n_l$, $\gamma_l \in (0, 1)$, and assuming that $\tilde{n}_l/n_l$ converges to p_l in stratum l and $v_l = n_l/n$ converges a constant as n goes to ∞ ($p_l = \lim \tilde{n}_l/n_l$), under H_0 and some regularity conditions, $n^{-1/2}W_n$ has an asymptotic normal distribution: $n^{-1/2}W_n \rightarrow_D N(0, \sigma^2 + \psi)$, where $\sigma^2 = \sum_{l=1}^{L} v_l \sigma_l^{\,2}$ and $\psi = \sum_{l=1}^{L} v_l \psi_l$

where $\sigma_l^{\,2}$ and ψ_l correspond to the asymptotic variance of the log-rank test based on stratum l in the full cohort and the variation resulting from sampling stratum l into the sub-cohort. Under the null hypothesis H_0: $\Lambda_{l1}(t) = \Lambda_{l2}(t) = \Lambda_l(t)$, $l = 1, \ldots L$, $t \in [0, \Gamma]$, let $S_l(t) = S_{lj}(t) = P(T_{lj} \geq t)$, and $\pi_{lj}(t) = P(C_{lj} \geq t)$, then by the results in Self and Prentice (1988),

$$\psi_l = \iint \frac{\omega(t)}{\gamma_l \pi_{l1}(t) + (1-\gamma_l)\pi_{l2}(t)} \frac{\omega(w)}{\gamma_l \pi_{l1}(w) + (1-\gamma_l)\pi_{l2}(w)} Q_l(t,w) \frac{dS_l(w)}{S_l(w)} \frac{dS_l(t)}{S_l(t)}, \text{ where}$$

$$Q_l(t,w) = \frac{1-p_l}{p_l} \gamma_l (1-\gamma_l) S_l(t \vee w)[(\gamma_l \pi_{l1}(t)\pi_{l1}(w)\pi_{l2}(t \vee w) + (1-\gamma_l)\pi_{l2}(t)\pi_{l2}(w)\pi_{l1}(t \vee w)]$$

with $a \vee b$ denoting the maximum of a and b.

The estimator for the asymptotic variance for W_n, $\hat{\sigma}^2_{W_n}$, can be derived based on the similar arguments as in Cai and Zeng (2004). Specifically, $\hat{\sigma}^2_{W_n}$ is given by $\hat{\sigma}^2_{W_n} = \hat{\sigma}^2 + \hat{\psi}$, where

$$\hat{\psi} = \frac{1}{n}\sum_{l=1}^{L} 2(1-\hat{p}_l)\sum_{j=1}^{2}\sum_{i=1}^{n_{lj}}\left\{\frac{\Delta_{lij}\omega(X_{lij})\widetilde{Y}_{l1}(X_{lij})\widetilde{Y}_{l2}(X_{lij})}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^2}\times\sum_{j'=1}^{2}\sum_{i'=1}^{n_{lj'}}\frac{\Delta_{li'j'}\omega(X_{li'j'})I(X_{li'j'}\le X_{lij})}{\widetilde{Y}_{l1}(X_{li'j'})+\widetilde{Y}_{l2}(X_{li'j'})}\right\}$$

$$-\frac{1}{n}\sum_{l=1}^{L}(1-\hat{p}_l)\sum_{j=1}^{2}\sum_{i=1}^{n_{lj}}\frac{\Delta_{lij}\omega(X_{lij})^2\widetilde{Y}_{l1}(X_{lij})\widetilde{Y}_{l2}(X_{lij})}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^3}, \qquad (3)$$

where $\hat{p}_l = \widetilde{n}_l / n_l$ is the estimate of p_l, and $\hat{\sigma}^2$ is the estimate of σ^2 given by

$$\hat{\sigma}^2 = \frac{1}{n}\left\{\sum_{l=1}^{L}\sum_{i=1}^{n_{l1}}\frac{\Delta_{li1}\omega(X_{li1})\widetilde{Y}_{l2}(X_{li1})^2}{(\widetilde{Y}_{l1}(X_{li1})+\widetilde{Y}_{l2}(X_{li1}))^2}+\sum_{l=1}^{L}\sum_{i=1}^{n_{l2}}\frac{\Delta_{li2}\omega(X_{li2})\widetilde{Y}_{l1}(X_{li2})^2}{(\widetilde{Y}_{l1}(X_{li2})+\widetilde{Y}_{l2}(X_{li2}))^2}\right\}.$$

Since all the quantities expressed above contribute to $\hat{\sigma}^2$ and $\hat{\psi}$ only when $\Delta_{li1} = 1$ or $\Delta_{li2} = 1$, $\hat{\sigma}^2_{W_n}$ can be obtained from the observed data. The derivations are given in Appendix A.

To test the equality of the cumulative hazard function of the event time between the two groups in SCC, i.e., to test the null hypothesis H_0: $\Lambda_{l1}(t) = \Lambda_{l2}(t)$, $l = 1, \ldots L$, $t \in [0, \Gamma]$ vs. the alternative hypothesis H_A: $\Lambda_{l1}(t) \neq \Lambda_{l2}(t)$ (two-sided), $l = 1, \ldots L$, $t \in [0, \Gamma]$ at the significance level α, we reject H_0 if $\left| n^{-1/2} W_n \Big/ \sqrt{\hat{\sigma}^2_{W_n}} \right| > z_{1-\alpha/2}$, where z_α is the $(100\alpha)^{\text{th}}$ percentile of the standard normal distribution.

Sample Size and Power Estimation

The sample size and power estimation formula is derived and simplified based on the alternative hypothesis H_A: $\Lambda_{l1}(t) = e^{\theta}\Lambda_{l2}(t)$ where $\theta = O(1/\sqrt{n})$. Assume the following

conditions in the observed data: (i) the censoring distributions are the same in the two groups; (ii) the number of events is very small in the full cohort but much larger than one; and (iii) there are no ties of event times, i.e. all the observed event times are different. For the sample size and power calculation, we consider the test statistic with $\omega(t) = 1$. Under the alternative hypothesis H_A, the asymptotic expectation of $n^{-1/2}W_n$ is the same as the asymptotic expectation of the usual log-rank test statistic for the full cohort under H_A and can be approximated by $n^{-1/2}\sum_{l=1}^{L}\int_0^{\Gamma}\frac{\bar{Y}_{l1}(t)\bar{Y}_{l2}(t)}{\bar{Y}_{l1}(t)+\bar{Y}_{l2}(t)}[d\Lambda_{l1}(t)-d\Lambda_{l2}(t)]$

$\approx n^{-1/2}\sum_{l=1}^{L}\theta(1-\gamma_l)D_{l1}$, where D_{lj} is the total number of failures in group $j\,(j=1,2)$ in stratum *l*. $\hat{\sigma}^2$ can be approximated by $1/n\sum_{l=1}^{L}((1-\gamma_l)^2 D_{l1}+\gamma_l{}^2 D_{l2})$ following the exact approximation and algebra as Cai and Zeng (2004) for each stratum. To simplify $\hat{\psi}$, we assume that failures are very few, and approximate $\sum_{j=1}^{2}\bar{Y}_{lj}(t)$ by $(n_l - D_l/2)$, where $D_l = D_{l1} + D_{l2}$. Since the size of risk set in stratum *l* of the sub-cohort is about p_l times the size of the risk set in stratum *l* of the full cohort, $\hat{\psi}$ can be approximated by $\frac{1}{n}\sum_{l=1}^{L}\frac{(1-p_l)}{(n_l - D_l/2)p_l}\gamma_l(1-\gamma_l)(D_{l1}+D_{l2})^2$. Hence, the non-centrality parameter for $n^{-1/2}W_n / \sqrt{\hat{\sigma}^2_{W_n}}$ under the alternative is approximately

$$\frac{n^{-1/2}\sum_{l=1}^{L}\theta(1-\gamma_l)D_{l1}}{\sqrt{1/n\sum_{l=1}^{L}((1-\gamma_l)^2 D_{l1}+\gamma_l{}^2 D_{l2})+1/n\sum_{l=1}^{L}\frac{(1-p_l)}{(n_l-D_l/2)p_l}\gamma_l(1-\gamma_l)(D_{l1}+D_{l2})^2}}$$ which can be simplified as $$\frac{n^{1/2}\theta\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}{\sqrt{\sum_{l=1}^{L}\left\{\gamma_l(1-\gamma_l)p_{Dl}v_l(1+\frac{(1-p_l)}{(1-p_{Dl}/2)p_l}p_{Dl})\right\}}},$$ where p_{Dl} is the failure

proportion in stratum l and v_l is the proportion of stratum l in the full cohort ($v_l = n_l / n$). Hence, the power function is

$$\Phi\left(z_{\alpha/2} + n^{1/2}|\theta| \frac{\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}{\sqrt{\sum_{l=1}^{L}\left\{(\gamma_l(1-\gamma_l)p_{Dl}v_l)(1+\frac{1-p_l}{(1-p_{Dl}/2)p_l}p_{Dl})\right\}}} \right), \quad (4)$$

where n is the number of subjects in L strata in the stratified full cohort, θ the log hazard ratio, α the significance level, p_{Dl} the failure proportion in stratum l, v_l the proportion of stratum l, γ_l the proportion of subjects in group 1 and $(1-\gamma_l)$ the proportion of subjects in group 2 in stratum l, p_l the sub-cohort sampling fraction in stratum l, $l = 1, \ldots L$. By dropping $p_{Dl}/2$, the formula (4) can be further simplified as

$$\Phi\left(z_{\alpha/2} + n^{1/2}|\theta| \frac{\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}{\sqrt{\sum_{l=1}^{L}\{(\gamma_l(1-\gamma_l)p_{Dl}v_l)(1+(1/p_l-1)p_{Dl})\}}} \right).$$

When $L = 1$, the above function can be further simplified as

$\Phi\left(z_{\alpha/2} + \tilde{n}^{1/2}|\theta|\sqrt{\frac{\gamma(1-\gamma)p_D}{p+(1-p)P_D}} \right)$, in which $\tilde{n} = np$. It is same to the power function for the case-cohort design in Cai and Zeng's paper (2004). When $p_l = 1$, the power of the stratified log-rank test for full cohort is

$$\Phi(z_{\alpha/2} + n^{1/2}|\theta|\sqrt{\textstyle\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}). \quad (5)$$

Proportional, Balanced, and Optimal Designs

We consider the power issues for the proportional and balanced two commonly used designs in this section. We also consider an allocation strategy which maximizes the power.

Proportional Design

Proportional design is commonly used in stratified studies. Under proportional design, the number of subjects in the sub-cohort at each stratum is proportional to the size of the stratum in the population. For example, consider the full cohort size $n = 2{,}000$, and there are 4 strata with the strata proportion of 0.1, 0.2, 0.3 and 0.4, respectively. Or, equivalently, there are 200, 400, 600, and 800 subjects in the 4 strata, respectively. The sub-cohort consists of 200 subjects. With the proportional sampling method, the numbers of samples in each stratum are 20, 40, 60, and 80, respectively. Under such a design, the sub-cohort sampling proportions are the same for all strata, i.e., $p_l = p$ for $l = 1, \ldots, L$.

To detect a log hazard ratio of θ with power β and significance level α, the required total sub-cohort size is $\tilde{n} = \left[\frac{n\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}^2 v_l /\left(1-p_{Dl}/2\right)\right)}{B_2^2 - \sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}v_l\left(1-p_{Dl}/\left(1-p_{Dl}/2\right)\right)\right)} \right]$, where [x] denotes the smallest integer that is bigger than x and $B_2 = \frac{n^{1/2}\theta\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}{z_{1-\alpha/2}+z_\beta}$. The sampling proportion $p = \frac{\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}^2 v_l /\left(1-p_{Dl}/2\right)\right)}{B_2^2 - \sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}v_l\left(1-p_{Dl}/\left(1-p_{Dl}/2\right)\right)\right)}$. The required

number of subjects in stratum *l'*

$$\widetilde{n}_{l'} = pnv_{l'} = \left[\frac{nv_{l'} \sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl}^2 v_l / \left(1 - p_{Dl}/2\right)\right)}{B_2^2 - \sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl} v_l \left(1 - p_{Dl} / \left(1 - p_{Dl}/2\right)\right)\right)} \right], l' = 1,2,..., L. \qquad (6)$$

The total SCC sample size

$$n_{SCC} = \left[n \sum_{l=1}^{L} p_{Dl} v_l + \frac{n \sum_{l=1}^{L} \left((1-p_{Dl}) v_l\right) \sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl}^2 v_l / \left(1 - p_{Dl}/2\right)\right)}{B_2^2 - \sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl} v_l \left(1 - p_{Dl} / \left(1 - p_{Dl}/2\right)\right)\right)} \right].$$

Note that the denominator in (6) needs to be positive. This condition can be re-written as $|\theta| > \theta_0 \equiv (z_{1-\alpha/2} + z_\beta) \dfrac{\sqrt{\sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l) - \sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl}^2 v_l / \left(1 - p_{Dl}/2\right)\right)}}{n^{1/2} \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l)}$.

Since the failure rate p_{Dl} is usually very small for the case-cohort studies,

$p_{Dl} - p_{Dl}^2 / \left(1 - \dfrac{p_{Dl}}{2}\right) \approx p_{Dl}$. Hence, $\theta_0 \approx \dfrac{(z_{1-\alpha/2} + z_\beta)}{n^{1/2} \sqrt{\sum_{l=1}^{L} \left(\gamma_l (1-\gamma_l) p_{Dl} v_l \right)}}$, the log-hazard ratio that can be detected with the entire cohort. This condition implies that the case-cohort design will not be able to detect a hazard ratio smaller than the one that can be detected by using the entire cohort.

Balanced Design

Another popular design for stratified samples is balanced design. To apply this idea to SCC, we allow the number of subjects in the sub-cohort be the same across the strata. For example, the full cohort size n is 2,000 with 4 strata. If we require a total of 200 subjects for the sub-cohort, with the balanced sampling method, each stratum would contain 50 sampled subjects. Under the balanced sampling $\widetilde{n}_l = \widetilde{n} / L$ and $p_l = \widetilde{n}_l / n_l = \widetilde{n} / Lnv_l$, where

L is the number of strata. To detect a log hazard ratio θ with power β and significance level α, the required total sub-cohort size $\tilde{n}$ is at least

$$\left[\frac{Ln\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}^2v_l^2/\left(1-p_{Dl}/2\right)\right)}{B_2^2-\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}v_l\left(1-p_{Dl}/\left(1-p_{Dl}/2\right)\right)\right)}\right]$$ with the balanced method. The number of subjects in each stratum of the sub-cohort is

$$\tilde{n}_l=\left[\frac{n\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}^2v_l^2/\left(1-p_{Dl}/2\right)\right)}{B_2^2-\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}v_l\left(1-p_{Dl}/\left(1-p_{Dl}/2\right)\right)\right)}\right]. \tag{7}$$

The total SCC size is

$$n_{SCC}=\left[n\sum_{l=1}^{L}p_{Dl}v_l+\frac{n\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}^2v_l^2/\left(1-p_{Dl}/2\right)\right)\left(L-\sum_{l=1}^{L}p_{Dl}\right)}{B_2^2-\sum_{l=1}^{L}\left(\gamma_l(1-\gamma_l)p_{Dl}v_l\left(1-p_{Dl}/\left(1-p_{Dl}/2\right)\right)\right)}\right].$$

In order to apply the balanced sampling method, the smallest stratum in the full cohort needs to be larger than the required number of subjects in the sub-cohort for each stratum. For instance, if a full cohort has 100, 500, 700, and 1,000 subjects in 4 strata, and a total of 500 subjects needs to be sampled into the sub-cohort. By the balanced method, 125 subjects need to be selected from each stratum. However, the number of 125 exceeds the total number of subjects (i.e., 100) in the first stratum. Consequently, the balanced method is not applicable in this situation.

Optimal Design

In many studies, the number of subjects that can be included in sub-studies is limited because of financial and resource constraints. In these studies, we are given the total number of subjects to be included in the sub-cohort. The distribution of the number of subjects to each of the stratum needs to be determined. We consider an optimal design

strategy which provides the highest power under such situation. Specifically, we propose an optimal design with a set of p_l *(*l = 1, 2, …,*L)* which provides the highest power for a given $\tilde{n}$. This optimization problem can be solved by using Lagrange multipliers method.

Maximizing the power function for a given $\tilde{n}$ in formula (4) is equivalent to minimizing the denominator $\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)(1+\frac{1-p_l}{(1-p_{Dl}/2)p_l}p_{Dl})$, a function of p_l, subject to $\sum_{l=1}^{L}p_l v_l = \tilde{n}/n,$ a constraint function of p_l. We obtain the Lagrange function $\Lambda(p_l,\lambda)=\sum_{l=1}^{L}\left((\gamma_l(1-\gamma_l)p_{Dl}v_l)(1+\frac{1-p_l}{(1-p_{Dl}/2)p_l}p_{Dl})\right)+\lambda*\left(\sum_{l=1}^{L}(p_l v_l)-\frac{\tilde{n}}{n}\right)$. Furthermore, we have

$$\frac{\partial\Lambda(p_l,\lambda)}{\partial p_l}=\frac{\partial\sum_{l=1}^{L}\left((\gamma_l(1-\gamma_l)p_{Dl}v_l)(1+\frac{1-p_l}{(1-p_{Dl}/2)p_l}p_{Dl})\right)}{\partial p_l}+\lambda*\frac{\partial\left(\sum_{l=1}^{L}(p_l v_l)-\frac{\tilde{n}}{n}\right)}{\partial p_l}$$

$$=-\frac{\gamma_l(1-\gamma_l)p_{Dl}v_l p_{Dl}/(1-p_{Dl}/2)}{p_l^2}+\lambda*v_l=0,\text{ and}$$

$$\frac{\partial\Lambda(p_l,\lambda)}{\partial\lambda}=\frac{\partial\lambda\left(\sum_{l=1}^{L}p_l v_l-\tilde{n}/n\right)}{\partial\lambda}=\sum_{l=1}^{L}p_l v_l-\frac{\tilde{n}}{n}=0.$$

By solving the above two equations, we obtain the optimal sub-cohort sampling proportion

$$p_l=\frac{\tilde{n}\sqrt{\gamma_l(1-\gamma_l)/(1-p_{Dl}/2)}p_{Dl}}{n\sum_{l=1}^{L}(\sqrt{\gamma_l(1-\gamma_l)/(1-p_{Dl}/2)}p_{Dl}v_l)}. \quad (8)$$

Hence, the optimal power for a given $\tilde{n}$ can be calculated as

$$\Phi\left(Z_{\alpha/2}+\frac{\tilde{n}^{1/2}|\theta|\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}{\sqrt{\tilde{n}/n\sum_{l=1}^{L}((\gamma_l(1-\gamma_l)p_{Dl}v_l(1-p_{Dl}/(1-p_{Dl}/2)))+\left(\sum_{l=1}^{L}\left(\sqrt{\gamma_l(1-\gamma_l)/(1-p_{Dl}/2)}p_{Dl}v_l\right)\right)^2}}\right).$$

To achieve power β with a significance level α based on the optimal design, the required total sub-cohort size is given by

$$\tilde{n} = \left[\frac{n\left(\sum_{l=1}^{L} \sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl} v_l\right)^2}{B_2^2 - \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l (1 - p_{Dl}/(1-p_{Dl}/2)))} \right], \text{ where } B_2 = \frac{n^{1/2}\theta \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l)}{z_{1-\alpha/2} + z_\beta}.$$

$$\text{Therefore, } p_l = \frac{\left(\sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl}\right)\left(\sum_{l=1}^{L} \sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl} v_l\right)}{B_2^2 - \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l (1 - p_{Dl}/(1-p_{Dl}/2)))},$$

$$\tilde{n}_l = p_l n v_l = \left[\frac{n v_l \left(\sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl}\right)\left(\sum_{l=1}^{L} \sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl} v_l\right)}{B_2^2 - \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l (1 - p_{Dl}/(1-p_{Dl}/2)))} \right], \quad (9)$$

$$n_{SCC} = \left[n \sum_{l=1}^{L} p_{Dl} v_l + \frac{n\left(\sum_{l=1}^{L} \sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl} v_l (1-p_{Dl})\right)\left(\sum_{l=1}^{L} \sqrt{\gamma_l (1-\gamma_l)/(1-p_{Dl}/2)}\, p_{Dl} v_l\right)}{B_2^2 - \sum_{l=1}^{L} (\gamma_l (1-\gamma_l) p_{Dl} v_l (1 - p_{Dl}/(1-p_{Dl}/2)))} \right].$$

From formula (8), we observe that when γ_l is the same over strata, and p_{Dl} is homogeneous over strata, then the optimal p_l is $\tilde{n}/n$ which is same across strata. The sub-cohort size at stratum *l* is proportional to the size of stratum *l*. On another note, when γ_l is the same over strata, i.e., there are L equally partitioned strata, and the event rate p_{Dl} is very small (disease is rare), then the optimal p_l is proportional to the event rate p_{Dl}. The following section addresses the homogeneous event rate situation in greater detail.

Relative Efficiency and Cost Efficiency

Because the SCC contains less information than the full cohort data, the log-rank test from SCC is less efficient than the test from the full cohort if the full cohort data is available. To evaluate the efficiency of SCC, we compare the asymptotic variances between the SCC and the full cohort. The relative efficiency of SCC compared to the full

cohort (Ω) is defined as the ratio of the asymptotic variances of the full cohort (σ^2) over that of the SCC design ($\sigma^2+\psi$). By applying the estimation formula for $\hat{\sigma}^2$ and $\hat{\psi}$, we obtained an estimator

$$\hat{\Omega}=\frac{\hat{\sigma}^2}{\hat{\sigma}^2+\hat{\psi}}=\frac{\frac{1}{n}\sum_{l=1}^{L}((1-\gamma_l)^2 D_{l1}+\gamma_l{}^2 D_{l2})}{\frac{1}{n}\sum_{l=1}^{L}((1-\gamma_l)^2 D_{l1}+\gamma_l{}^2 D_{l2})+\frac{1}{n}\sum_{l=1}^{L}\frac{(1-p_l)}{(n_l-D_l/2)p_l}\gamma_l(1-\gamma_l)(D_{l1}+D_{l2})^2}$$

$$=\frac{\frac{1}{n}\sum_{l=1}^{L}((1-\gamma_l)^2\gamma_l p_{Dl}n_l+\gamma_l{}^2(1-\gamma_l)p_{Dl}n_l)}{\frac{1}{n}\sum_{l=1}^{L}((1-\gamma_l)^2\gamma_l p_{Dl}n_l+\gamma_l{}^2(1-\gamma_l)p_{Dl}n_l)+\frac{1}{n}\sum_{l=1}^{L}\frac{(1-p_l)}{(n_l-p_{Dl}/2*n_l)p_l}\gamma_l(1-\gamma_l)(p_{Dl}n_l)^2}$$

$$=\frac{\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}n_l)}{\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}n_l)+\sum_{l=1}^{L}\frac{(1-p_l)}{p_l(1-p_{Dl}/2)}\gamma_l(1-\gamma_l)p_{Dl}^2 n_l}. \quad (10)$$

It is also of interest to examine the cost saved by SCC compared to random samples. We assume that the cost for collecting the data for each individual is the same regardless of the design. The cost efficiency is then calculated as the ratio of the required sample sizes for the stratified random sample and SCC to achieve the same power.

The required number of subjects to achieve power β with significance level α for the stratified random sample design n_{SRS} can be obtained based on formula (5), i.e., $n_{SRS}=\frac{(z_{1-\alpha/2}+z_\beta)^2}{\theta^2\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}$. With a stratified case-cohort design, the number of subjects needed for evaluating exposure status is equal to the size of the sub-cohort plus the

additional failures at each stratum for all strata, which is $n_{SCC} = \sum_{l=1}^{L}(\tilde{n}_l + (1-p_l)p_{Dl}\tilde{n}_l / p_l)$.

Hence, the cost efficiency *R* is obtained by

$$R = \frac{n_{SRS}}{n_{SCC}}, \tag{11}$$

where n_{SCC} is dependent on the design strategy. n_{SCC} for the proportional, balanced, or optimal design are given in the previous sections. If $n_{SRS} > n_{SCC}$, $R > 1$, fewer subjects are needed in the SCC design than in the stratified random sample design. In other words, $R > 1$ represents cost savings for SCC design. Larger *R* is associated with more savings.

Numeric Results

Theoretical Power

Based on the power formula derived in the previous sections, Table 1 shows the theoretical power of SCC, as well as the power based on the full cohort and the sub-cohort. The power function (4) is used to calculate P_{SCC}, the power of SCC, while formula (5) is used to calculate P_{Full}, the power of full cohort. The sub-cohort power P_{Sub} is obtained by substituting *n* with $\tilde{n}$ in the full cohort power function, where $\tilde{n}$ is the sub-cohort size, $\tilde{n} = \sum_{l=1}^{L}\tilde{n} = \sum_{l=1}^{L} n_l p_l = n\sum_{l=1}^{L} v_l p_l$. The power P_{Full}, P_{SCC}, and P_{Sub} are calculated for different combinations of the full cohort size n, event proportion p_{Dl}, group 1 proportion γ_l, log-hazard ratio θ, and sub-cohort sampling fraction p_l in stratum *l*. The significant level is set at $\alpha = 0.05$ and the number of strata is $L = 4$. The event proportion p_D in the table is an average proportion over strata. Specifically, at the level of

10%, p_{Dl} over the strata are set to 9%, 8%, 11%, and 10% for each stratum. Similarly, 4%, 5%, 4.5%, and 6% are set for p_D = 5%; 0.8%, 1%, 1.2%, and 0.9% are set for p_D = 1%. In the example where the full cohort size n = 2,000, the event proportion p_D = 5%, the group 1 proportion γ_1 = 0.3, and the log-hazard ratio θ = 0.5, the SCC with the 10% sub-cohort sampling yields the power of 0.479 against the full cohort power of 0.643 and that of the stratified random sample 0.110. In another example where the full cohort size n = 10,000, the event proportion p_D = 1%, the group 1 proportion γ_1 = 0.3, and the log-hazard ratio θ = 0.5, the SCC sample with the 1% sub-cohort sampling yields the power of 0.365 while the powers for the full cohort and for the stratified random sample are 0.630 and 0.042, respectively. The results from the theoretical power comparisons in Table 1 suggest that the SCC design is efficient and an attractive solution in situations with low event proportions and small sub-cohort sampling fractions.

Table 1. Theoretical Power of Stratified Case-Cohort Design

n	p_D	γ_l	θ	p_l	P_{Full}	P_{SCC}	P_{Sub}
2,000	10%	0.3	0.5	10%	0.894	0.634	0.172
				20%	0.894	0.769	0.300
			1.0	10%	1.000	0.996	0.527
				20%	1.000	1.000	0.818
		0.5	0.5	10%	0.938	0.710	0.197
				20%	0.938	0.836	0.347
			1.0	10%	1.000	0.999	0.600
				20%	1.000	1.000	0.879
	5%	0.3	0.5	10%	0.643	0.479	0.110
				20%	0.643	0.559	0.179
			1.0	10%	0.996	0.968	0.312
				20%	0.996	0.988	0.548
		0.5	0.5	10%	0.718	0.548	0.124
				20%	0.718	0.633	0.205
			1.0	10%	0.999	0.986	0.361
				20%	0.999	0.996	0.621
4,000	5%	0.3	0.5	1%	0.908	0.256	0.051
				2%	0.908	0.406	0.067
			1.0	1%	1.000	0.742	0.096
				2%	1.000	0.931	0.152
		0.5	0.5	1%	0.948	0.296	0.055
				2%	0.948	0.468	0.073
			1.0	1%	1.000	0.813	0.107
				2%	1.000	0.964	0.172
	1%	0.3	0.5	1%	0.305	0.174	0.035
				2%	0.305	0.218	0.040
			1.0	1%	0.826	0.533	0.047
				2%	0.826	0.657	0.061
		0.5	0.5	1%	0.352	0.199	0.036
				2%	0.352	0.251	0.041
			1.0	1%	0.885	0.606	0.050
				2%	0.885	0.732	0.065

(Continued)

Table 1. (continued)

n	p_D	γ_l	θ	p_l	P_{Full}	P_{SCC}	P_{Sub}
10,000	5%	0.3	0.5	1%	0.999	0.541	0.075
				2%	0.999	0.777	0.110
			1.0	1%	1.000	0.985	0.179
				2%	1.000	1.000	0.312
		0.5	0.5	1%	1.000	0.615	0.082
				2%	1.000	0.844	0.124
			1.0	1%	1.000	0.995	0.205
				2%	1.000	1.000	0.361
	1%	0.3	0.5	1%	0.630	0.365	0.042
				2%	0.630	0.464	0.051
			1.0	1%	0.996	0.898	0.067
				2%	0.996	0.962	0.095
		0.5	0.5	1%	0.705	0.421	0.044
				2%	0.705	0.532	0.054
			1.0	1%	0.999	0.941	0.072
				2%	0.999	0.983	0.105

Note. n = full cohort size, p_D = average event proportion, γ_l = group 1 proportion, θ = log-hazard ratio, p_l = sub-cohort sampling fraction in stratum l. P_{SCC} = theoretical power of SCC, P_{Full} = theoretical power of full cohort, and P_{Sub} = theoretical power of sub-cohort. Significant level $\alpha = 0.05$.

Empirical Type I Error and Power of Stratified Log-rank Test

Simulation studies are conducted to evaluate the empirical type I error and empirical power for the stratified log-rank test using the SCC, full cohort, and sub-cohort data.

Table 2 shows the empirical type I error for the stratified log-rank test using the SCC (T_{SCC}), full cohort (T_{Full}), and sub-cohort data (T_{Sub}). The significance level α is set at 0.05 and the number of strata $L = 4$. Various values are considered for the full cohort size n, stratum proportion v_l, average event proportion p_D, group 1 proportion γ_l, and sub-cohort sampling fraction p_l in stratum *l.* The following procedures/parameters are set up for the simulation:

- There are 4 strata in the full cohort with size n = 2,000, 4,000, or 10,000, with the stratum proportions of 0.1, 0.2, 0.3, and 0.4;
- All subjects are assigned to one of the two groups and the group 1 proportion γ_l (0.3 or 0.5) is the same over 4 strata;
- The average event proportion p_D (1%, 5%, or 10%) and sub-cohort sampling proportion p_l (0.1, 0.2, 0.01, or 0.02) are the same over 4 strata;
- The event time is generated from the exponential distribution with $\lambda_{l1} = \lambda_{l2}$ at each stratum with the values of 0.1, 0.2, 0.3, and 0.5 for stratum 1, 2, 3, and 4, respectively;
- The censoring time is generated from a uniform distribution between $[0, \Gamma]$, where Γ is varied with different censoring proportions in strata based on the given event proportions;

- The proposed stratified log-rank test for SCC is programmed in SAS. SAS procedure PROC LIFETEST for the stratified log-rank test is used for the full cohort and the sub-cohort data analysis; and
- Each simulation is repeated 2,000 times.

For n = 2,000, we considered the situation with the average event proportion p_D to be 5% or 10%. From the results in the first block of Table 2, we note that the empirical type I error rates in SCC are fairly close to the nominal 0.05 level. For n = 4,000, we considered the smaller average event proportion p_D (1% or 5%). In a couple of cases with small sub-cohort sampling proportion (1%), the empirical type I error rates are much higher than the nominal level. For example, the average event proportion p_D = 5%, the group 1 proportion γ_1 = 0.3, and sub-cohort sampling fraction p_1 = 1% in all strata and shows that the empirical type I error rates for SCC T_{SCC} = 0.107 meanwhile for full cohort T_{Full} = 0.049 and for sub-cohort T_{Sub} = 0.039. The results occur in these cases because some simulated SCC samples may contain no event in at least one of the strata. The numbers of such cases are large when the full cohort size is small, and the event rate and sampling fraction are low. However, the empirical type I error rate for SCC T_{SCC} is improved to 0.067 after the sample fraction is increased to 2% from 1%. Also the empirical type I error rate for SCC T_{SCC} is improved to 0.067 when the full cohort size is increased from 4,000 to 10,000.

When n is increased to 10,000, the empirical type I error rates are fairly close to the nominal level, especially with the 2% sampling rate. In some situations with low event rates (1%) and small sub-cohort sampling proportions (1% or 2%), a simulated stratified

sub-cohort sample could contain no events in all strata. Under this situation, the stratified log-rank test cannot be conducted for the sub-cohort. We excluded these cases when reporting the results for the sub-cohort. This is not an issue for the SCC sample, since by design all events in each stratum are included. In addition, we observed that the empirical type I error is nearly always slightly higher than the nominal level of 0.05 for SCC. This could be due to the small number of events in the setups considered in the simulations. We note that with the same disease proportions, as the sample size increases, the empirical error is closer to the nominal level.

Table 3 presents the empirical power for log-rank tests in SCC, full cohort, and sub-cohort. In addition, the theoretical power P_{SCC} is presented in comparison with the empirical power in SCC. The simulated samples are generated similarly to Table 2, with the exception that the event time is generated from the exponential distribution with $\lambda_{l1} =$ 1.5 λ_{l2} at each stratum which gives the values of 0.15, 0.3, 0.45, and 0.75 for stratum 1, 2, 3, and 4, respectively, while the λ_{l2} remains the same. A number of SCC samples are generated with different combinations of the full cohort size, event rate, and sub-cohort sampling rate in order to explore the efficiency of SCC power relative to the full cohort. The results in Table 3 indicate that the test based on the SCC design is more powerful than using the sub-cohort sample only and the power based on the full cohort provides an upper bound. In real studies, the full cohort is not available. In many cases in Table 3, using only a small fraction of the subjects, the power based on the SCC design is over 50% of the power with the full cohort, where n_{SCC} is calculated as $n_{scc} = n\sum_{l=1}^{L}(p_l v_l + (1-p_l)p_{Dl}v_l)$. As expected, when the sampling rate p_l increases, the

power for SCC increases. Overall, the empirical power T_{SCC} is very close to the theoretical powers P_{SCC}.

Table 2. Empirical Type I Error of Stratified Case-Cohort Design

n	γ_l	p_D	p_l	T_{Full}	T_{SCC}	T_{Sub}
2,000	0.3	10%	10%	0.058	0.057	0.049
			20%	0.058	0.054	0.046
		5%	10%	0.059	0.051	0.045
			20%	0.059	0.050	0.052
	0.5	10%	10%	0.059	0.059	0.056
			20%	0.059	0.043	0.045
		5%	10%	0.049	0.055	0.049
			20%	0.049	0.048	0.054
4,000	0.3	5%	1%	0.049	0.107	0.039
			2%	0.049	0.069	0.045
		1%	1%	0.056	0.069	0.040
			2%	0.056	0.050	0.044
	0.5	5%	1%	0.050	0.103	0.035
			2%	0.050	0.064	0.036
		1%	1%	0.051	0.070	0.018
			2%	0.051	0.051	0.009
10,000	0.3	5%	1%	0.053	0.067	0.043
			2%	0.053	0.062	0.049
		1%	1%	0.051	0.050	0.036
			2%	0.051	0.050	0.043
	0.5	5%	1%	0.057	0.068	0.040
			2%	0.057	0.059	0.048
		1%	1%	0.050	0.049	0.005
			2%	0.050	0.055	0.018

Note: n = full cohort size, p_D = average event proportion, γ_l = group 1 proportion, p_l = sub-cohort sampling fraction in stratum l. T_{SCC} = empirical type I error of SCC, T_{Full} = empirical type I error of full cohort, and T_{Sub} = empirical type I error of sub-cohort. Significant level $\alpha = 0.05$.

Table 3. Simulated Testing Power of Stratified Case-Cohort Design

n	γ_l	p_D	p_l	T_{Full}	T_{SCC}	T_{Sub}	P_{SCC}	$\frac{T_{SCC}}{T_{Full}}$	$\frac{n_{SCC}}{n}$
2,000	0.3	10%	10%	0.804	0.441	0.160	0.469	0.55	19.0%
			20%	0.804	0.579	0.265	0.597	0.72	28.0%
		5%	10%	0.514	0.312	0.112	0.336	0.61	14.5%
			20%	0.514	0.366	0.158	0.395	0.71	24.0%
	0.5	10%	10%	0.861	0.484	0.147	0.538	0.56	19.0%
			20%	0.861	0.647	0.276	0.672	0.75	28.0%
		5%	10%	0.557	0.350	0.083	0.389	0.63	14.5%
			20%	0.557	0.439	0.154	0.455	0.79	24.0%
4,000	0.3	5%	1%	0.771	0.250	0.065	0.186	0.32	6.0%
			2%	0.788	0.296	0.081	0.288	0.38	6.9%
		1%	1%	0.257	0.145	0.048	0.130	0.56	2.0%
			2%	0.257	0.142	0.051	0.159	0.55	3.0%
	0.5	5%	1%	0.796	0.238	0.021	0.213	0.30	6.0%
			2%	0.851	0.303	0.058	0.333	0.36	6.9%
		1%	1%	0.269	0.143	0.016	0.146	0.53	2.0%
			2%	0.269	0.155	0.014	0.181	0.58	3.0%
10,000	0.3	5%	1%	0.991	0.384	0.089	0.392	0.39	6.0%
			2%	0.990	0.555	0.090	0.601	0.56	6.9%
		1%	1%	0.504	0.233	0.062	0.260	0.46	2.0%
			2%	0.504	0.301	0.071	0.330	0.60	3.0%
		0.5%	1%	0.260	0.152	0.046	0.188	0.58	1.5%
			2%	0.260	0.162	0.053	0.216	0.62	2.5%
	0.5	5%	1%	0.994	0.379	0.063	0.452	0.38	6.0%
			2%	0.994	0.571	0.093	0.677	0.57	6.9%
		1%	1%	0.549	0.251	0.014	0.300	0.46	2.0%
			2%	0.549	0.352	0.021	0.382	0.64	3.0%
		0.5%	1%	0.279	0.180	0.009	0.215	0.65	1.5%
			2%	0.279	0.214	0.006	0.249	0.77	2.5%

Note: n = full cohort size, p_D = average event proportion, γ_l = group 1 proportion, p_l = sub-cohort sampling fraction in stratum l. T_{SCC} = simulated testing power of SCC, T_{Full} = simulated testing power of full cohort, and T_{Sub} = simulated testing power of sub-cohort, P_{SCC} = theoretical power of SCC, n_{SCC} = SCC sample size, Significant level $\alpha = 0.05$.

Sample Size and Power of Proportional, Balanced, and Optimal Designs

We compare the proportional, balanced, and optimal sampling methods in order to investigate which one is more efficient in the SCC design. Two situations where the event rates are relatively homogeneous or heterogeneous over the strata are considered for comparison. In the situation where the event rates are homogeneous, the event proportion p_{Dl} at each stratum is relatively similar to each other. In the situation where the event rates are heterogeneous, the event proportions p_{Dl} over the strata have a wide range. The corresponding analysis results in both homogeneous and heterogeneous situations are presented in Table 4.

In SCC with homogeneous event rates, a theoretical power based on proportional, balanced, and optimal sampling for SCC with various combinations of the full cohort size n, the event proportion p_{Dl}, the group 1 proportion γ_l, the log hazard ratio θ, and the sub-cohort size $\tilde{n}$ is presented. The number of strata is $L = 4$ with the stratum proportions (v_l) of 0.1, 0.2, 0.3, and 0.4, respectively. The event proportion p_D in the table is an average value over all strata. Specifically, at the level of 10%, p_{Dl} over the strata are set to 9%,8%, 11%, and 10% for each stratum. Similarly, 4%, 5%, 4.5%, and 6% are set for p_D = 5%; 0.8%, 1%, 1.2%, and 0.9% are set for p_D = 1%. The sub-cohort sampling fractions p_l in stratum l for the proportional, balanced, and optimal designs are calculated by $\tilde{n}/n, \tilde{n}/Lnv_l$, and the formula (8), respectively. The total SCC sizes n_{scc}*(prop),* n_{scc} *(bal),* and n_{scc}*(opt)* are then calculated using the formula $n\sum_{l=1}^{L}(p_l v_l + (1-p_l)p_{Dl}v_l)$. The

theoretical powers P_{prop} , P_{Bal} , and P_{opt} are calculated using the power formula (4). The power ratio (P_{Bal} vs. P_{prop}) is presented in percent (%).

Table 4 indicates that the total SCC sample sizes from the three methods are similar in general under homogeneous circumstances. For instance, where the full cohort size n = 2,000, the event proportion p_D = 10%, the group 1 proportion γ_l = 0.3, the log hazard ratio θ = 0.5, and the stratified sub-cohort size = 200, the total SCC sample sizes are 376, 377, and 376 for proportional, balanced, and optimal samplings, respectively. The results show that the power from proportional method P_{prop} is at least equal to or larger than P_{Bal} in all the situations and the power ratio (P_{Bal} vs. P_{prop}) has a range from 83% to 100%. These results suggest that the proportional sampling is more efficient than the balanced sampling when the event rates are homogeneous over the strata. Furthermore, we observe that the powers from the proportional method and the optimal design remain close, which indicates that the proportional method is close to the optimal sampling strategy when the event rates are homogeneous and the treatment group 1 proportion γ_l is the same over strata.

Table 4 also provides results for situations with heterogeneous event rates over strata. The set-up of the SCC samples is similar to the homogeneous situation, except that the event rates are set to a wide range over strata. We consider three sets of combination of p_{Dl} (l = 1, 2, 3, and 4). Set 1 gives the values of p_{Dl} to 9%, 30%, 5%, and 20% for the 4 strata; Set 2 gives the values of p_{Dl} to 4%, 25%, 10%, and 6%; and Set 3 to 0.8%, 10%, 2%, and 30% for the 4 strata, respectively. Table 4 indicates that for the given set up and given $\tilde{n}$, the total SCC sample sizes from the proportional and balanced methods are

similar in heterogeneous situation. The power for these two methods is similar with slightly more power for the balanced method in most of the situation considered in Set1 and Set2. In the situations where the sub-cohort sampling fraction p_l is small (1% or 2%) and the event rates exhibit a wide range (Set3), the proportional method results in more power. As expected, the optimal design yields the highest theoretical power (P_{opt}) with the smallest total SCC sample size among all three methods. For instance, where the full cohort size n = 2,000, the event proportion p_{Dl} as in Set1, the group 1 proportion γ_l = 0.3, the log hazard ratio θ = 0.5, and the stratified sub-cohort size = 200, the powers (n_{scc}) are 0.637 (495), 0.590 (496), and 0.731 (485) for the proportional, balanced, and optimal samplings, respectively. Thus, the optimal design indeed is the best in comparison with the proportional and balanced sampling strategies under the heterogeneous event rate situation.

Table 4. Proportional, Balanced, and Optimal Sampling in SCC Samples

with Homogeneous Event Rates

					Proportional			Balanced		$\frac{P_{Bal}}{P_{prop}}$	Optimal	
n	p_D	γ_1	θ	$\tilde{n}$	p_1	n_{scc}	P_{prop}	n_{scc}	P_{Bal}		n_{scc}	P_{Opt}
2,000	10%	0.3	0.5	200	10%	376	0.634	377	0.581	92%	376	0.637
				400	20%	557	0.769	558	0.732	96%	556	0.770
			1	200	10%	376	0.996	377	0.991	100%	376	0.996
				400	20%	557	1.000	558	0.999	100%	556	1.000
		0.5	0.5	200	10%	376	0.710	377	0.656	93%	376	0.713
				400	20%	557	0.836	558	0.804	96%	556	0.838
			1	200	10%	376	0.999	377	0.997	100%	376	0.999
				400	20%	557	1.000	558	1.000	100%	556	1.000
	5%	0.3	0.5	200	10%	293	0.479	293	0.442	92%	292	0.482
				400	20%	482	0.559	484	0.533	95%	482	0.561
			1	200	10%	293	0.968	293	0.952	98%	292	0.969
				400	20%	482	0.988	484	0.983	100%	482	0.988
		0.5	0.5	200	10%	293	0.548	293	0.507	93%	292	0.551
				400	20%	482	0.633	484	0.606	96%	482	0.635
			1	200	10%	293	0.986	293	0.977	99%	292	0.987
				400	20%	482	0.996	484	0.994	100%	482	0.996
4,000	5%	0.3	0.5	40	1%	244	0.256	244	0.214	84%	244	0.260
				80	2%	282	0.406	282	0.345	85%	282	0.411
			1	40	1%	244	0.742	244	0.646	87%	244	0.750
				80	2%	282	0.931	282	0.877	95%	282	0.935
		0.5	0.5	40	1%	244	0.296	244	0.246	83%	244	0.300
				80	2%	282	0.468	282	0.399	85%	282	0.474
			1	40	1%	244	0.813	244	0.722	89%	244	0.820
				80	2%	282	0.964	282	0.926	96%	282	0.966
	1%	0.3	0.5	40	1%	80	0.174	80	0.162	93%	80	0.175
				80	2%	119	0.218	119	0.207	95%	119	0.220
			1	40	1%	80	0.533	80	0.494	93%	80	0.537
				80	2%	119	0.657	119	0.629	96%	119	0.660
		0.5	0.5	40	1%	80	0.199	80	0.184	93%	80	0.200
				80	2%	119	0.251	119	0.238	95%	119	0.253
			1	40	1%	80	0.606	80	0.565	93%	80	0.610
				80	2%	119	0.732	119	0.704	96%	119	0.735

(continued)

Table 4. (continued)

with Heterogeneous Event Rates

n	p_D	γ_l	θ	$\tilde{n}$	Proportional			Balanced		$\frac{P_{Bal}}{P_{prop}}$	Optimal	
					p_l	n_{scc}	P_{prop}	n_{scc}	P_{Bal}		n_{scc}	P_{Opt}
2,000	Set1	0.3	0.5	200	10%	495	0.637	496	0.590	93%	485	0.731
				400	20%	662	0.837	664	0.803	96%	641	0.894
			1	200	10%	495	0.996	496	0.992	100%	485	0.999
				400	20%	662	1.000	664	1.000	100%	641	1.000
		0.5	0.5	200	10%	495	0.712	496	0.665	93%	485	0.802
				400	20%	662	0.894	664	0.867	97%	641	0.938
			1	200	10%	495	0.999	496	0.998	100%	485	1.000
				400	20%	662	1.000	664	1.000	100%	641	1.000
	Set2	0.3	0.5	200	10%	394	0.553	393	0.574	104%	384	0.663
				400	20%	573	0.732	571	0.748	102%	553	0.809
			1	200	10%	394	0.987	393	0.990	100%	384	0.997
				400	20%	573	0.999	571	1.000	100%	553	1.000
		0.5	0.5	200	10%	394	0.627	393	0.649	104%	384	0.739
				400	20%	573	0.803	571	0.818	102%	553	0.871
			1	200	10%	394	0.995	393	0.997	100%	384	0.999
				400	20%	573	1.000	571	1.000	100%	553	1.000
4,000	Set2	0.3	0.5	40	1%	468	0.197	467	0.209	106%	466	0.269
				80	2%	503	0.334	503	0.355	106%	499	0.452
			1	40	1%	468	0.600	467	0.632	105%	466	0.766
				80	2%	503	0.865	503	0.888	103%	499	0.957
		0.5	0.5	40	1%	468	0.226	467	0.240	106%	466	0.310
				80	2%	503	0.387	503	0.410	106%	499	0.519
			1	40	1%	468	0.676	467	0.708	105%	466	0.834
				80	2%	503	0.917	503	0.934	102%	499	0.980
	Set3	0.3	0.5	40	1%	621	0.168	623	0.123	73%	617	0.262
				80	2%	655	0.285	659	0.202	71%	646	0.450
			1	40	1%	621	0.513	623	0.360	70%	617	0.754
				80	2%	655	0.796	659	0.613	77%	646	0.956
		0.5	0.5	40	1%	621	0.191	623	0.139	72%	617	0.303
				80	2%	655	0.330	659	0.232	70%	646	0.517
			1	40	1%	621	0.585	623	0.416	71%	617	0.823
				80	2%	655	0.860	659	0.689	80%	646	0.980

Note. p_D = average event proportion, γ_l = group 1 proportion, θ = log-hazard ratio, $\tilde{n}$ = sub-cohort size, n_{SCC} = SCC sample size, p_l = sub-cohort sampling rate in stratum l. P_{prop}(P_{Bal} , P_{Opt}) = proportional (balanced, optimal) power. *Set1 (Set2; Set3):* event proportion = 9%, 30%, 5%, and 20% (4%, 25%, 10%, and 6%; 0.8%, 10%, 2%, and 30%) for strata 1-4.

Relative Efficiency

Figure 1 displays the relative efficiencies of SCC vs. full cohort over the sub-cohort sampling fractions p_l changing from 0 to 1. The event proportions p_{Dl} of 0.01, 0.05, 0.1, and 0.2 and full cohort size of 1,000 are considered for analysis. The other parameters are set up similarly to Table 1. The formula (10) is used to calculate the relative efficiencies. It is observed that the relative efficiency becomes larger while the event proportion is smaller. The relative efficiency is near 1 once the event proportion is at 0.05.

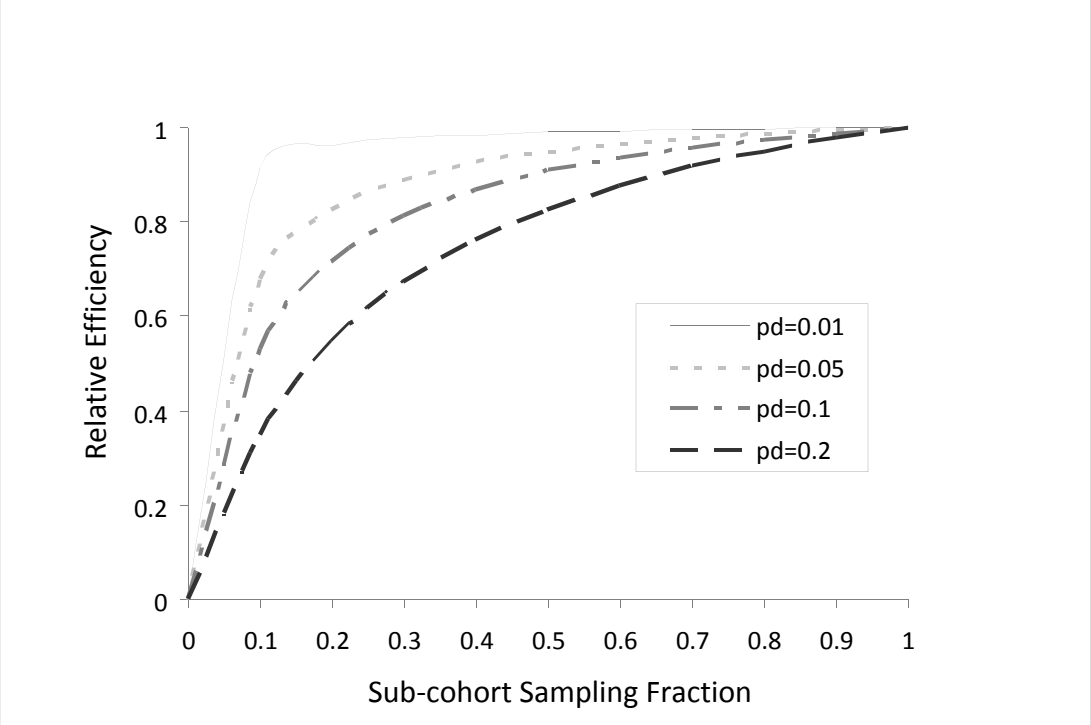
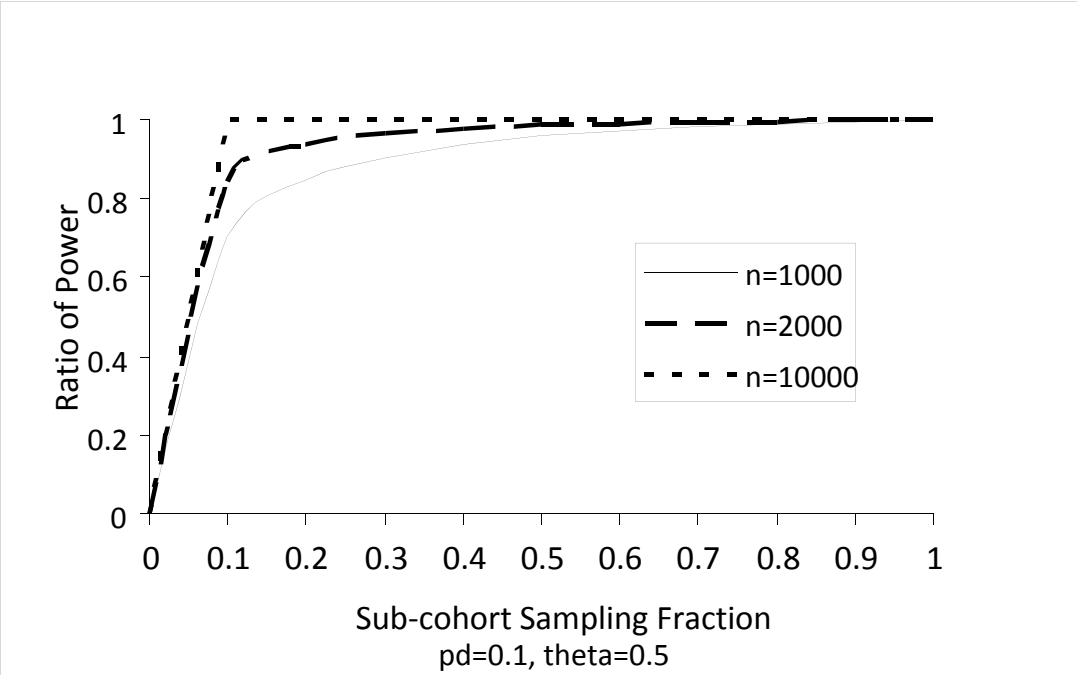


Figure 1 Relative Efficiency of Stratified Case-Cohort vs. Full Cohort

Theoretical Power Ratio

Figure 2 provide a set of theoretical power ratios between the SCC and full cohort with sampling proportions changing from 0 to 1. The power ratios are examined with the full cohort size n = 1,000, 2,000, or 10,000, θ (theta) = 0.5 or 1, γ_l = 0.5, and p_{Dl} (pd) = 0.1 or 0.2. The other parameters are set up similarly to Figure 1. The figures show that the power ratio is close to 1 in most of the situations even though the event rate is low, and the sampling proportion is as small as 0.1. These plots indicate that the SCC with the small sampling proportion yields similar powers as the full cohort. When the event proportion p_{Dl} or the log-hazard ratio θ increases, the power ratio increases, which implies the SCC yields a higher power close to the full cohort.



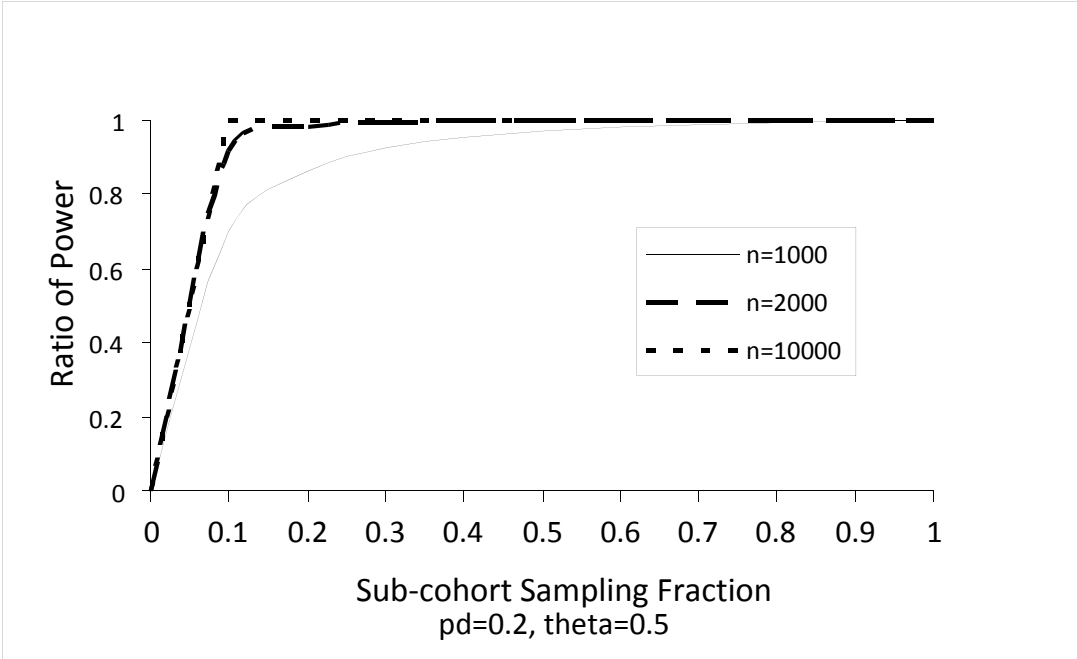
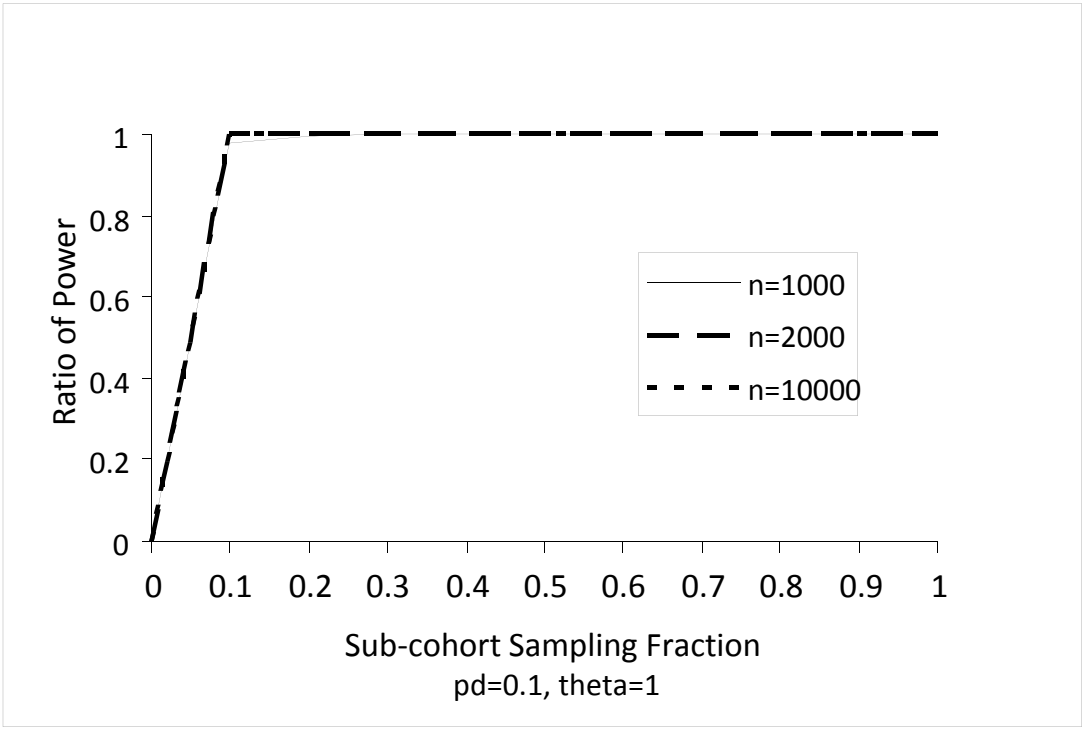


Figure 2 Power of Stratified Case-Cohort vs. Full Cohort



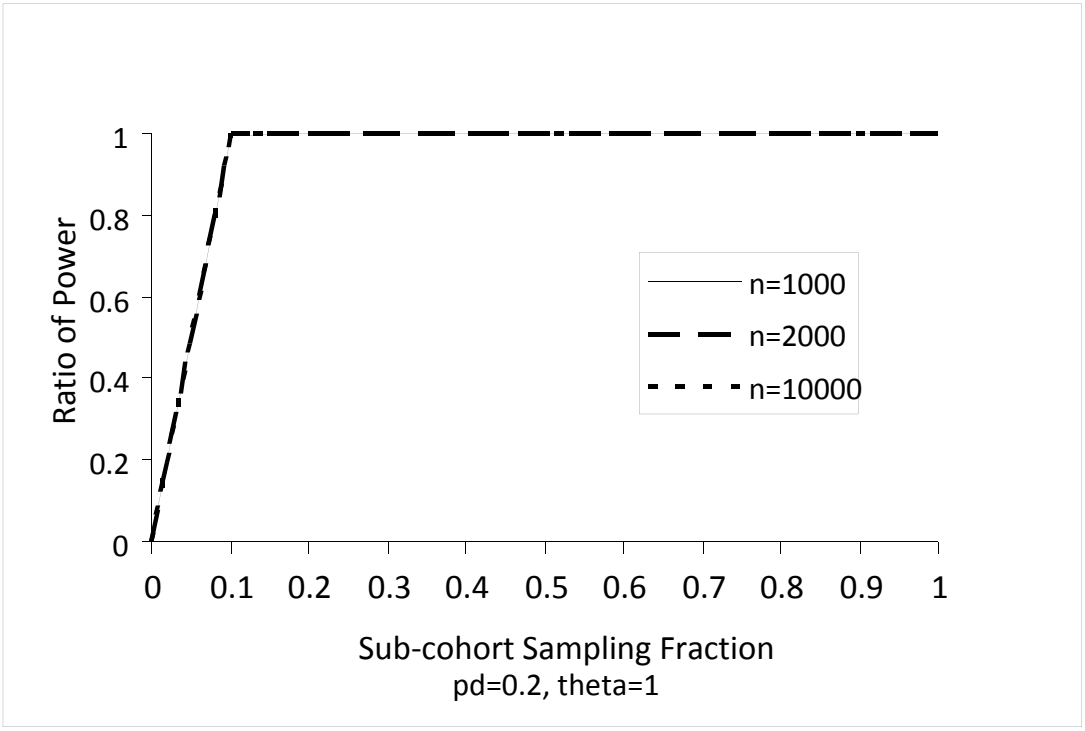
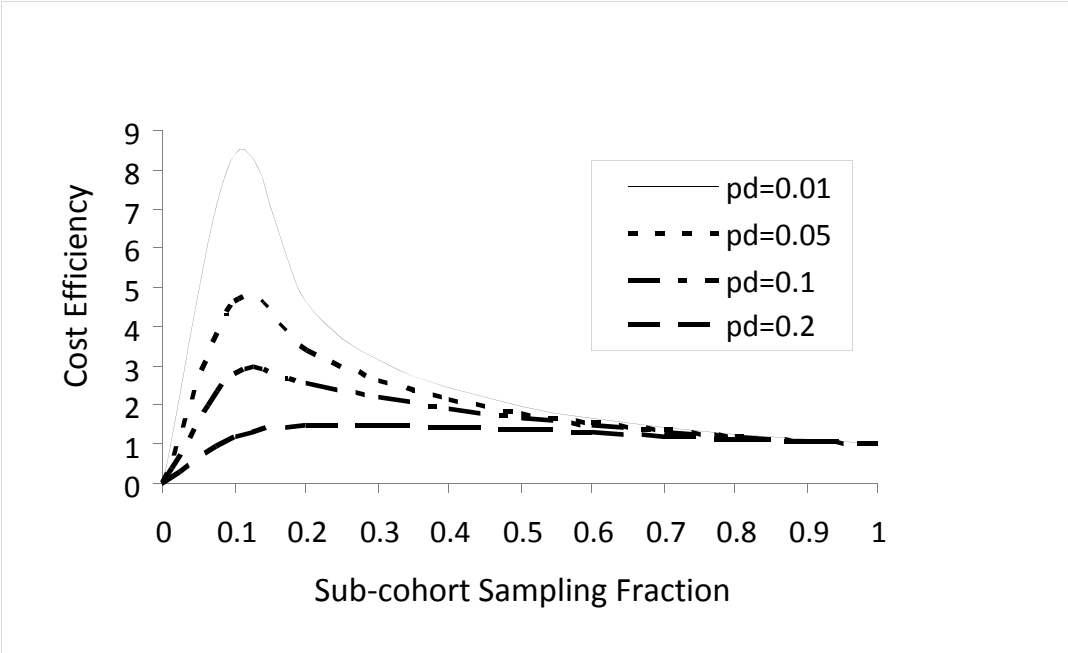


Figure 2 Power of Stratified Case-Cohort vs. Full Cohort

Cost Efficiency

Figure 3 provides the cost efficiencies vs. sampling proportion changing from 0 to 1 in different SCC samples with event rates of 0.01, 0.05, 0.1 and 0.2. The other parameters are set up similarly to Figure 1. Formula (11) is used for calculation. The first plot of Figure 3 shows that the lower event rate is associated with higher efficiency for SCC. When the sampling proportion is small (< 0.2), the SCC design is much more efficient than the stratified simple random sample. When the sampling proportion is large (>= 0.6), the cost of the two designs are similar.

The second plot of Figure 3 shows the efficiency region of SCC, in which the area below the line represents the region that the total sample size for SCC is smaller than the stratified random sample to achieve the same power. We note that the most cost saving occurs when the event rate is lower than 0.3 and the sampling fraction rate is smaller than 0.2.



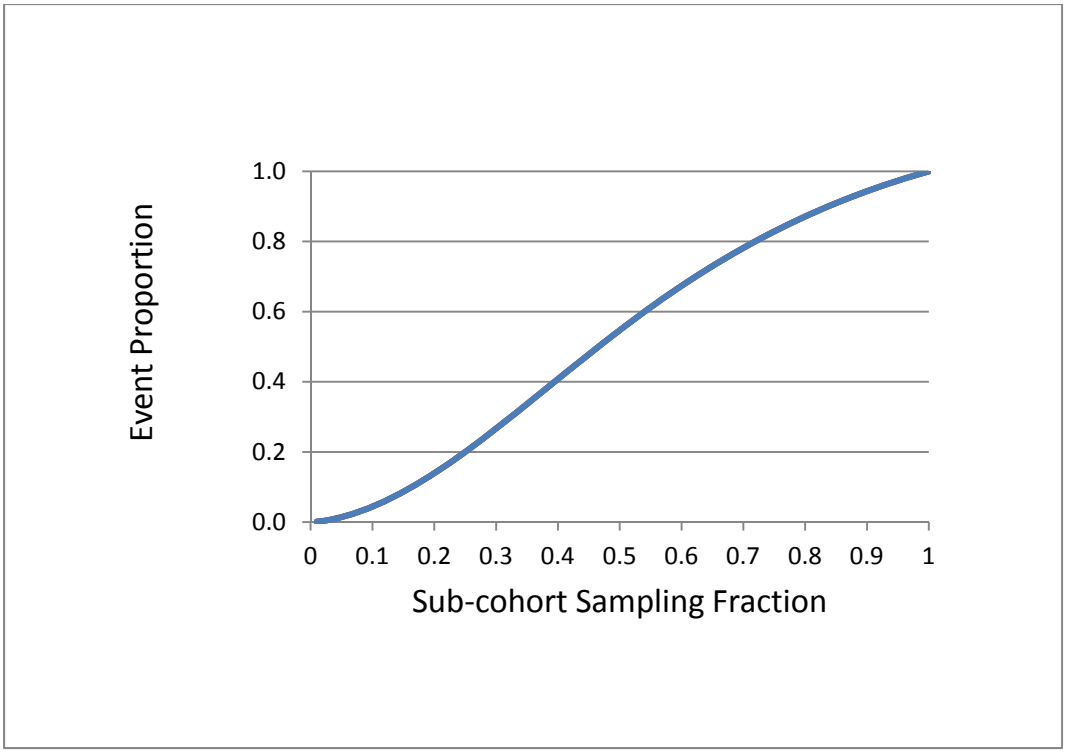


Figure 3 Cost Efficiency of Stratified Case-Cohort Design

Illustration of Sample Size Calculation

In this section, we consider the MORGAM (MONICA, Risk, Genetics, Archiving, and Monograph) study (Kulathinal et al., 2007) as another example to illustrate the stratified sub-cohort sample size calculation. The MORGAM study is a multinational collaborative cohort study prospectively followed for the development of CHD and stroke events in order to investigate the relationship between their genetic risk factor and CHD incidence. A total of 4,559 subjects including 2,282 males and 2,277 females were assessed at the baseline visit in 1997 and followed. Ninety-six CHD events were observed in males and 24 in females by 2003. Since the CHD incidence rate which differs by gender was low and genotyping was expensive, we consider the stratified case-cohort design to examine the relationship between genetic risk factors and CHD incidence. We design the study with 80% power and 0.05 significance level. We assume the genetic risk factor frequency was about 0.2 for both male and female strata. Gender was considered a stratification factor.

Assume that a hazard ratio of 2 is to be detected. This hazard ratio is checked against the minimal detectable hazard ratio. The minimal hazard ratio has a value of 1.9 which is calculated by using the formula in the previous section. The hazard ratio of 2 is appropriate as it is greater than the minimal detectable value.

Table 5 presented the sample size calculation using the proportional, balanced, and optimal sampling methods. Under the optimal design (proportional) design, a total of 153 (209) subjects is required for the sub-cohort, one hundred and twenty-three (105) of which from the male stratum and 30 (104) from the female stratum; the total SCC sample size is 269 (325). , The balanced design requires the similar sample size as the

proportional method. It is because 2,282 subjects in the male stratum and 2,277 in the female stratum which results to the similar strata proportion v_l for the male and female coincidently ($v_1 = v_2 = 0.5$). The proportional (balanced) method requires 20% more sub-cohort size comparing to the optimal design.

It is also observed that under the optimal design, the sub-cohort size at stratum l is proportional to the number of the events at stratum l vs. all events, that is $\tilde{n}_l = \frac{D_l}{D}\tilde{n}$. For instance, in the male stratum, there are 96 events, 80% of 124 the total number of events in the full cohort. The required sub-cohort size at the male stratum is 153, presenting 80% of 153 the overall sub-cohort size.

The non-event vs. event ratio has been examined for all three sampling methods. All methods yield a ratio greater than 1 to ensure the good precision of testing. The optimal method has the smallest overall non-event vs. event ratio of 1.2 among all methods, supporting the conclusion that the optimal method is the most efficient among others.

Table 5. MORGAM Study Sample Size Calculation: SCC Design

Full Cohort and Strata Information

n	*Stratum No.*	*Strata Description*	n_l	*Event*	p_{Dl}	v_l	γ_l
4,559	1	Male	2,282	96	0.042	0.5	0.2
	2	Female	2,277	24	0.011	0.5	0.2
Overall			4,559	120	0.026	1.0	0.2

Sample Size Calculation $\theta = 0.693$

	Stratum No.	n_l	*Sub-cohort*	p_l	*Non-event*	*Event*	*NE:E ratio*	n_{scc}
Proportional	1	2,282	105	0.046	101	96	1.0	197
	2	2,277	104	0.046	104	24	4.4	128
Overall		4,559	209	0.046	205	120	1.7	325
Balanced	1	2,282	105	0.046	101	96	1.0	197
	2	2,277	104	0.046	104	24	4.4	128
Overall		4,559	209	0.046	205	120	1.7	325
Optimal	1	2,282	123	0.054	118	96	1.2	214
	2	2,277	30	0.013	31	24	1.3	55
Overall		4,559	153	0.034	149	120	1.2	269

Note. n = full cohort size, n_l = size of stratum l in full cohort, v_l = proportion of stratum l, p_{Dl} = event proportion in stratum l, γ_l = group 1 proportion, θ = log-hazard ratio, p_l = sub-cohort sampling fraction in stratum l, n_{SCC} = SCC sample size, *Sub-cohort* = sub-cohort size at stratum l, *Non-event* = number of subjects with non-event in stratum l in SCC, *Event* = number of subjects with event in stratum l in SCC, *NE:E = Non-event : Event*. Significant level $\alpha = 0.05$. *Power* = 80%. The sample size is rounded up to the nearest integer as appropriate.

Discussion and Conclusion

We have introduced the stratified log-rank type test statistic for the SCC design and derived the power calculation formula. We considered proportional, balanced, and optimal sampling methods, and derived the corresponding sample size calculation formulas. The optimal method is derived by using Lagrange multiplier method. Our simulation studies show that the proposed stratified log-rank type test statistic is valid for data from SCC design in finite samples. The simulations also show that the power of SCC can be fairly high compared to the power of the full cohort when the event rate is low. The empirical power is similar to the theoretical power.

Simulation studies are also conducted to compare the proportional, balanced, and optimal samplings methods. The results show that when the event rates are relatively homogeneous across strata, the proportional method is superior to the balanced method and is close to the optimal method. However, in the situation that the event rates are heterogeneous over the strata, either the proportional or balanced method can possibly yield higher power than the other. The optimal method yields the highest power with the smallest required sample size among all three methods.

Sample size calculation has been illustrated by using two real studies ARIC and MORGAM projects. The sample sizes from three sampling designs have been generated and compared. It is observed that the sample size from the optimal design is desirable in overall. In addition, the relative efficiency and cost efficiency analyses have been performed and the results suggest that the SCC design is efficient in finite SCC samples.

3 Sample Size/Power Calculation for Generalized Stratified Case-Cohort Design (GSCC)

Introduction

Time to event is a commonly used endpoint for the risk factor assessment in epidemiologic studies or disease prevention trials (Kalbfleisch and Lawless, 1988; ARIC Investigators, 1989; Schouten et al., 1993; Liao et al., 1997; Savitz et al., 2000; Ballantyne, 2004). Case-cohort design (CC) has been often used in studying this endpoint when the disease is rare (Prentice, 1986; Barlow and Prentice, 1988; Self. and Prentice, 1988; Lin and Ying, 1993; Barlow et al., 1994; Borgan et al., 1995; Barlow et al., 1999; Chen and Lo, 1999; Chen, 2001; Chen, 2001a; Chen, 2001b; Cai and Zeng, 2004). A typical case-cohort sample consists of a simple random sample (sub-cohort) from the full cohort and all events in the full cohort. However, when disease incidence is not low, it may not be necessary to include all events in the case-cohort. Cai and Zeng (2007) advocated a generalized case-cohort design (GCC) for this situation. A GCC sample contains a sub-cohort from the full cohort and a random sample from the remaining events without replacement. Cai and Zeng (2007) proposed a general log-rank type of test for the GCC design. The authors also addressed the asymptotic normality property of the test statistics and provided the explicit form for the power calculation. The simulation studies in the paper indicated that the asymptotic approximation performs well in finite samples, and the GCC design was cost-effective and desirable under the non-rare events situation.

In many situations, the study population is not homogenous so that stratification is needed (Boice and Monson, 1977; Hrubec et al., 1989; Borgan et al., 2000; Langholz and Jiao, 2006; Breslow et al., 2009). However, the sample size/power calculation issues for the stratified version of the GCC design have not been addressed before. In this section, we consider a generalized stratified case-cohort design (GSCC) and propose a general log-rank type of test. We discuss the asymptotic normality property of the test statistic and the sample size/power estimation methods, and address the practical proportional and balanced sampling techniques. Simulation studies are conducted to examine the performance of the proposed test and the sample size/power formulas in finite samples. The ARIC study is presented to illustrate the sample size calculation in the proportional and balanced methods under the GSCC design. Recommendations are made in the discussion and conclusion section.

Generalized Stratified Case-Cohort Design

As mentioned in the previous chapter, the following procedures are taken to assemble a GSCC sample: in each stratum, first generate a random sample without replacement from all subjects in the stratum; secondly, generate another random sample without replacement from the remaining events in the stratum; all subjects from two random samples in all strata consist of a GSCC sample.

Stratified Log-rank Test

Notation

Let n denote the number of subjects in the entire cohort. Assume that there are 2 groups and L strata, and there are n_{lj} subjects in group j (j = 1, 2) and stratum l (l = 1,..., L). Let T_{lij} and C_{lij} denote the potential event and censoring time for subject i in group j and stratum l (i = 1,…, n_{lj}) and they are assumed to be independent. Let $X_{lij} = T_{lij} \wedge C_{lij}$ denote the observed time, where $a \wedge b$ denotes the minimum of a and b, and $\Delta_{lij} = I(T_{lij} \leq C_{lij})$ the event indicator variable. Assume that the full cohort size is $n = \sum_{l=1}^{L} n_l$ where $n_l = n_{l1} + n_{l2}$ is the summation of number of subjects in group1 and group 2 in stratum l; and the sub-cohort size is $\tilde{n} = \sum_{l=1}^{L} \tilde{n}_l$, in which $\tilde{n}_l$ is the sub-cohort size in stratum l. Denote γ_l as the proportion of subjects in group 1 and $(1-\gamma_l)$ the proportion of subjects in group 2 in stratum l. Let p_{Dl} be the observed failure rate after censoring in stratum l. Let ξ_{lij} be the indicator that subject i of group j and stratum l is sampled into the sub-cohort and p_l be the sub-cohort sampling probability in stratum l. Let η_{lij} be the indicator that subject i of group j and stratum l is selected into the random sample from the remaining events and q_l be the corresponding sampling probability. The mean number of subjects in sub-cohort in stratum l is $n_l p_l$; and the mean number of sampled additional events in stratum l is $n_l(1-p_l)p_{Dl}q_l$. Let $\hat{n}_l$ be the total number of events not selected into the sub-cohort in stratum l. All subjects in the sub-cohort and random sample from the remaining events make up the generalized stratified case-cohort sample.

Test Statistic

We consider a log-rank type of test to compare the hazard rates between the two groups based on data from a GSCC study. The null hypothesis is H_0: $\Lambda_{l1}(t) = \Lambda_{l2}(t)$, $t \in [0, \Gamma]$, where Γ is the length of study period and $\Lambda_{lj}(t)$ the cumulative hazard function of the event time in group j in stratum l.

Consider the stratified log-rank test statistic T_n:

$$T_n = \sum_{l=1}^{L} \int_0^{\Gamma} \frac{\omega(t)\widetilde{Y}_{l1}(t)\widetilde{Y}_{l2}(t)}{\widetilde{Y}_{l1}(t)+\widetilde{Y}_{l2}(t)} \left\{ \frac{d\widetilde{N}_{l1}(t)}{\widetilde{Y}_{l1}(t)} - \frac{d\widetilde{N}_{l2}(t)}{\widetilde{Y}_{l2}(t)} \right\}, \tag{12}$$

where $\omega(t)$ is a weight function, $\widetilde{Y}_{lj}(t)$ is the approximated risk set of the full cohort in group j by using the sampled subjects in stratum l, and $\widetilde{N}_{lj}(t)$ is the approximated counting process of the full cohort in group j by using the sampled cases in stratum l.

Specifically, $\widetilde{Y}_{lj}(t) = \sum_{i=1}^{n_l} I(X_{lij} \geq t) \left\{ \frac{\Delta_{lij}(1-\xi_{lij})\eta_{lij}}{q_l} + (\Delta_{lij} + \frac{1-\Delta_{lij}}{p_l})\xi_{lij} \right\}$,

$$\widetilde{N}_{lj}(t) = \sum_{i=1}^{n_l} \Delta_{lij} I(X_{lij} \leq t) \left\{ \xi_{lij} + \frac{(1-\xi_{lij})\eta_{lij}}{q_l} \right\}.$$

The formula above indicates that only the subjects selected into the GSCC sample contributed to the calculation; the subjects not selected into the GSCC have a value of zero for ξ_{lij} and η_{lij} thus do not contribute to the summation. Therefore T_n is computable based on the GSCC sample. Essentially inverse sampling probability weighting is used to approximate the at risk and counting process. Plugging in the expression for $\widetilde{N}_{lj}(t)$ into equation (12), we obtain the test statistic T_n in a form

$$\sum_{l=1}^{L}\sum_{i=1}^{n_{l1}}\frac{\Delta_{li1}\omega(X_{li1})\widetilde{Y}_{l2}(X_{li1})\{\xi_{li1}+\frac{(1-\xi_{li1})\eta_{li1}\}}{q_l}}{\widetilde{Y}_{l1}(X_{li1})+\widetilde{Y}_{l2}(X_{li1})}-\sum_{l=1}^{L}\sum_{i=1}^{n_{l2}}\frac{\Delta_{li2}\omega(X_{li2})\widetilde{Y}_{l1}(X_{li2})\{\xi_{li2}+\frac{(1-\xi_{li2})\eta_{li2}\}}{q_l}}{\widetilde{Y}_{l1}(X_{li2})+\widetilde{Y}_{l2}(X_{li2})}. \quad (13)$$

Asymptotic Variance

Following similar arguments as in Cai and Zeng (2007), it can be shown that $n^{-1/2}T_n$ has an asymptotically normal distribution with the asymptotic variance equal to

$$\sigma^2+\sum_{l=1}^{L}v_l\frac{1-p_l}{p_l}Var(\varepsilon_l(X,J)(1-\Delta))+\sum_{l=1}^{L}v_l\frac{P_l(\Delta=1)(1-p_l)(1-q_l)}{q_l}Var(u_l(X,J)\,|\,\Delta=1,\xi=0),$$

where $v_l=n_l/n$, weight of the stratum *l* in the full cohort, and

$$\varepsilon_l(X,J)=-\int_0^X\omega(t)(1-\alpha_l(t))d\Lambda_l(t)I(J=1)+\int_0^X\omega(t)\alpha_l(t)d\Lambda_l(t)I(J=2), \text{ and}$$

$$u_l(X,J)=\varepsilon_l(X,J)+\omega(X)(1-\alpha_l(X))I(J=1)-\omega(X)\alpha_l(X)I(J=2),$$

in which $\alpha_l(t)=P_l(C\geq t,J=1)/P_l(C\geq t)$ at stratum *l,* where *C* denotes the right-censoring time. σ^2 can be estimated by

$$\hat{\sigma}^2=\frac{1}{n}\sum_{l=1}^{L}\left\{\sum_{i=1}^{n_l}\frac{\Delta_{li1}\omega(X_{li1})\widetilde{Y}_{l2}^2(X_{li1})\{\xi_{li1}+\frac{(1-\xi_{li1})\eta_{li1}\}}{q_l}}{(\widetilde{Y}_{l1}(X_{li1})+\widetilde{Y}_{l2}(X_{li1}))^2}\right.$$

Furthermore, let $\hat{u}_l(X,J)=\int_0^X\frac{\omega(t)\{\widetilde{Y}_{l1}(t)I(J=2)-\widetilde{Y}_{l2}(t)I(J=1)\}\{d\widetilde{N}_{l1}(t)+d\widetilde{N}_{l2}(t)\}}{(\widetilde{Y}_{l1}(t)+\widetilde{Y}_{l2}(t))^2}$

$$-\frac{\omega(X)\{\widetilde{Y}_{l1}(X)I(J=2)-\widetilde{Y}_{l2}(X)I(J=1)\}}{\widetilde{Y}_{l1}(X)+\widetilde{Y}_{l2}(X)},$$

and $\hat{\varepsilon}_l(X,J)=\int_0^X\frac{\omega(t)\{\widetilde{Y}_{l1}(t)I(J=2)-\widetilde{Y}_{l2}(t)I(J=1)\}\{d\widetilde{N}_{l1}(t)+d\widetilde{N}_{l2}(t)\}}{(\widetilde{Y}_{l1}(t)+\widetilde{Y}_{l2}(t))^2}$,

$\hat{u}_l(X,J)$ and $\hat{\varepsilon}_l(X,J)$ uniformly converge to $u_l(X,J)$ and $\varepsilon_l(X,J)$. $\hat{u}_l(X,J)$ and $\hat{\varepsilon}_l(X,J)$ can be expressed in the following equivalent summation form:

$$\hat{\varepsilon}_l(X,J)=\sum_{j=1}^{2}\sum_{i=1}^{n_l}\frac{\Delta_{lij}\omega(X_{lij})\widetilde{Y}_{l1}(X_{lij})\{\xi_{lij}+\frac{(1-\xi_{lij})\eta_{lij}}{q_l}\}I(X_{lij}\le X)I(J=2)}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^2}$$

$$-\sum_{j=1}^{2}\sum_{i=1}^{n_l}\frac{\Delta_{lij}\omega(X_{lij})\widetilde{Y}_{l2}(X_{lij})\{\xi_{lij}+\frac{(1-\xi_{lij})\eta_{lij}}{q_l}\}I(X_{lij}\le X)I(J=1)}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^2},$$

$$\hat{u}_l(X,J)=\hat{\varepsilon}_l(X,J)+\frac{\omega(X)\widetilde{Y}_{l2}(X)I(J=1)}{\widetilde{Y}_{l1}(X)+\widetilde{Y}_{l2}(X)}-\frac{\omega(X)\widetilde{Y}_{l1}(X)I(J=2)}{\widetilde{Y}_{l1}(X)+\widetilde{Y}_{l2}(X)}.$$

Hence, the asymptotic variance of $n^{-1/2}T_n$ can be estimated by using the following equation:

$$\hat{\sigma}_T^2=\hat{\sigma}^2+\frac{1}{n}\sum_{l=1}^{l}n_l\frac{1-\hat{p}_l}{\hat{p}_l}\left\{\frac{1}{n_l\hat{p}_l}\sum_{i=1}^{n_l}(\hat{\varepsilon}_l(X_{li},J_{li})^2(1-\Delta_{li})\xi_{li})\right.$$

$$\left.-[\frac{1}{n_l\hat{p}_l}\sum_{i'=1}^{n_l}(\hat{\varepsilon}_l(X_{li'},J_{li'})(1-\Delta_{li'})\xi_{li'})]^2\right\}$$

$$+\frac{1}{n}\sum_{l=1}^{l}\frac{\widehat{n}_l(1-\hat{q}_l)}{\hat{q}_l}\left\{\frac{1}{\widehat{n}_l\hat{q}_l}\sum_{i=1}^{n_l}(\hat{u}_l(X_{li},J_{li})^2\Delta_{li}(1-\xi_{li})\eta_{li})\right.$$

$$\left.-[\frac{1}{\widehat{n}_l\hat{q}_l}\sum_{i'=1}^{n_l}(\hat{u}_l(X_{li'},J_{li'})\Delta_{li'}(1-\xi_{li'})\eta_{li'})]^2\right\}. \qquad (14)$$

Given the test statistic and asymptotic variance in the sections above, we construct the stratified log-rank test for GSCC as below:

To test the equality of the cumulative hazard function of the event time between the two groups in GSCC, i.e. to test the null hypothesis H_0: $\Lambda_{l1}(t) = \Lambda_{l2}(t)$, $t \in [0, \Gamma]$ vs. the

alternative hypothesis H_A : $\Lambda_{l1}(t) \neq \Lambda_{l2}(t)$ (two-sided), or $\Lambda_{l1}(t) = e^{\theta} \Lambda_{l2}(t), \theta$= the log-hazard ratio for two groups, we reject H_0 at the significant level of α if $\left| n^{-1/2} T_n \Big/ \sqrt{\hat{\sigma}^2_{T_n}} \right| > z_{1-\alpha/2}$, where z_α is the $(100\alpha)^{\text{th}}$ percentile of standard normal distribution.

Power Calculation

Under the alternative hypothesis H_A: $\Lambda_{l1}(t) = e^{\theta} \Lambda_{l2}(t)$, where $\theta = O(n^{-1/2})$, $t \in [0, \Gamma]$, we consider $\omega(t) = 1$. Using similar arguments for the GCC in Cai and Zeng (2007), we can show that $n^{-1/2} T_n$ can be approximated by $n^{-1/2} \sum_{l=1}^{L} \int_0^{\Gamma} \alpha_l(t) \bar{Y}_{l2}(t) \{d\Lambda_{l1}(t) - d\Lambda_{l2}(t)\}$

$$= n^{1/2} \theta \sum_{l=1}^{L} \{(1-\gamma_l) \int_0^{\Gamma} \alpha_l(t) P_l(C \geq t \mid J = 2) e^{-\Lambda_l(t)} d\Lambda_l(t)\} + o_p(1).$$

Let $P_l(C \geq t \mid J = 1) = S_{cl}(t)$ indicate the survival function for the censoring random variable in group 1, and $P_l(C \geq t \mid J = 2) = \beta_l P_l(C \geq t \mid J = 1)$ in stratum l, in which β_l means that the subjects in group 2 are β_l -times as likely to be censored as in group 1. The event proportion of group 1 is $\gamma_l / (\gamma_l + \beta_l(1-\gamma_l))$, and for group 2 is $(\beta_l(1-\gamma_l)) / (\gamma_l + \beta_l(1-\gamma_l))$. Based on the arguments for the GCC in Cai and Zeng (2007), we obtain σ_T^2 the asymptotic variance of $n^{-1/2} T_n$

$$\sigma_T^2 = \sum_{l=1}^{L} \left\{ \frac{v_l \beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} \{1 + \frac{(1-p_l)(1-q_l)}{q_l}\} \int_0^{\Gamma} S_{cl}(t) e^{-\Lambda_l(t)} d\Lambda_l(t) \right\}$$

$$+ \sum_{l=1}^{L} \left\{ \frac{2 v_l \beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} \{\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}\} \int_0^{\Gamma} S_{cl}(t) \Lambda_l(t) e^{-\Lambda_l(t)} d\Lambda_l(t) \right\}$$

$$+\sum_{l=1}^{L}\left\{\frac{v_l\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}\{-\frac{1-p_l}{p_l}+\frac{(1-p_l)(1-q_l)}{q_l}\}\int_0^{\Gamma}S_{cl}(t)\Lambda_l(t)^2e^{-\Lambda_l(t)}d\Lambda_l(t)\right\}.$$

We order the failures from the smallest to the largest and assume no two failures are tied to each other, then $\int_0^{\Gamma}dN_l(t)\approx D_l$, the total failures in stratum l, and

$\int_0^{\Gamma}\int_0^{t}\frac{dN_l(s)}{Y_l(s)}dN_l(t)\approx\sum_{k_l=1}^{D_l}\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}$, where n_{k_l} represents the risk set size for the k^{th} failure in stratum l. Similarly, we obtain

$$\int_0^{\Gamma}\{\int_0^{t}\frac{dN_l(s)}{Y_l(s)}\}^2dN_l(t)\approx\sum_{k_l=1}^{D_l}\{\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2,\ Y_l(t)/(n_l(\gamma_l+(1-\gamma_l)\beta_l))\to S_{cl}(t)e^{-\Lambda_l(t)},$$

$\int_0^{t}dN_l(s)/Y_l(s)\to\Lambda_l(t)$, $Y_l(t)=Y_{l1}(t)+Y_{l2}(t)$, $N_l(t)=N_{l1}(t)+N_{l2}(t)$, and $v_l=n_l/n$. The above equation can be written as

$$\sigma_T^2\approx\frac{1}{n}\sum_{l=1}^{L}\left\{\frac{\beta_l\gamma_l(1-\gamma_l)}{(\gamma_l+\beta_l(1-\gamma_l))^2}[\frac{D_l}{p_l}-\{\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}\}\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2]\right\}. \qquad (15)$$

The derivation is shown in Appendix B.

We obtain $n^{-1/2}T_n\ \approx n^{1/2}\theta\sum_{l=1}^{L}\frac{v_l\beta_l\gamma_l(1-\gamma_l)p_{Dl}}{\gamma_l+\beta_l(1-\gamma_l)}$, where $p_{Dl}=D_l/n_l$. Under the alternative hypothesis, the power is calculated from the proposed test statistic $n^{-1/2}T_n\Big/\sqrt{\hat{\sigma}_{T_n}^2}$. The power function is given as

$$\text{Power}=\Phi\left(z_{\alpha/2}+\frac{n\theta}{\sqrt{\chi}}\sum_{l=1}^{L}\frac{v_l\beta_l\gamma_l(1-\gamma_l)p_{Dl}}{\gamma_l+\beta_l(1-\gamma_l)}\right), \qquad (16)$$

where $\chi=\sum_{l=1}^{L}\left\{\frac{n_l\beta_l\gamma_l(1-\gamma_l)}{(\gamma_l+\beta_l(1-\gamma_l))^2}[\frac{p_{Dl}}{p_l}-\{\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}\}\frac{1}{n_l}\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2]\right\}$.

$\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2$ can be approximated by $n_l\left[p_{Dl}-(1-p_{Dl})\{\log(1-p_{Dl})\}^2\right]$. The derivation is shown in Appendix C.

Sample Size Calculation of Proportional and Balanced Designs

We consider two commonly used allocation strategies across strata. One is to select the number of subjects in each stratum proportional to the stratum size in the population (proportional design) and the other is to select the same number of subjects in each stratum (balanced design). We consider the power issues for the proportional and balanced designs. The expected number of subjects with non-event (*NE*) and the expected number with event (*E*) in stratum *l* are stratum $NE_l = n_l p_l(1-p_{Dl})$ and $E_l = n_l p_{Dl}(p_l+(1-p_l)q_l)$, respectively. The expected total number of subjects in stratum *l* in GSCC sample $\tilde{n}_{Gl} = NE_l + E_l = (1+M_l^{-1})n_l p_l(1-p_{Dl})$, where $M_l = NE_l : E_l$.

The total number of subjects in GSCC sample $n_{GSCC} = \sum_{l=1}^{L}\tilde{n}_{Gl} = \sum_{l=1}^{L}\{(1+M_l^{-1})n_l p_l(1-p_{Dl})\}$.

Let $A_l = \frac{n_l\beta_l\gamma_l(1-\gamma_l)}{(\gamma_l+\beta_l(1-\gamma_l))^2}$, and $\frac{1}{n_l}\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2 = p_{Dl}-B_l$,

where $B_l = (1-p_{Dl})\{\log(1-p_{Dl})\}^2$. The power function (16) is transformed to the following algebraic equation

$$\sum_{l=1}^{L}\left\{A_l p_{Dl}(p_l+\frac{1-p_l}{q_l})\right\}+\sum_{l=1}^{L}\left\{A_l B_l(\frac{1}{p_l}+\frac{1-p_l}{q_l}-p_l)\right\}=\left(\frac{n\theta\Delta}{(z_{1-\alpha/2}+z_\beta)}\right)^2, \qquad (17)$$

where $\Delta = \sum_{l=1}^{L}\frac{v_l\beta_l\gamma_l(1-\gamma_l)p_{Dl}}{\gamma_l+\beta_l(1-\gamma_l)}$.

In the next sub-sections, we will derive the formulas for the sample size calculation in the proportional and balanced designs, respectively.

Proportional Design

Under the proportional design, $p_l = p$, and $q_l = q$. Denote D the total number of cases in full cohort, and p_D the overall event proportion across strata, then $NE = p(n-D)$, $E = Dp + D(1-p)q$, and $n_{GSCC} = np + qD - pqD$. Let $NE : E = M$, the ratio of non-events and events in the case-cohort sample and assume it is pre-specified. Then $n_{Gscc} = (1 + M^{-1})p(n-D)$ and $q = \frac{p}{1-p}(\frac{n-D}{MD} - 1)$. Let $B_G^2 = \left(\frac{n\theta\Delta}{(z_{1-\alpha/2} + z_\beta)} \right)^2$. The Equation (17) becomes

$$\left\{ p + \frac{(1-p)^2}{p}(\frac{n-D}{MD} - 1)^{-1} \right\} \sum_{l=1}^{L} (A_l p_{Dl}) + \left\{ \frac{1}{p} + \frac{(1-p)^2}{p}(\frac{n-D}{MD} - 1)^{-1} - p \right\} \sum_{l=1}^{L} (A_l B_l) = B_G^2.$$

Let $F_1 = \sum_{l=1}^{L} (A_l p_{Dl}), F_2 = (\frac{n-D}{MD} - 1)^{-1}$, and $F_3 = \sum_{l=1}^{L} (A_l B_l)$, we obtain 2 solutions of p $p^{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, in which $a = F_1(1 + F_2) + (F_2 - 1), b = -2F_1F_2 - 2F_2F_3 - B_G^2$, and $c = F_1F_2 + (1 + F_2)F_3$. The meaningful p shall be $0 \le p \le 1$. If both $p^{1,2}$ are between 0 to 1, the smaller one shall be chosen. We further obtain q and n_{GSCC} from the expression above. The total number of subjects in each stratum is $\tilde{n}_{Gl} = n_l(p + (1-p)p_{Dl}q)$, including the sub-cohort size $n_l p$ and the size from the remaining events $n_l(1-p)p_{Dl}q$ in each stratum.

Balanced Design

Under the balanced design, the GSCC sample size is equally distributed to each stratum in L strata. The sample size at stratum l $\tilde{n}_{Gl}$ has $\tilde{n}_{Gl} = \frac{n_{GSCC}}{L}$. Let $NE_l : E_l = M_l$, we obtain $\tilde{n}_{Gl} = (1 + M_l^{-1}) n_l p_l (1 - p_{Dl}) = \frac{n_{GSCC}}{L}$, in which M_l can be pre-specified. After transformation, we have $p_l = \frac{n_{GSCC}}{L n_l (1 + M_l^{-1})(1 - p_{Dl})}$, and $q_l = \frac{p_l}{1 - p_l}\left[\frac{(1 - p_{Dl})}{M_l p_{Dl}} - 1\right]$. An appropriate M_l shall be $M_l \le \frac{(1 - p_{Dl})}{p_{Dl}}$ to ensure $q_l \ge 0$. The meaningful q_l shall be the minimum of $\frac{p_l}{1 - p_l}\left[\frac{(1 - p_{Dl})}{M_l p_{Dl}} - 1\right]$ and 1. Furthermore, let $F_{1l} = A_l p_{Dl}$, $F_{2l} = L n_l (1 + M_l^{-1})(1 - p_{Dl})$, $F_{3l} = \frac{(1 - p_{Dl})}{M_l p_{Dl}} - 1$, $F_{4l} = A_l B_l$, from these expressions and the Equation (17), we obtain 2 solutions $n_{GSCC}^{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$, in which

$$a = \sum_{l=1}^{L} \frac{F_{1l}}{F_{2l}} - \sum_{l=1}^{L} \frac{F_{4l}}{F_{2l}} + \sum_{l=1}^{L} \frac{F_{1l}}{F_{2l} F_{3l}} + \sum_{l=1}^{L} \frac{F_{4l}}{F_{2l} F_{3l}},\quad b = -2\sum_{l=1}^{L} \frac{F_{1l}}{F_{3l}} - 2\sum_{l=1}^{L} \frac{F_{4l}}{F_{3l}} - B_G^2, \text{ and}$$

$$c = \sum_{l=1}^{L} \frac{F_{1l} F_{2l}}{F_{3l}} + \sum_{l=1}^{L} F_{2l} F_{4l} + \sum_{l=1}^{L} F_{4l}.$$ The meaningful n_{GSCC} shall be greater than 0. If both are greater than 0, the smaller one shall be chosen. We further obtain p_l and q_l using the expressions above.

The GSCC optimal design is not addressed because there are 2 sets of variables (p_l and q_l) under GSCC design while there is only 1 set of variables (p_l) in SCC design. Because of the two sets of the freely varying variables, there are too many assumptions

that need to be made in order to derive an optimal design. Because of the many imposed constrains, such design will be too restrictive to be of general use. Therefore we do not pursue an optimal allocation plan under the GSCC design.

Numeric Results

We conduct the simulation studies to evaluate the empirical type 1 error and empirical power for the finite GSCC samples. We also calculate the theoretical power to compare with the simulated empirical power. Both the homogeneous and heterogeneous GSCC samples are generated with different overall event proportion rate (0.2 and 0.4), hazard ratio (1, 1.5, and 2), drop-out rate (0 and 0.2), and sub-cohort sampling proportion (0.1 and 0.2). We present the proportional and balanced designs in each sample. In each GSCC sample for power analysis, the total sample size n_{GSCC}, the size of sub-cohort, the size of the remaining event sample, the total number of non-event vs. event and its ratio are provided. Furthermore, the ARIC study is presented as an example to illustrate the sample size calculation under the GSCC design with both the proportional and balanced sampling methods.

Simulation Studies

Simulation studies are conducted to evaluate the empirical type 1 error and empirical power for the stratified log-rank test in finite GSCC samples (T_{GSCC}). The empirical power in full cohort (T_{Full}) is provided for reference. The formula (13) is used to calculate the GSCC test statistic and the formula (14) for the GSCC asymptotic variance. T_{Full} is calculated using the SAS procedure PROC LIFETEST for stratified log-rank test.

Furthermore, we present the theoretical power of GSCC (P_{GSCC}) and of full cohort (P_{Full}) for each set up. P_{GSCC} is calculated based on the simulated full cohort using the power formula (16). P_{Full} is calculated using the formula $\Phi(z_{\alpha/2} + n^{1/2}|\theta|\sqrt{\sum_{l=1}^{L}(\gamma_l(1-\gamma_l)p_{Dl}v_l)}$. Various values are considered for the full cohort size n , stratum proportion v_l , event proportion p_{Dl} , group 1 proportion γ_l , sub-cohort sampling fraction p_l , pre-mature dropout rate p_{lc} , and event sampling fraction q_l in stratum *l.* The follow procedures/parameters are set up for the simulation:

- The significant level α is set at 0.05 and the number of strata L = 4. The full cohort size n = 1,000, with stratum proportions of 0.1, 0.2, 0.3, and 0.4. n = 10,000 is used for the empirical type 1 error samples for higher precision.
- Samples with both homogeneous and heterogeneous event rates are generated. In the homogeneous GSCC samples, the event rates across strata are similar. For the set up where the overall event proportion, p_D , is 21% (41%), the event proportions (p_{Dl}) are 0.15, 0.2, 0.25, and 0.2 (0.35, 0.4, 0.45, and 0.4) in stratum 1-4, respectively. In the heterogeneous GSCC samples, the event rates across strata are different. For the set up where the overall event proportion p_D is 21% (41%), the event proportions (p_{Dl}) are 0.15, 0.2, 0.35, and 0.125 (0.2, 0.4, 0.45, and 0.44) in stratum 1-4, respectively.
- Both proportional and balanced sampling methods are examined. Under the proportional method, the sub-cohort sampling proportion p (0.1 or 0.2) is the same over 4 strata. The event sampling fraction q is set to 0.3 (0.1) for the GSCC

samples with p_D = 21% (41%) to ensure the non-event vs. event ratio *M* is around 1 or larger. Under the balanced method, the same total sample size as in the proportional method is equally distributed to each stratum; p_l and q_l are set appropriately so that the overall *M* is comparable to the proportional design. The combination of p_l and q_l is indicated as Set1-4 for the balanced design, the values of which are presented in Table 6.

- All subjects are assigned to one of the two groups. The group 1 proportion γ_l (0.3) is the same over 4 strata. Assume the two groups have the same censoring time (β_l = 1). The event time is generated from the exponential distribution with the values of e^{θ} (or hazard ratio (*HR*) = 1, 1.5, or 2.0).
- The censoring time is from a mixture distribution with probability p_{lc} being generated from uniform distribution in $[0, \Gamma]$ and probability $(1 - p_{lc}) = \Gamma$. $p_{lc} = 0$ indicates no pre-mature dropout occurs and $p_{lc} = 0.2$ means the 20% pre-mature dropout in study.
- Each simulation is repeated 1,000 times.

Table 6 displays the sample size and *p* and *q* combination associated with the set up above. The numbers of sample size are rounded up to the integer as appropriate. The empirical type 1 error and empirical and theoretical power of the proportional and balanced designs in the homogeneous and heterogeneous GSCC samples are reported in Table 7 and Table 8, respectively.

Table 7 shows the empirical type I error for the stratified log-rank test using the GSCC (T_{GSCC}) and full cohort (T_{Full}) data. It is observed that both T_{GSCC} and T_{Full} are close

to 0.05 for both the proportional and balanced designs. For instance, in Sample No. 1, where the full cohort size n = 10,000, the event proportion p_D = 0.21, the sub-cohort sampling proportion p = 0.2, the event sampling proportion q = 0.3, the pre-mature dropout rate p_{lc} = 0, and the hazard ratio θ = 1, we obtain the empirical type 1 error for GSCC T_{GSCC} = 0.054 (0.051) for the proportional (balanced) design and for full cohort T_{Full} = 0.046. The theoretical power for GSCC P_{GSCC} and full cohort P_{Full} = 0.050. The results from these finite samples indicate that the stratified log rank test for GSCC works well in finite samples.

Table 8 shows the empirical power and theoretical power under the alternative hypothesis HR (θ) = 1.5 and 2. For instance, in Sample No. 23 , where the full cohort size n = 1,000, the event proportion p_D = 0.21, the sub-cohort sampling proportion p = 0.2, the event sampling proportion q = 0.3, the pre-mature dropout rate p_{lc} = 0, and the hazard ratio θ = 2, we obtain empirical power T_{GSCC} = 0.611 (0.544) and theoretical power P_{GSCC} = 0.681 (0.624) for the proportional (balanced) design for the GSCC sample and T_{Full} = 0.983 and P_{Full} = 0.996 for the full cohort. In overall, the simulated empirical power is in align with the theoretical power within either the proportional or the balanced design.

Proportional and balanced design comparison

The proportional and balanced designs are compared for the homogeneous and heterogeneous GSCC samples, respectively. The results from Table 8 suggest that the proportional method yields slightly higher power than the balanced method in majority of GSCC samples. The balanced design has the disadvantage in some GSCC samples. Some

strata may not achieve the required event size so that the desired non-event vs. event ratio is not satisfied. One may combine the strata as appropriate to ensure the adequate events in the combined strata.

Table 6: GSCC Samples Set up of Proportional and Balanced Designs

n	p_D	*p*	*q*	n_{GSCC}	*Sub-cohort*	*Event Sample*	*Non-event*	*Event*	*M (NE/E)*
Proportional Design with Homogeneous and Heterogeneous Event Rates									
1,000	0.21	0.1	0.3	157	100	57	79	78	1.0
	0.21	0.2	0.3	250	200	50	158	92	1.7
	0.41	0.1	0.1	137	100	37	59	78	0.8
	0.41	0.2	0.1	233	200	33	118	115	1.0
Balanced Design with Homogeneous Event Rates									
1,000	0.21	Set1		157	104	53	83	74	1.1
	0.21	Set2		250	176	74	142	109	1.3
	0.41	Set3		137	114	22	68	68	1.0
	0.41	Set4		233	195	38	116	116	1.0
Balanced Design with Heterogeneous Event Rates									
1,000	0.21	Set1		157	105	51	83	74	1.1
	0.21	Set2		250	179	71	142	109	1.3
	0.41	Set3		137	111	25	68	68	1.0
	0.41	Set4		233	201	32	126	107	1.2

(Continued)

Table 6: (continued)

p_l and q_l Combination in Balanced Design					
Homogeneous	p_l	q_l	*Heterogeneous*	p_l	q_l
Set1	0.28	1.00	Set1	0.28	1.00
	0.12	0.42		0.12	0.42
	0.09	0.19		0.10	0.10
	0.06	0.20		0.06	0.36
Overall	0.10	0.28	Overall	0.11	0.27
Set2	0.56	1.00	Set2	0.56	1.00
	0.20	0.73		0.20	0.73
	0.14	0.32		0.16	0.16
	0.10	0.33		0.09	0.59
Overall	0.18	0.56	Overall	0.18	0.41
Set3	0.26	0.31	Set3	0.21	0.82
	0.14	0.08		0.14	0.08
	0.10	0.03		0.10	0.03
	0.07	0.04		0.08	0.02
Overall	0.11	0.06	Overall	0.11	0.06
Set4	0.45	0.69	Set4	0.48	1.00
	0.24	0.16		0.24	0.16
	0.18	0.05		0.18	0.05
	0.12	0.07		0.13	0.04
Overall	0.19	0.11	Overall	0.20	0.08

Note. n = full cohort size, p_D = average event proportion, p_l = sub-cohort sampling fraction in stratum l, p = overall sub-cohort sampling fraction, q_l = sampling proportion from remaining events in stratum l, q = overall sampling proportion from remaining events, n_{GSCC} = GSCC sample size, *Sub-cohort* = sub-cohort size, *Event sample* = number of subjects sampled from remaining events, *Non-event* = number of subjects with non-event in GSCC, *Event* = number of subjects with event in GSCC, *M (NE/E)* = ratio of *Non-event vs. Event*.

Table 7: Empirical Type I Error of Proportional and Balanced Designs in GSCC Samples

			Proportional Design				Balanced Design				
Sample No.	p_D	p_{lc}	p	q	T_{GSCC}	P_{GSCC}	p, q	T_{GSCC}	P_{GSCC}	T_{full}	P_{full}
With Homogeneous Events											
1	0.21	0	0.1	0.3	0.054	0.050	Set1	0.051	0.050	0.046	0.050
2	0.21	0.2	0.1	0.3	0.051	0.050	Set1	0.043	0.050	0.055	0.050
3	0.21	0	0.2	0.3	0.055	0.050	Set2	0.050	0.050	0.046	0.050
4	0.21	0.2	0.2	0.3	0.059	0.050	Set2	0.068	0.050	0.055	0.050
5	0.41	0	0.1	0.1	0.054	0.050	Set3	0.058	0.050	0.051	0.050
6	0.41	0.2	0.1	0.1	0.062	0.050	Set3	0.072	0.050	0.050	0.050
7	0.41	0	0.2	0.1	0.061	0.050	Set4	0.048	0.050	0.051	0.050
8	0.41	0.2	0.2	0.1	0.061	0.050	Set4	0.040	0.050	0.050	0.050
With Heterogeneous Events											
9	0.21	0	0.1	0.3	0.049	0.050	Set1	0.065	0.050	0.047	0.050
10	0.21	0.2	0.1	0.3	0.055	0.050	Set1	0.051	0.050	0.064	0.050
11	0.21	0	0.2	0.3	0.057	0.050	Set2	0.050	0.050	0.047	0.050
12	0.21	0.2	0.2	0.3	0.052	0.050	Set2	0.067	0.050	0.064	0.050
13	0.41	0	0.1	0.1	0.057	0.050	Set3	0.055	0.050	0.048	0.050
14	0.41	0.2	0.1	0.1	0.054	0.050	Set3	0.075	0.050	0.050	0.050
15	0.41	0	0.2	0.1	0.055	0.050	Set4	0.055	0.050	0.048	0.050
16	0.41	0.2	0.2	0.1	0.052	0.050	Set4	0.065	0.050	0.050	0.050

Note. n = full cohort size, p_D = average event proportion, p_{lc} = pre-mature dropout rate in study, p = overall sub-cohort sampling fraction, q = overall sampling proportion from remaining events, T_{GSCC} = empirical Type I error of GSCC, T_{Full} = empirical Type I error of full cohort, T_{Sub} = empirical Type I error of sub-cohort, P_{GSCC} = theoretical power of GSCC, P_{Full} = theoretical power of full cohort. Significant level $\alpha = 0.05$.

Table 8: Empirical and Theoretical Power of Proportional and Balanced Designs in GSCC Samples

Sample				Proportional Design				Balanced Design				
No.	p_D	*HR*	p_{lc}	p	q	T_{GSCC}	P_{GSCC}	p, q	T_{GSCC}	P_{GSCC}	T_{full}	P_{full}
With Homogeneous Events												
1	0.21	1.5	0	0.1	0.3	0.231	0.239	Set1	0.201	0.200	0.720	0.768
2	0.21	1.5	0.2	0.1	0.3	0.193	0.211	Set1	0.160	0.177	0.618	0.673
3	0.21	1.5	0	0.2	0.3	0.281	0.311	Set2	0.294	0.299	0.720	0.768
4	0.21	1.5	0.2	0.2	0.3	0.233	0.264	Set2	0.242	0.261	0.626	0.673
5	0.21	2	0	0.1	0.3	0.518	0.571	Set1	0.431	0.485	0.987	0.996
6	0.21	2	0.2	0.1	0.3	0.410	0.507	Set1	0.366	0.425	0.959	0.985
7	0.21	2	0	0.2	0.3	0.605	0.707	Set2	0.643	0.689	0.987	0.996
8	0.21	2	0.2	0.2	0.3	0.522	0.622	Set2	0.563	0.619	0.959	0.985
9	0.41	1.5	0	0.1	0.1	0.253	0.228	Set3	0.188	0.147	0.958	0.964
10	0.41	1.5	0.2	0.1	0.1	0.211	0.192	Set3	0.199	0.123	0.907	0.920
11	0.41	1.5	0	0.2	0.1	0.285	0.294	Set4	0.268	0.232	0.958	0.964
12	0.41	1.5	0.2	0.2	0.1	0.261	0.238	Set4	0.212	0.187	0.907	0.920
13	0.41	2	0	0.1	0.1	0.531	0.546	Set3	0.342	0.343	1.000	1.000
14	0.41	2	0.2	0.1	0.1	0.435	0.462	Set3	0.300	0.278	1.000	1.000
15	0.41	2	0	0.2	0.1	0.618	0.679	Set4	0.502	0.554	1.000	1.000
16	0.41	2	0.2	0.2	0.1	0.494	0.568	Set4	0.418	0.449	1.000	1.000

(continued)

Table 8: (continued)

Sample				Proportional design				Balanced design				
No.	p_D	*HR*	p_{lc}	p	q	T_{GSCC}	P_{GSCC}	p, q	T_{GSCC}	P_{GSCC}	T_{full}	P_{full}
With Heterogeneous Events												
17	0.21	1.5	0	0.1	0.3	0.223	0.225	Set1	0.195	0.180	0.716	0.771
18	0.21	1.5	0.2	0.1	0.3	0.197	0.200	Set1	0.158	0.155	0.630	0.676
19	0.21	1.5	0	0.2	0.3	0.264	0.298	Set2	0.222	0.264	0.716	0.771
20	0.21	1.5	0.2	0.2	0.3	0.237	0.254	Set2	0.192	0.225	0.630	0.676
21	0.21	2	0	0.1	0.3	0.517	0.541	Set1	0.379	0.431	0.983	0.996
22	0.21	2	0.2	0.1	0.3	0.425	0.481	Set1	0.323	0.367	0.960	0.985
23	0.21	2	0	0.2	0.3	0.611	0.681	Set2	0.544	0.624	0.983	0.996
24	0.21	2	0.2	0.2	0.3	0.518	0.603	Set2	0.451	0.540	0.960	0.985
25	0.41	1.5	0	0.1	0.1	0.244	0.227	Set3	0.207	0.117	0.953	0.964
26	0.41	1.5	0.2	0.1	0.1	0.235	0.181	Set3	0.229	0.101	0.893	0.920
27	0.41	1.5	0	0.2	0.1	0.314	0.293	Set4	0.211	0.190	0.953	0.964
28	0.41	1.5	0.2	0.2	0.1	0.257	0.237	Set4	0.198	0.154	0.893	0.920
29	0.41	2	0	0.1	0.1	0.515	0.543	Set3	0.323	0.259	1.000	1.000
30	0.41	2	0.2	0.1	0.1	0.439	0.460	Set3	0.295	0.214	0.999	1.000
31	0.41	2	0	0.2	0.1	0.628	0.678	Set4	0.400	0.457	1.000	1.000
32	0.41	2	0.2	0.2	0.1	0.506	0.566	Set4	0.344	0.361	0.999	1.000

Note. n = full cohort size, p_D = average event proportion, p_{lc} = pre-mature dropout rate in study, HR = hazard ratio, p = overall sub-cohort sampling fraction, q = overall sampling proportion from remaining events, T_{GSCC} = empirical testing power of GSCC, T_{Full} = empirical testing power of full cohort, T_{Sub} = empirical testing power of sub-cohort, P_{GSCC} = theoretical power of GSCC, P_{Full} = theoretical power of full cohort. Significant level $\alpha = 0.05$.

Illustration of GSCC Sample Size Calculation and Comparison with SCC

In this section, we provide an example for the Atherosclerosis Risk in Communities Study (ARIC) to illustrate the sample size calculation under the GSCC design using the proportional and balanced methods, respectively. In addition, we provide the SCC sample size calculation under the optimal, proportional, and balanced designs for comparison.

The ARIC study, sponsored by the National Heart, Lung and Blood Institute (NHLBI), is a prospective epidemiologic study conducted in four U.S. communities to investigate the etiology and natural history of atherosclerosis and the etiology of clinical atherosclerotic diseases (ARIC Investors, 1989). A total of 15,972 participants completed a home interview and clinic examinations and were prospectively followed for the development of coronary heart disease (CHD) and other vascular events. It was of interest to examine the relationship between platelet PI^{A2} allele, a potential genetic risk factor, and CHD incidence. After some exclusion criteria, the total number of participants is 14,239 (full cohort n). The combination of gender (male and female), age group (<= 54-yrs and >= 55-yrs), and carotid artery intima-media thickness (IMT) (thin IMT or not thin IMT) were considered as the stratification factor and all subjects in the full cohort were stratified into 8 strata with the stratum proportion v_l of 0.19, 0.06, 0.17, 0.15, 0.19, 0.02, 0.17, and 0.05, respectively. The strata information is presented in Table 9. The numbers of subjects in each stratum n_l were 2,703, 830, 2,487, 2,066, 2,690, 295, 2,386, and 782, respectively. The CHD event rates p_{Dl} over the strata were 0.037, 0.068, 0.114, 0.073, 0.051, 0.029, 0.142, and 0.083. Based on the results of Batalla et al. (2004), we use the prevalence of PI^{A2} allele carriers γ_l to be 25% across strata.

Based on the sample size formulas in the previous sections, we calculate the required sample sizes under the proportional and balanced designs. We set the power at 80%, the significant level at 0.05, and the minimum detected hazard ratio at 1.6 ($\theta = 0.47$). The required sample sizes are presented in Table 9. The numbers are rounded up to the integers.

Under the proportional design a total of $n_{GSCC} = 497$ subjects comprising 71, 28, 103, 69, 78, 8, 112, and 28 subjects in 8 strata, respectively. The overall sampling proportion of 0.017 is required in order to generate 270 subjects from the sub-cohort of which 51, 16, 47, 39, 51, 6, 45, and 15 are required from each of strata 1-8. The event sampling proportion of 0.2 will provide a total of 227 subjects to be sampled from the remaining events, of which 20, 12, 56, 30, 27, 2, 67, and 13 subjects will be from strata 1-8, respectively. The overall ratio of non-event vs. event = 1 (249:248).

Under the balanced design 96 subjects including 48 events would be required in each stratum. However, there are only 9 subjects in Stratum 6. We combined Stratum 6 with Stratum 5. The overall sampling proportion of 0.026 is required in order to generate 365 subjects from the sub-cohort of which 50, 51, 54, 52, 50, 56, and 52 are required from each of strata 1-7. The overall event sampling proportion of 0.277 will provide a total of 307 subjects to be sampled from the remaining events, of which 46, 45, 42, 44, 46, 40, and 44 subjects will be from strata 1-7, respectively. The overall ratio of non-event vs. event = 1.0 (336:336). As a result, the proportional design requires approximate 1/4 less subjects than the balanced design so is more cost-effective.

In order to compare the GSCC with SCC design, we perform the sample size calculation for the ARIC study using the SCC design under the same $\alpha, \beta, \theta,$ and other specifications. The optimal, proportional, and balanced sampling methods are used and the sample sizes are displayed in Table 9. In the proportional design, the sub-cohort sampling proportion from GSCC is comparable with that from SCC (0.019 vs. 0.02), however, the event sampling proportion is quite different between GSCC and SCC (0.2 vs. 1, considering 1 is the special case of GSCC by including all events outside of sub-cohort). It indicates that the GSCC design saves 80% of events outside of sub-cohort than the SCC design. In the balanced design, the similar sub-cohort sampling proportion is observed in GSCC and SCC (0.029 vs. 0.026) while approximately 70% of events outside of sub-cohort are saved in GSCC compared with SCC. The sample size required for the optimal design is larger than either of the GSCC design. As a result, the GSCC design saves the sample size comparing with SCC design in the ARIC study.

Table 9. ARIC Study Sample Size Calculation: GSCC and SCC Designs

Full Cohort and Strata Information

n	Stratum l	*Strata Description*	n_l	p_{Dl}	v_l	γ_l
14,239	1	female, age >= 55-yrs, not thin IMT	2,703	0.037	0.19	0.25
	2	female, age >= 55-yrs, thin IMT	830	0.068	0.06	0.25
	3	female, age =< 54-yrs, not thin IMT	2,487	0.114	0.17	0.25
	4	female, age =< 54-yrs, thin IMT	2,066	0.073	0.15	0.25
	5	male, age >= 55-yrs, not thin IMT	2,690	0.051	0.19	0.25
	6	male, age >= 55-yrs, thin IMT	295	0.029	0.02	0.25
	7	male, age =< 54-yrs, not thin IMT	2,386	0.142	0.17	0.25
	8	male, age =< 54-yrs, thin IMT	782	0.083	0.05	0.25

Sample Size Calculation: Proportional Design

θ	l	p_l	q_l	n_{GSCC}	*Sub-cohort*	*Event sample*	*Non-event*	*Event*	*Ratio (NE/E)*
0.47	1	0.019	0.2	71	51	20	49	22	2.2
	2	0.019	0.2	28	16	12	15	13	1.2
	3	0.019	0.2	103	47	56	42	61	0.7
	4	0.019	0.2	69	39	30	36	33	1.1
	5	0.019	0.2	78	51	27	48	30	1.6
	6	0.019	0.2	8	6	2	6	2	3.0
	7	0.019	0.2	112	45	67	39	73	0.5
	8	0.019	0.2	28	15	13	14	14	1.0
Overall		0.019	0.2	497	270	227	249	248	1.0

Sample Size Calculation: Balanced Design

θ	l	p_l	q_l	n_{GSCC}	*Sub-cohort*	*Event sample*	*Non-event*	*Event*	*Ratio (NE/E)*
0.47	1	0.018	0.465	96	50	46	48	48	1.0
	2	0.061	0.836	96	51	45	48	48	1.0
	3	0.022	0.149	96	54	42	48	48	1.0
	4	0.025	0.300	96	52	44	48	48	1.0
	5,6	0.017	0.316	96	50	46	48	48	1.0
	7	0.023	0.120	96	56	40	48	48	1.0
	8	0.066	0.710	96	52	44	48	48	1.0
Overall		0.026	0.313	672	365	307	336	336	1.0

(continued)

Table 9. (continued)

Sample Size Calculation: SCC Design

			Optimal		*Proportional*		*Balanced*	
θ	l	n_l	*Sub-cohort*	p_l	*Sub-cohort*	p_l	*Sub-cohort*	p_l
0.47	1	2,703	20	0.007	55	0.02	47	0.017
	2	830	12	0.014	17	0.02	47	0.057
	3	2,487	57	0.023	51	0.02	47	0.019
	4	2,066	32	0.015	42	0.02	47	0.023
	5	2,690	28	0.010	55	0.02	47	0.017
	6	295	2	0.007	6	0.02	47	0.159
	7	2,386	72	0.030	49	0.02	47	0.020
	8	782	12	0.015	16	0.02	47	0.060
Overall		14,239	235	0.017	291	0.02	376	0.026
n_{SCC}			1,351		1,407		1,484	

Note. n = full cohort size, n_l = size of stratum l in full cohort, v_l = proportion of stratum l, p_{Dl} = event proportion in stratum l, γ_l = group 1 proportion, θ = log-hazard ratio, p_l = sub-cohort sampling fraction in stratum l, n_{SCC} = SCC sample size, *Sub-cohort* = sub-cohort size at stratum l, *Non-event* = number of subjects with non-event in stratum l, *Event* = number of subjects with event in stratum l, *NE:E = Non-event : Event*, *Event sample* = number of subjects sampled from remaining events in stratum l, q_l = sampling proportion from remaining events in stratum l, n_{GSCC} = GSCC sample size in stratum l, Significant level $\alpha = 0.05$. *Power* = 80%. The sample size is rounded up to the nearest integer as appropriate.

Discussion and Conclusion

We considered the GSCC design for the situation where the event is not-rare. We proposed the general stratified log-rank type test for the GSCC design and derived the power function. Our simulation studies show that the proposed stratified log-rank type test is valid in finite GSCC samples. The proportional design has a slightly higher power than the balanced design in both the homogeneous and heterogeneous situation. However, the balanced design has the disadvantage that some strata may not have adequate events for sampling so that it requires to be combined with other strata. The ARIC study is used to illustrate the sample size calculation from the proportional and balanced designs. It is observed that the proportional design saves about 1/4 sample size compared with the balanced design.

The GSCC design is compared with the SCC design in the sample size calculation for the ARIC study. The GSCC design requires smaller sample size by including only a portion of subjects with events in the ARIC study instead of including all the events as in the SCC design.

4 Future Research Plans

Introduction

In previous chapters, we used the log-rank test to detect the difference of hazard function between two groups in the stratified case-cohort data. In this chapter, we are interested in other tests which may be more sensitive against the alternatives under different configurations. To simplify the problem, we only consider the case-cohort data without stratification. Assume the survival in two groups is stochastically ordered, the alternative hypothesis of interest is $S_1(t) \geq S_2(t)$ for all t ($t \in [0, \Gamma]$), $S_1(.) \neq S_2(.)$, in which $S_1(t)$ and $S_2(t)$ are the survival functions at time t in group 1 and 2, respectively. Because the log-rank test is sensitive to alternatives of ordered hazard functions however not necessarily to ordered survival functions, we consider the Weighted Kaplan-Meier test (WKM) which is directly based on the estimated survival functions (Kaplan and Meier, 1958; Gill, 1980; Breslow et al., 1984; Fleming et al., 1987; Pepe and Fleming, 1989; Fleming and Harrington, 1991). Because the variance of the WKM test statistic is unknown, we will use the permutation test. The Renyi-type test based on the supremum versions of WKM test statistics will be considered as well (Gill, 1980; Fleming et al., 1987; Cai and Shen, 2000). We will also present the permutation tests for log-rank and its associated Renyi-type test statistics in comparison with WKM test statistics. Four configurations of the alternative hypotheses, survival difference between two groups in proportional and in early, middle or late stage of study, are investigated by using WKM and log-rank tests. Simulation studies will be conducted. Empirical type I error and empirical testing power resulting from these permutation tests will be compared among WKM, log-rank, and their

Renyi-type tests. Recommendation of preferred test with regards to each configuration will be made.

Notation

Assume that there are n subjects in a full cohort. Assume that there are 2 groups and n_j subjects in group j (j = 1, 2) with $n = n_1 + n_2$. Let T_{ij} represent the event time and C_{ij} the censoring time for subject i in group j ($i = 1,\ldots, n_j$), and they are independent of each other. Let J_{ij} be the dichotomous variable indicating the exposure status, $X_{ij} = T_{ij} \wedge C_{ij}$ be the observed time, and $\Delta_{ij} = I(T_{ij} \leq C_{ij})$ be the failure indicator, in which $\Delta_{ij} = 1$ denotes failure and $\Delta_{ij} = 0$ denotes censoring. Assume $\tilde{n}$ subjects are randomly sampled into a sub-cohort from n subjects without replacement. Let $\xi_{ij} = 1$ denote that subject i in group j is selected into the sub-cohort and $\xi_{ij} = 0$ otherwise. Denote by γ the proportion of group 1 and $(1-\gamma)$ the proportion of group 2. All subjects in the sub-cohort and all events in the full cohort consist of the case-cohort sample. The observed data in the case-cohort is

$$(J_{ij}(\xi_{ij} + (1-\xi_{ij})\Delta_{ij}), X_{ij}, \Delta_{ij}), i = 1,\ldots, n_j; j = 1,2,$$

where $(\xi_{ij} + (1-\xi_{ij})\Delta_{ij}) = 1$ indicates that J_{ij} is observable and $(\xi_{ij} + (1-\xi_{ij})\Delta_{ij}) = 0$ otherwise.

Case-Cohort Weighted Kaplan-Meier Test Statistics

Kaplan-Meier estimator (KM) of survival is considered as a natural statistic to measure the difference in stochastically ordered survival between two groups (Kaplan and

Meier, 1958). However, it is known that the KM estimator can be unstable for t close to t_0 in the heavy censoring data; and the appropriate weight function was recommended to ensure the statistic stability (Pepe and Fleming, 1989). The Weighted Kaplan-Meier test statistics (WKM) can be defined as

$$WKM = \int_0^t \hat{\omega}(s)[\hat{S}_1(s) - \hat{S}_2(s)]ds,$$

in which $\hat{\omega}(s)$ is the weight function, $\hat{S}_1(s)$ and $\hat{S}_2(s)$ are the KM estimator of the survival function for the treatment group 1 and 2 at time s, respectively, by using the case-cohort data.

It is known that given a full cohort data, the KM estimator of survival function can be obtained by $\hat{S}(t) = \prod_{s<t}\left[1 - \frac{d\overline{N}(s)}{\overline{Y}(s)}\right]$, where $\overline{Y}(s)$ is the number of subjects at risk at time s ($s < t$) and $\overline{N}(s)$ the counting process for the event before time s ($s < t$) in the full cohort (Kaplan and Meier, 1958). However, in a case-cohort sample which includes all events and a sub-cohort, $\overline{N}(s)$ is known, and $\overline{Y}(s)$ can be estimated by $\widetilde{Y}(s)/p$, where $\widetilde{Y}(s)$ is the number of subjects at risk at time s in sub-cohort and p is the sampling proportion of sub-cohort (Prentice, 1986; Self and Prentice, 1988; Samuelsen, 2010). Thus the case-cohort KM estimator of survival function can be obtained as $\hat{S}(t) = \prod_{s<t}\left[1 - (p)\frac{d\overline{N}(s)}{\widetilde{Y}(s)}\right]$. Similarly, the KM estimator of the survival function for the treatment group 1 and 2 can be obtained as $\hat{S}_1(t) = \prod_{s<t}\left[1 - (p)\frac{d\overline{N}_1(s)}{\widetilde{Y}_1(s)}\right]$ and $\hat{S}_2(t) = \prod_{s<t}\left[1 - (p)\frac{d\overline{N}_2(s)}{\widetilde{Y}_2(s)}\right]$.

We consider 4 common configurations of alternative hypotheses, proportional, early, middle, and late differences in survival functions between two groups (Cai and Shen, 2000). Four weight functions are considered accordingly in the following:

1. $\hat{\omega}_1(t) = 1$ for proportional difference;
2. $\hat{\omega}_2(t) = \hat{S}(t)$ for early difference;
3. $\hat{\omega}_3(t) = \sqrt{\hat{S}(t)(1-\hat{S}(t))}$ for middle difference;
4. $\hat{\omega}_4(t) = (1-\hat{S}(t))$ for late difference.

In addition, we investigate the Renyi-type test based on the supremum of WKM statistics, in a format of $WKMR = \sup_{t\geq 0} WKM$. The Renyi-type test considers the maximal deviation at each time *t* and may be more powerful against a variety of configurations of alternative hypotheses (Gill, 1980; Fleming et al., 1987; Cai and Shen, 2000).

Case-Cohort Log-rank Test Statistics

Cai and Zeng (2004) proposed a log-rank type of test statistic for the case-cohort study, of which the test statistic was equivalent to the score test based on the pseudo-partial likelihood function in Self and Prentice (1988). The test statistics was given as

$$SP = \int_0^t \frac{\hat{\omega}(s)\widetilde{Y}_1(s)\widetilde{Y}_2(s)}{\widetilde{Y}_1(s)+\widetilde{Y}_2(s)}\left\{\frac{d\overline{N}_1(s)}{\widetilde{Y}_1(s)} - \frac{d\overline{N}_2(s)}{\widetilde{Y}_2(s)}\right\},$$

in which $\hat{\omega}(s)$, $\overline{N}(s)$ and $\widetilde{Y}(s)$ are defined as in the previous section.

Similarly, the Renyi-type test based on the supremum of log-rank test statistics SP will be conducted as $SPR = \sup_{t\geq 0} SP$ in order to compare with WKM.

Permutation Tests

Denote by WKM_0 the permutation WKM test statistics at $t = \Gamma$, $WKMR_0$ the permutation WKM-associated Renyi-type test statistics, SP_0 the permutation log-rank (SP) test statistics at $t = \Gamma$, and SPR_0 the permutation SP-associated Renyi-type test statistics using the observed case-cohort data. Let C denote the number of permutation samples (c = 1, 2,….C). In each sample the event time and the censoring indicator will remain the same while the treatment will be assigned at random. Each sample shares the equal probability of $1/C$ in the permutation distribution. Denote WKM_c, $WKMR_c$, SP_c and SPR_c the test statistics obtained from the c^{th} permutation sample generated from the observed data, then the p-values of WKM, SP, and its Renyi-type permutation tests will be calculated as

$$\hat{P} = \sum_{c=1}^{C} I\{WKM_c \geq WKM_0\}/C \text{, or } \hat{P} = \sum_{c=1}^{C} I\{WKMR_c \geq WKMR_0\}/C$$

$$\hat{P} = \sum_{c=1}^{C} I\{SP_c \geq SP_0\}/C \text{, or } \hat{P} = \sum_{c=1}^{C} I\{SPR_c \geq SPR_0\}/C .$$

Simulation Studies

Simulation Procedures

A series of simulation studies will be conducted to evaluate the performance of WKM and log-rank permutation tests. The empirical size and power of the permutation tests will be compared. The case-cohort samples will be generated by using the following procedures/parameters:

1) Various full cohort sizes (1,000 or 4,000), event proportions (1%, 5%, or 10%), and sub-cohort sampling fractions (0.1 or 0.2) will be considered. All subjects will be assigned to one of the two groups (denote $Z = 1$ for group 1, and $Z = 0$ for group 2). The proportion of group 1 will be set to 0.5. The significant level α will be set at 0.05.

2) The failure time will be generated from the piecewise exponential distribution by using the following procedures:

   a. For each of 4 configurations, assign appropriate $V_1 < V_2 ... < V_H$, of which V_h's ($h = 1, 2, \ldots H$) are the cut points for the time interval associated with the piecewise exponential distribution with $V_1 \equiv 0$ and $V_{H+1} \equiv \infty$. Assign ˚ppropriate λ for each treatment group and θ (log-hazard ratio) in each time internal. To examine the empirical size, data will be generated with no difference between two groups so that two groups will have the same λ and $\theta = 0$ in each time interval.

   b. Generate an independent uniform (0, 1) variant u.

   c. Calculate failure time $t = \sum_{h=1}^{H} \varsigma_h * I(V_h \le \varsigma_h \le V_{h+1})$, in which

$$\varsigma_h = (-In(1-u) - \sum_{m=1}^{h} (V_m - V_{m-1}) e^{\theta_{m-1} Z} \lambda_{m-1}) e^{-\theta_h Z} / \lambda_h + V_h (m = 1,2,...H)$$

   and $I(A)$ is the indicator function for A.

3) The censoring time is generated from a uniform distribution between $[0, \Gamma]$, where Γ is varied with different censoring proportions based on the given event proportions.

4) Each simulation will be repeated 2,000 times.

5) Empirical type I error will be calculated for WKM and log-rank permutation tests and corresponding Renyi-type tests under no difference set up.
6) Empirical testing power will be calculated for WKM and log-rank permutation tests and corresponding Renyi-type tests under the proportional, early, middle, and late difference situations.
7) SAS Version 9.2 and/or above will be used for programming.

Simulation Results

The simulation results will be compared in order to evaluate the performance of WKM and log-rank permutation tests. It will be observed if the empirical type I error results from WKM test, log-rank test, and their Renyi-type tests are comparable to the level of significance. The empirical testing power from WKM and log-rank tests will be compared to show which one yields a higher power than the others under the situation that the survival in two treatment groups is stochastically ordered. Furthermore, WKM tests with different weight functions $\hat{\omega}_1(t), \hat{\omega}_2(t), \hat{\omega}_3(t), \text{and } \hat{\omega}_4(t)$ will be evaluated to examine which one has higher empirical testing power than the others for the configurations with proportional, early, middle, and late differences. Finally, the Renyi-type tests will be compared against their counterparts for each simulation sample.

Conclusion and Discussion

If the results from the empirical type I error are observed to be close to the level of significance α, it can be concluded that the corresponding tests, WKM test, log-rank test, and/or their Renyi-type tests, are valid for the case-cohort study in testing the null

hypothesis of no difference in survival between two treatment groups. Moreover, the closeness to the α will reflect the sensitivity of a test in testing the null hypothesis.

The higher empirical testing power from WKM test or log-rank test will support the conclusion that the corresponding test is more powerful and efficient than the other in testing the stochastically ordered survivals in two treatment groups. Furthermore, if the WKM test associated with the weight function of $\hat{\omega}_1(t), \hat{\omega}_2(t), \hat{\omega}_3(t), \text{or } \hat{\omega}_4(t)$ appears to have a higher power than the others for a specific configuration of alternative hypothesis, this WKM test associated with the choice of weight function is recommended for testing the alternative hypothesis with the specific configuration. For instance, if it is observed that the empirical testing power from WKM test with weight function of $\hat{\omega}_4(t)$ has the highest power among all other tests with different weight functions in testing the difference in survival between two groups, in which the difference appears to be a late difference, WKM test with weight function of $\hat{\omega}_4(t)$ would be recommended to test the late difference configuration situation.

The Renyi-type tests will be compared against their counterparts for each situation discussed above. In testing the null hypothesis of no difference in survival between two groups, whether the Renyi-type tests are more sensitive than their counterparts WKM and/or log-rank tests will be revealed by their empirical type I error if it is closer to α than the others. In testing the alternative hypothesis with different configuration situations, Renyi-type tests will be recommended if they are more powerful than their counterparts by comparing their empirical testing power results in each specific situation.

References

Andersen, V., Østergaard, M., Christensen, J., Overvad, K., Tjønneland, A., and Vogel, U. (2009). Polymorphisms in the xenobiotic transporter Multidrug Resistance 1 (MDR1) and interaction with meat intake in relation to risk of colorectal cancer in a Danish prospective case-cohort study. *BMC Cancer* 9:407 doi:10.1186/1471-2407-9-407.

ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: Design and Objectives. *American Journal of Epidemiology* 129, 687-702.

Ballantyne, C. M., Hoogeveen, R. C., Bang, H. J., et al. (2004). Lipoprotein associated phospholipase A(2), high-sensitivity C-reactive protein, and risk for ischemic stroke in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. *Circulation* 109 (7), 837–842.

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

Barlow, W. E. and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* 75, 65-74.

Barlow, W. E., Ichikawa, L, Rosner, D., et al. (1999). Analysis of case cohort designs. *Journal of Clinical Epidemiology* 52 (12), 1165–1172.

Batalla A, Gonzalez F, Suarez E, Martinez J (1999). Transient complete auriculoventricular block in Lyme disease. *Revista Espanola de Cardiologia* 52, 529–531.

Beelen R., Hoek, G., van den Brandt, P. A., Goldbohm, R. A., Fischer, P., Schouten, L. J., Jerrett, M., Hughes, E., Armstrong, B., and Brunekreef, B. (2008). Long-Term Effects of Traffic-Related Air Pollution on Mortality in a Dutch Cohort (NLCS-AIR Study). *Environ Health Perspect* 116:196–202. doi:10.1289/ehp.10767.

Boice, J. and Monson, R. (1977). Breast cancer in women after repeated fluoroscopic examinations of the chest. *Journal of National Cancer Institute* 59 (3), 823–32.

Borgan, O., Goldstein L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *The Annals of Statistics* 23, 1749-1778.

Borgan, Ø., Langholz, B., Samuelsen, S., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* 6, 39–58.

Breslow, N., Lumley, T., Ballantyne, C., Chambless, L., and Kulich, M (2009). Using the Whole Cohort in the Analysis of Case-Cohort Data. *American Journal of Epidemiology* 169 (11), 1398–1405.

Breslow, N. E., Edler, L., and Berger, J. (1984). A two-sample censored data rank test for acceleration. *Biometrics* 40, 1049-1062.

Cai, J. and Shen, Y. (2000). Permutation tests for comparing marginal survival functions with clustered failure time data. *Statistics in Medicine* 19, 2963-2973.

Cai, J. and Zeng, D. (2004). Sample Size/Power Calculation for Case-Cohort Studies. *Biometrics* 60, 1015-1024.

Cai, J. and Zeng, D. (2007). Power Calculation for Case-Cohort Studies with Non-rare Events. *Biometrics* 63, 1288-1295.

Chen, K. N. (2001). Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B* 63, 791-809.

Chen, H. (2001a). Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. *Journal of American Statistical Association* 96, 1446-1457.

Chen, H. (2001b). Fitting semiparametric transformation regression models to data from a modified case-cohort design. *Biometrika* 88, 255-268.

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

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

Deville J. C., and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association* 87 (418), 376–382.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Pennsylvania: *Society for Industrial and Applied Mathematics*.

Fleming, T. R., Harrington D. P., and O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* 82, 312-320.

Fleming, T. R. and Harrington D. P. (1991). Counting processes and survival analysis. New Jersey: Wiley.

Gill, R. D. (1980). *Censoring and stochastic integrals*. Mathematical Centre Tracts 124. Amsterdam: Mathematical Centrum.

Herder C., Baumert J., Zierer A., Roden M., Meisinger C., et al. (2011). Immunological and Cardiometabolic Risk Factors in the Prediction of Type 2 Diabetes and Coronary Events: MONICA/KORA Augsburg Case-Cohort Study. *PLoS ONE* 6(6): e19852. doi:10.1371/journal.pone.0019852

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association* 47 (260), 663–685.

Hrubec, Z., Boice, J., Monson, R., and Rosenstein, M. (1989). Breast cancer after multiple chest flouroscopies: Second follow-up of Massachusetts women with tuberculosis. *Cancer Research* 49, 229–34.

Kalbfleisch, J. and Lawless, J. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine* 7, 149–60.

Kang S.; Cai J. (2009). Marginal hazards regression for retrospective studies within cohort with possibly correlated failure time data. *Biometrics* 65, 405-14.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53, 457-481.

Kulathinal, S., Karvanen, J., Saarela, O., and Kuulasmaa, K. (2007). Case-cohort design in practice – experiences from the MORGAM. *Project Epidemiologic Perspectives & Innovations*, 4:15 doi: 10.1186/1742-5573-4-15.

Langholz, B., and Jiao, J. (2006). Computational methods for case-cohort studies. *Computational Statistics & Data Analysis* 51 (8), 3737-3748.

Liao, D., Cai, J., Rosamond, W. D., Barnes, R. W., Hutchinson, R. G., Whitsel, E. A., Rautaharju, P., Heiss, G. (1997). Cardiac Autonomic Function and Incident Coronary Heart Disease: A Population Based Case-Cohort Study - The ARIC Study. *American Journal of Epidemiology* 145, 696-706.

Lin, D. Y. and Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of American Statistical Association* 88, 1341-1349.

Lin, D. Y., and Wei, L. J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association* 84, 1074-1078.

McElrath, M. J., De Rosa, S. C., Moodie, Z., Dubey, S., Kierstead, L., Janes, H., Defawe, O. D., Carter, D. K., Hural, J., Akondy, R., Step Study Protocol Team et al. (2008). HIV-1 vaccine-induced immunity in the test-of-concept Step Study: A case-cohort analysis. *Lancet* 372, pp. 1894–1905.

Pepe, M. P. and Fleming, T. R. (1989). Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics* 45, 497-507.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73, 1-11.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63, 581-592.

Samuelsen, S. O. (2010). Nested case-control and case-cohort studies: an overview. Slides, Graduate School in Biostatistics, The University of Oslo.

Sarndal, C. E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York, NY, *Springer-Verlag*.

Savitz, D. A., Cai, J., van Wijngaarden, E., Loomis, D., Mihlan, G., Dufort, V., Kleckner, R.C., Nylander-French, L.A., Kromhout, H. and Zhou, H (2000). Case-Cohort Analysis of Brain Cancer and Leukemia in Electric Utility Workers Using a Refined Magnetic Field Job-Exposure Matrix. *American Journal of Industrial Medicine* 38, 417-425.

Schouten, E., Dekker, J., Kok, F., Le Cessie, S., Van Houwelingen, H., Pool, J., and Vandenbroucke, J. (1993). Risk ratio and rate ratio estimation in case-cohort designs: Hypertension and cardiovascular mortality. *Statistics in Medicine* 12, 1733-1745.

Self, S. G., and Prentice, R. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics* 16, 64-81.

Sinner, M. F., Reinhard, W., Muller, M., Beckmann, B., Martens, E., Perz, S., Pfeufer, A., Winogradow, J., Stark, K., Meisinger, C., Wichmann, H., Peters, A., Riegger, G. A., Steinbeck, G., Christian Hengstenberg, C., Stefan Kaab, S. (2010). Association of Early Repolarization Pattern on ECG with Risk of Cardiac and All-Cause Mortality: A Population-Based Prospective Cohort Study (MONICA/KORA). *PLoS Med* 7(7): e1000314 doi:10.1371/journal.pmed.1000314.

Szilagyi, P. G., Fairbrother, G., Griffin, M. R., Hornung, R. W., Donauer, S., Morrow, A., Altaye, M., Zhu, Y., Ambrose, S., Edwards, K. M., Poehling, K. A., Lofthus, G., Holloway, M., Finelli, L., Iwane, M., Staat, M. A. (2008). Influenza Vaccine Effectiveness Among Children 6 to 59 Months of Age During 2 Influenza Seasons A Case-Cohort Study. *Arch Pediatr Adolesc Med*. 162(10): 943-951.

Appendices

A. SCC Asymptotic Variance Derivation

From the expression of ψ_l in the previous sections and note that $d\Lambda_l(t) = \frac{dS_l(t)}{S_l(t)}$,

$\Lambda_l(t) \approx p_l \int_0^{\tau} \frac{d\overline{N}_{l1}(t) + d\overline{N}_{l2}(t)}{\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t)}$, $\gamma_l \pi_{l1}(t) S_l(t) \approx \frac{\widetilde{Y}_{l1}(t)}{\widetilde{n}_l}$, and $(1 - \gamma_l)\pi_{l2}(t) S_l(t) \approx \frac{\widetilde{Y}_{l2}(t)}{\widetilde{n}_l}$, ψ can be approximated by

$$\hat{\psi} = \frac{1}{n}\sum_{l=1}^{L} n_l \left\{ \frac{\hat{p}_l(1-\hat{p}_l)}{\widetilde{n}_l} \iint \frac{\omega(t)\omega(w)}{\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t)} \frac{\widetilde{Y}_{l1}(t)\widetilde{Y}_{l1}(w)\widetilde{Y}_{l2}(t \vee w) + \widetilde{Y}_{l2}(t)\widetilde{Y}_{l2}(w)\widetilde{Y}_{l1}(t \vee w)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} \right.$$

$$\left. \times \frac{d\overline{N}_{l1}(w) + d\overline{N}_{l2}(w)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} \frac{d\overline{N}_{l1}(t) + d\overline{N}_{l2}(t)}{\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t)} \right\}$$

$$= \frac{1}{n}\sum_{l=1}^{L} \left\{ (1-\hat{p}_l) \int \frac{\omega(t)}{(\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t))^2} \left[\widetilde{Y}_{l1}(t) \int_0^t \frac{\omega(w)\widetilde{Y}_{l1}(w)\widetilde{Y}_{l2}(t \vee w)}{(\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w))^2} [d\overline{N}_{l1}(w) + d\overline{N}_{l2}(w)] \right.\right.$$

$$\left.\left. + \widetilde{Y}_{l2}(t) \int_0^t \frac{\omega(w)\widetilde{Y}_{l2}(w)\widetilde{Y}_{l1}(t \vee w)}{(\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w))^2} [d\overline{N}_{l1}(w) + d\overline{N}_{l2}(w)] \right] [d\overline{N}_{l1}(t) + d\overline{N}_{l2}(t)] \right\}$$

$$= \frac{1}{n}\sum_{l=1}^{L} \left\{ (1-\hat{p}_l) \int \frac{\omega(t)}{(\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t))^2} \left[2\widetilde{Y}_{l1}(t)\widetilde{Y}_{l2}(t) \{ \int_0^t \frac{\omega(w) I(w \le t)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} [d\overline{N}_{l1}(w) + d\overline{N}_{l2}(w)] \right.\right.$$

$$\left.\left. - \frac{\omega(w) I(w = t)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} \} [d\overline{N}_{l1}(w) + d\overline{N}_{l2}(w)] \right] [d\overline{N}_{l1}(t) + d\overline{N}_{l2}(t)] \right\}$$

$$= \frac{1}{n}\sum_{l=1}^{L} \left\{ 2(1-\hat{p}_l) \int \frac{\omega(t)\widetilde{Y}_{l1}(t)\widetilde{Y}_{l2}(t)}{(\widetilde{Y}_{l1}(t) + \widetilde{Y}_{l2}(t))^2} \left[\int_0^t \frac{\omega(w) I(w \le t)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} [\sum_{i=1}^{n_{l1}} dN_{li1}(w) + \sum_{i=1}^{n_{l2}} dN_{li2}(w)] \right.\right.$$

$$\left.\left. - \frac{\omega(w) I(w = t)}{\widetilde{Y}_{l1}(w) + \widetilde{Y}_{l2}(w)} [\sum_{i=1}^{n_{l1}} dN_{li1}(w) + \sum_{i=1}^{n_{l2}} dN_{li2}(w)] \right] [\sum_{i=1}^{n_{l1}} dN_{li1}(t) + \sum_{i=1}^{n_{l2}} dN_{li2}(t)] \right\}.$$

Since $\widetilde{Y}_{lj}(t)=\sum_{i=1}^{\widetilde{n}_l} I(X_{lij}\geq t),\ \overline{N}_{lj}(t)=\sum_{i=1}^{n_{lj}} N_{lij}(t), \text{and } N_{lij}(t)=\Delta_{lij} I(X_{lij}\leq t),$

$$\hat{\psi}=\frac{1}{n}\sum_{l=1}^{L}\left\{2(1-\hat{p}_l)\left[\sum_{j=1}^{2}\sum_{i=1}^{n_{lj}}\frac{\Delta_{lij}\omega(X_{lij})\widetilde{Y}_{l1}(X_{lij})\widetilde{Y}_{l2}(X_{lij})}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^2}\{\sum_{j'=1}^{2}\sum_{i'=1}^{n_{lj'}}\frac{\Delta_{li'j'}\omega(X_{li'j'})I(X_{li'j'}\leq X_{lij})}{\widetilde{Y}_{l1}(X_{li'j'})+\widetilde{Y}_{l2}(X_{li'j'})}\right.\right.$$

$$\left.\left.-\frac{1}{2}\frac{\omega(X_{lij})}{\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij})}\}\right]\right\}$$

$$=\frac{1}{n}\sum_{l=1}^{L}2(1-\hat{p}_l)\sum_{j=1}^{2}\sum_{i=1}^{n_{lj}}\left\{\frac{\Delta_{lij}\omega(X_{lij})\widetilde{Y}_{l1}(X_{lij})\widetilde{Y}_{l2}(X_{lij})}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^2}\times\sum_{j'=1}^{2}\sum_{i'=1}^{n_{lj'}}\frac{\Delta_{li'j'}\omega(X_{li'j'})I(X_{li'j'}\leq X_{lij})}{\widetilde{Y}_{l1}(X_{li'j'})+\widetilde{Y}_{l2}(X_{li'j'})}\right\}$$

$$-\frac{1}{n}\sum_{l=1}^{L}(1-\hat{p}_l)\sum_{j=1}^{2}\sum_{i=1}^{n_{lj}}\frac{\Delta_{lij}\omega(X_{lij})^2\widetilde{Y}_{l1}(X_{lij})\widetilde{Y}_{l2}(X_{lij})}{(\widetilde{Y}_{l1}(X_{lij})+\widetilde{Y}_{l2}(X_{lij}))^3}.$$

B. GSCC Asymptotic Variance Derivation

From the expression in the previous sections, the GSCC asymptotic variance of $n^{-1/2}T_n, \sigma_T^2$, consists of 3 terms, namely Term A, B, and C. Term A (σ^2) is the asymptotic variance of the usual log-rank test statistic in a form of

$$\sigma^2 = \sum_{l=1}^{L} v_l \left\{ \gamma_l \int_0^{\Gamma} (1-\alpha_l(t))^2 P_l(C > t \mid J = 1)\, e^{-\Lambda_l(t)} d\Lambda_1(t) \right.$$

$$\left. + (1-\gamma_l) \int_0^{\Gamma} \alpha_l(t)^2 P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right\}.$$

Also by knowing

$$\varepsilon_l(X,J) = -\int_0^X \omega(t)(1-\alpha_l(t))d\Lambda_l(t)I(J=1) + \int_0^X \omega(t)\alpha_l(t)d\Lambda_l(t)I(J=2),$$

$$u_l(X,J) = v_l(X,J) + \omega(X)(1-\alpha_l(X))I(J=1) - \omega(X)\alpha_l(X)I(J=2)$$

$$= -\int_0^X \omega(t)(1-\alpha_l(t))d\Lambda_l(t)I(J=1) + \int_0^X \omega(t)\alpha_l(t)d\Lambda_l(t)I(J=2)$$

$$+ \omega(X)(1-\alpha_l(X))I(J=1) - \omega(X)\alpha_l(X)I(J=2), \text{and } \varepsilon_l(X,J)(1-\Delta) =$$

$F_l(X,J,\Delta)$, we can derive the Term B as below:

$$\sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} Var(\varepsilon_l(X,J)(1-\Delta))$$

$$= \sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} \left\{ \int (\varepsilon_l(X,J)(1-\Delta))^2 dF_l(X,J,\Delta) - [\int (\varepsilon_l(X,J)(1-\Delta)) dF_l(X,J,\Delta)]^2 \right\}$$

$$= \sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} \left\{ \int (\varepsilon_l(X,J))^2 dF_l(X,J,\Delta=0) - [\int (\varepsilon_l(X,J)) dF_l(X,J,\Delta=0)]^2 \right\}$$

$$= \sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} \left\{ \int [-\int_0^{T \wedge C} (1-\alpha_l(w)) d\Lambda_l(w)]^2 dF_l(T \wedge C, J=1, C \le T) \right.$$

$$\left. + \int [\int_0^{T \wedge C} (\alpha_l(w)) d\Lambda_l(w)]^2 dF_l(T \wedge C, J=2, C \le T) \right.$$

$$-\left[\int[-\int_0^{T\hat{}C}(1-\alpha_l(w))d\Lambda_l(w)]dF_l(T\hat{}C, J=1, C\le T)\right.$$

$$\left.+\int[\int_0^{T\hat{}C}(\alpha_l(w))d\Lambda_l(w)]dF_l(T\hat{}C, J=2, C\le T)\right]^2\Bigg\}$$

$$=\sum_{l=1}^{L} v_l \frac{1-p_l}{p_l}\Bigg\{\gamma_l\int[-\int_0^{C}(1-\alpha_l(w))d\Lambda_l(w)]^2 dF_l(C, C\le T\mid J=1)$$

$$+(1-\gamma_l)\int[\int_0^{C}(\alpha_l(w))d\Lambda_l(w)]^2 dF_l(C, C\le T\mid J=2)$$

$$-\left[\gamma_l\int[-\int_0^{C}(1-\alpha_l(w))d\Lambda_l(w)]dF_l(C, C\le T\mid J=1)\right.$$

$$\left.+(1-\gamma_l)\int[\int_0^{C}(\alpha_l(w))d\Lambda_l(w)]dF_l(C, C\le T\mid J=2)\right]^2\Bigg\}$$

$$=\sum_{l=1}^{L} v_l \frac{1-p_l}{p_l}\Bigg\{\gamma_l\int_0^{\Gamma}[-\int_0^{t}(1-\alpha_l(w))d\Lambda_l(w)]^2 e^{-\Lambda_l(t)}dP_l(C\le T\mid J=1)$$

$$+(1-\gamma_l)\int_0^{\Gamma}[\int_0^{t}\alpha_l(w)d\Lambda_l(w)]^2 e^{-\Lambda_l(t)}dP_l(C\le T\mid J=2)$$

$$-\left[\gamma_l\int_0^{\Gamma}[-\int_0^{t}(1-\alpha_l(w))d\Lambda_l(w)]e^{-\Lambda_l(t)}dP_l(C\le T\mid J=1)\right.$$

$$\left.+(1-\gamma_l)\int_0^{\Gamma}[-\int_0^{t}\alpha_l(w)d\Lambda_l(w)]e^{-\Lambda_l(t)}dP_l(C\le T\mid J=2)\right]^2\Bigg\}.$$

Given that

$$P_l(C\le T\mid J=1)=1-P_l(C>T\mid J=1), P_l(C\le T\mid J=2)=1-P_l(C>T\mid J=2),$$

after integration, we obtain Term B as

$$\sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} Var(\varepsilon_l(X,J)(1-\Delta))$$

$$=\sum_{l=1}^{L} v_l \frac{1-p_l}{p_l}\Bigg\{\gamma_l\int_0^{\Gamma}\Bigg[2\int_0^{t}(1-\alpha_l(w))d\Lambda_l(w))(1-\alpha_l(t))$$

$$-\left(\int_0^t (1-\alpha_l(w))d\Lambda_l(w)\right)^2\Big]P_l(C>t\mid J=1)e^{-\Lambda_l(t)}d\Lambda_l(t)$$

$$+(1-\gamma_l)\int_0^{\Gamma}\Big[2\int_0^t \alpha_l(w)d\Lambda_l(w)\alpha_l(t)-\left(\int_0^t \alpha_l(w)d\Lambda_l(w)\right)^2\Big]P_l(C>t\mid J=2)e^{-\Lambda_l(t)}d\Lambda_l(t)$$

$$-\Big[-\gamma_l\int_0^{\Gamma}\Big[(1-\alpha_l(t))-\int_0^t(1-\alpha_l(w))d\Lambda_l(w)\Big]P_l(C>t\mid J=1)e^{-\Lambda_l(t)}d\Lambda_l(t)$$

$$+(1-\gamma_l)\int_0^{\Gamma}\Big[\alpha_l(t)-\int_0^t\alpha_l(w)d\Lambda_l(w)\Big]P_l(C>t\mid J=2)e^{-\Lambda_l(t)}d\Lambda_l(t)\Big]^2\Big\}.$$

Similarly, Consider $(u_l(X,J)\mid\Delta=1,\xi=0)=M_l(X,J,\Delta,\xi)$, Term C can be derived

by $\sum_{l=1}^{L}v_l\dfrac{P_l(\Delta=1)(1-p_l)(1-q_l)}{q_l}Var(u_l(X,J)\mid\Delta=1,\xi=0)$

$$=\sum_{l=1}^{L}v_l\frac{P_l(\Delta=1)(1-p_l)(1-q_l)}{q_l}\Big\{\int(u_l(X,J)\mid(\Delta=1,\xi=0))^2\,dM_l(X,J,\Delta,\xi)$$

$$-[\int(u_l(X,J)\mid(\Delta=1,\xi=0))dM_l(X,J,\Delta,\xi)]^2\Big\}$$

$$=\sum_{l=1}^{L}v_l\frac{P_l(\Delta=1)(1-p_l)(1-q_l)}{q_l}\left\{\int\frac{(u_l(X,J))^2}{P(\Delta=1)}dM_l(X,J,\Delta=1,\xi=0)\right.$$

$$\left.-[\int(\frac{u_l(X,J)}{P_l(\Delta=1)})dM_l(X,J,\Delta=1,\xi=0)]^2\right\}$$

$$=\sum_{l=1}^{L}v_l\frac{(1-p_l)(1-q_l)}{q_l}\Big\{\int[-\int_0^{T\hat{}C}(1-\alpha_l(w))d\Lambda_l(w)+(1-\alpha_l(T\hat{}C))]^2\,dM_l(T\hat{}C,J=1,C>T,\xi=0)$$

$$+\int[\int_0^{T\hat{}C}(\alpha_l(w))d\Lambda_l(w)-\alpha_l(T\hat{}C)]^2\,dM_l(T\hat{}C,J=2,C>T,\xi=0)$$

$$-\frac{1}{P_l(\Delta=1)}\Big[\int[-\int_0^{T\hat{}C}(1-\alpha_l(w))d\Lambda_l(w)+(1-\alpha_l(T\hat{}C))]dM_l(T\hat{}C,J=1,C>T,\xi=0)$$

$$+\int[\int_0^{T\hat{}C}(\alpha_l(w))d\Lambda_l(w)-\alpha_l(T\hat{}C)]dM_l(T\hat{}C,J=2,C>T,\xi=0)\Big]^2\Big\}$$

$$= \sum_{l=1}^{L} v_l \frac{(1-p_l)(1-q_l)}{q_l} \Big\{ \gamma_l \int [-\int_0^T (1-\alpha_l(w))d\Lambda_l(w) + (1-\alpha_l(T))]^2 \, dM_l(T, C > T \mid J = 1)$$

$$+ (1-\gamma_l) \int [\int_0^T (\alpha_l(w))d\Lambda_l(w) - \alpha_l(T)]^2 \, dM_l(T, C > T \mid J = 2)$$

$$- \frac{1}{P_l(\Delta = 1)} \Big[\gamma_l \int [-\int_0^T (1-\alpha_l(w))d\Lambda_l(w) - (1-\alpha_l(T))] dM_l(T, C > T \mid J = 1)$$

$$+ (1-\gamma_l) \int [\int_0^T (\alpha_l(w))d\Lambda_l(w) - \alpha_l(T)] dM_l(T, C > T \mid J = 2) \Big]^2 \Big\}$$

$$= \sum_{l=1}^{L} v_l \frac{(1-p_l)(1-q_l)}{q_l} \Big\{ \gamma_l \int_0^\Gamma [-\int_0^t (1-\alpha_l(w))d\Lambda_l(w) + (1-\alpha_l(t))]^2 e^{-\Lambda_l(t)} dP_l(C > T \mid J = 1)$$

$$+ (1-\gamma_l) \int_0^\Gamma [\int_0^t \alpha_l(w)d\Lambda_l(w) - \alpha_l(t)]^2 e^{-\Lambda_l(t)} dP_l(C > T \mid J = 2)$$

$$- \frac{1}{P_l(\Delta = 1)} \Big[\gamma_l \int_0^\Gamma [-\int_0^t (1-\alpha_l(w))d\Lambda_l(w) + (1-\alpha_l(t))] e^{-\Lambda_l(t)} dP_l(C > T \mid J = 1)$$

$$+ (1-\gamma_l) \int_0^\Gamma [\int_0^t \alpha_l(w)d\Lambda_l(w) - \alpha_l(t)] e^{-\Lambda_l(t)} dP_l(C > T \mid J = 2) \Big]^2 \Big\}.$$

After integration, we obtain the Term C as

$$\sum_{l=1}^{L} v_l \frac{P_l(\Delta = 1)(1-p_l)(1-q_l)}{q_l} Var(u_l(X, J) \mid \Delta = 1, \xi = 0)$$

$$= \sum_{l=1}^{L} v_l \frac{(1-p_l)(1-q_l)}{q_l} \Big\{ \gamma_l \int_0^\Gamma \Big[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w))d\Lambda_l(w)) \Big]^2 P_l(C > t \mid J = 1) e^{-\Lambda_l(t)} d\Lambda_l(t)$$

$$+ (1-\gamma_l) \int_0^\Gamma \Big[\alpha_l(t) - \int_0^t \alpha_l(w)d\Lambda_l(w) \Big]^2 P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} d\Lambda_l(t)$$

$$- \frac{1}{P_l(\Delta = 1)} \Big[\gamma_l \int_0^\Gamma \Big[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w))d\Lambda_l(w) \Big] P_l(C > t \mid J = 1) e^{-\Lambda_l(t)} d\Lambda_l(t)$$

$$+ (1-\gamma_l) \int_0^\Gamma \Big[-\alpha_l(t) + \int_0^t \alpha_l(w)d\Lambda_l(w) \Big] P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} d\Lambda_l(t) \Big]^2 \Big\}.$$

Thus, the asymptotic variance of $n^{-1/2}T_n$ is to combine Term A, B and C in a form

$$\sigma_T^2 = \sigma^2 + \sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} Var(\varepsilon_l(X,J)(1-\Delta))$$

$$+ \sum_{l=1}^{L} v_l \frac{P_l(\Delta=1)(1-p_l)(1-q_l)}{q_l} Var(u_l(X,J) \mid \Delta=1, \xi=0)$$

$$= \sum_{l=1}^{L} v_l \left\{ \gamma_l \int_0^{\Gamma} (1-\alpha_l(t))^2 P_l(C>t \mid J=1)\, e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t}) \right.$$

$$\left. + (1-\gamma_l) \int_0^{\Gamma} \alpha_l(t)^2 P_l(C>t \mid J=2) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t}) \right\}$$

$$+ \sum_{l=1}^{L} v_l \frac{1-p_l}{p_l} \left\{ \gamma_l \int_0^{\Gamma} \left[2\int_0^t (1-\alpha_l(w)) d\Lambda_l(w))(1-\alpha_l(t)) \right.\right.$$

$$\left.\left. - (\int_0^t (1-\alpha_l(w)) d\Lambda_l(w))^2 \right] P_l(C>t \mid J=1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.$$

$$\left. + (1-\gamma_l) \int_0^{\Gamma} \left[2\int_0^t \alpha_l(w) d\Lambda_l(w) \alpha_l(t) - (\int_0^t \alpha_l(w) d\Lambda_l(w))^2 \right] P_l(C>t \mid J=2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.$$

$$\left. - \left[-\gamma_l \int_0^{\Gamma} \left[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w)) d\Lambda_l(w) \right] P_l(C>t \mid J=1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.\right.$$

$$\left.\left. + (1-\gamma_l) \int_0^{\Gamma} \left[\alpha_l(t) - \int_0^t \alpha_l(w) d\Lambda_l(w) \right] P_l(C>t \mid J=2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}$$

$$+ \sum_{l=1}^{L} v_l \frac{(1-p_l)(1-q_l)}{q_l} \left\{ \gamma_l \int_0^{\Gamma} \left[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w)) d\Lambda_l(w)) \right]^2 P_l(C>t \mid J=1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.$$

$$\left. + (1-\gamma_l) \int_0^{\Gamma} \left[\alpha_l(t) - \int_0^t \alpha_l(w) d\Lambda_l(w) \right]^2 P_l(C>t \mid J=2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.$$

$$\left. - \frac{1}{P_l(\Delta=1)} \left[\gamma_l \int_0^{\Gamma} \left[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w)) d\Lambda_l(w) \right] P_l(C>t \mid J=1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.\right.$$

$$\left.\left. + (1-\gamma_l) \int_0^{\Gamma} \left[-\alpha_l(t) + \int_0^t \alpha_l(w) d\Lambda_l(w) \right] P_l(C>t \mid J=2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}.$$

After integration by part, we obtain

$$\sigma_T^2 = \sum_{l=1}^{L} v_l \gamma_l \int_0^{\Gamma} \left\{ (1+\frac{(1-p_l)(1-q_l)}{q_l})(1-\alpha_l(t))^2 \right.$$

$$+ 2(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}) \int_0^t (1-\alpha_l(w)) d\Lambda_l(w)) (1-\alpha_l(t)$$

$$\left. + (-\frac{1-p_l}{P_l} + \frac{(1-p_l)(1-q_l)}{q_l}) \left\lfloor \int_0^t (1-\alpha_l(w)) d\Lambda_l(w)) \right\rfloor^2 \right\}$$

$$P_l(C > t \mid J = 1) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t})$$

$$+ \sum_{l=1}^{L} v_l (1-\gamma_l) \int_0^{\Gamma} \left\{ (1+\frac{(1-p_l)(1-q_l)}{q_l}) \alpha_l(t)^2 \right.$$

$$+ 2(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}) \int_0^t \alpha_l(w) d\Lambda_l(w)) \alpha_l(t)$$

$$\left. + (-\frac{1-p_l}{P_l} + \frac{(1-p_l)(1-q_l)}{q_l}) \left\lfloor \int_0^t \alpha_l(w) d\Lambda_l(w)) \right\rfloor^2 \right\}$$

$$P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t})$$

$$+ \sum_{l=1}^{L} v_l \left\{ \left(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)} \right) \left[\gamma_l \int_0^{\Gamma} \left[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w)) d\Lambda_l(w) \right] P_l(C > t \mid J = 1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right. \right.$$

$$\left. \left. + (1-\gamma_l) \int_0^{\Gamma} \left[-\alpha_l(t) + \int_0^t \alpha_l(w) d\Lambda_l(w) \right] P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}.$$

Since

$$P_l(\Delta = 1) = \gamma_l \int_0^{\Gamma} P_l(C > t \mid J = 1) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t}) + (1-\gamma_l) \int_0^{\Gamma} P_l(C > t \mid J = 2) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t}),$$

$$P_l(C > t \mid J = 2) = \beta_l P_l(C > t \mid J = 1), P_l(C > t \mid J = 1) = S_{cl}(t),$$

$$\alpha_l(t) = P_l(C > t \mid J = 1) / P_l(C > t)$$

$$= P_l(C > t \mid J = 1) / [P_l(C > t \mid J = 1) + P_l(C > t \mid J = 2)]$$

$$= \gamma_l / (\gamma_l + \beta_l (1-\gamma_l)),$$

$$1 - \alpha_l(t) = \beta_l (1-\gamma_l) / (\gamma_l + \beta_l (1-\gamma_l)),$$

we find that

$$\sum_{l=1}^{L} v_l \left\{ \left(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)} \right) \left[\gamma_l \int_0^{\Gamma} \left[(1-\alpha_l(t)) - \int_0^t (1-\alpha_l(w)) d\Lambda_l(w) \right] P_l(C>t \mid J=1) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.\right.$$
$$\left.\left. + (1-\gamma_l) \int_0^{\Gamma} \left[-\alpha_l(t) + \int_0^t \alpha_l(w) d\Lambda_l(w) \right] P_l(C>t \mid J=2) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}$$

$$= \sum_{l=1}^{L} v_l \left\{ \left(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)} \right) \left[\gamma_l \int_0^{\Gamma} \left[\frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} - \frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} \Lambda_l(t) \right] S_{cl}(t) e^{-\Lambda_l(t)} d\Lambda_l(t) \right.\right.$$
$$\left.\left. + (1-\gamma_l) \int_0^{\Gamma} \left[-\frac{\gamma_l}{(\gamma_l + \beta_l(1-\gamma_l)} + \frac{\gamma_l}{(\gamma_l + \beta_l(1-\gamma_l)} \Lambda_l(t) \right] \beta_l S_{cl}(t) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}$$

$$= \sum_{l=1}^{L} v_l \left\{ \left(\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)} \right) \right.$$
$$\left. \left[[\gamma_l \frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} - (1-\gamma_l) \frac{\gamma_l \beta_l}{\gamma_l + \beta_l(1-\gamma_l)}] \int_0^{\Gamma} [1-\Lambda_l(t)] S_{cl}(t) e^{-\Lambda_l(t)} d\Lambda_l(t) \right]^2 \right\}$$

$$= 0.$$

Integrating by part, we obtain

$$\sigma_T^2 = \sum_{l=1}^{L} v_l \gamma_l \int_0^{\Gamma} \left\{ (\frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)})^2 + \frac{1-p_l}{p_l} \left[2(\frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)})^2 \Lambda_l(t) \right.\right.$$
$$\left. - (\frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} \Lambda_l(t))^2 \right]$$
$$\left. + \frac{(1-p_l)(1-q_l)}{q_l} \left[\frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} - \frac{\beta_l(1-\gamma_l)}{\gamma_l + \beta_l(1-\gamma_l)} \Lambda_l(t) \right]^2 \right\} S_{cl}(t) e^{-\Lambda_l(t)} \mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l(1-\gamma_l)\int_0^{\Gamma}\left\{(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2+\frac{1-p_l}{p_l}\left[2(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2\Lambda_l(t)-(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)}\Lambda_l(t))^2\right]\right.$$

$$\left.+\frac{(1-p_l)(1-q_l)}{q_l}\left[\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)}-\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)}\Lambda_l(t)\right]^2\right\}\beta_l S_{cl}(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l\left\{\left(\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)}\right)\left[\gamma_l\int_0^{\Gamma}\left[\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}-\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}\Lambda_l(t)\right]S_{cl}(t)e^{-\Lambda_l(t)}d\Lambda_l(t)\right.\right.$$

$$\left.\left.+(1-\gamma_l)\int_0^{\Gamma}\left[-\frac{\gamma_l}{(\gamma_l+\beta_l(1-\gamma_l)}+\frac{\gamma_l}{(\gamma_l+\beta_l(1-\gamma_l)}\Lambda_l(t)\right]\beta_l S_{cl}(t)e^{-\Lambda_l(t)}d\Lambda_l(t)\right]^2\right\}$$

$$=\sum_{l=1}^{L} v_l\{\gamma_l(\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)})^2+\beta_l(1-\gamma_l)(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2\}\int_0^{\Gamma}S_{cl}(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} 2v_l\frac{1-p_l}{p_l}[\gamma_l(\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)})^2+\beta_l(1-\gamma_l)(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2]\int_0^{\Gamma}S_{cl}(t)\Lambda_l(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} 2v_l\frac{1-p_l}{p_l}[-\gamma_l(\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)})^2-\beta_l(1-\gamma_l)(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2]\int_0^{\Gamma}S_{cl}(t)\Lambda_l(t)^2e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l\frac{(1-p_l)(1-q_l)}{q_l}[\gamma_l(\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)})^2$$

$$+\beta_l(1-\gamma_l)(\frac{\gamma_l}{\gamma_l+\beta_l(1-\gamma_l)})^2]\int_0^{\Gamma}S_{cl}(t)(1-\Lambda_l(t))^2e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l\left\{\left(\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l P_l(\Delta=1)}\right)\left[(\gamma_l\frac{\beta_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}\right.\right.$$

$$\left.\left.-\beta_l(1-\gamma_l)\frac{\gamma_l}{(\gamma_l+\beta_l(1-\gamma_l)})\int_0^{\Gamma}S_{cl}(t)(1-\Lambda_l(t))e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})\right]\right\}$$

$$=\sum_{l=1}^{L} v_l\frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}\int_0^{\Gamma}S_{cl}(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} 2v_l \frac{1-p_l}{p_l} \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)} \int_0^{\Gamma} S_{cl}(t)\Lambda_l(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$-\sum_{l=1}^{L} 2v_l \frac{1-p_l}{p_l} \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)} \int_0^{\Gamma} S_{cl}(t)\Lambda_l(t)^2 e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l \frac{(1-p_l)(1-q_l)}{q_l} \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)} \int_0^{\Gamma} S_{cl}(t)(1-\Lambda_l(t))^2 e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$=\sum_{l=1}^{L} v_l \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[1+\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} S_{cl}(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l \frac{2\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} S_{cl}(t)\Lambda_l(t)e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t})$$

$$+\sum_{l=1}^{L} v_l \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[-\frac{(1-p_l)}{p_l}+\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} S_{cl}(t)\Lambda_l(t)^2 e^{-\Lambda_l(t)}\mathrm{d}\Lambda_l(\mathrm{t}).$$

Furthermore, Given that

$Y_l(t)/(n_l(\gamma_l+(1-\gamma_l)\beta_l)) \to S_{cl}(t)e^{-\Lambda_l(t)}$, $\int_0^t dN_l(s)/Y_l(s) \to \Lambda_l(t)$, $Y_l(t)=Y_{l1}(t)+Y_{l2}(t)$,

$N_l(t)=N_{l1}(t)+N_{l2}(t)$, the above equation can be further derived as

$$\sigma_T^2=\sum_{l=1}^{L} v_l \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[1+\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} \frac{Y_l(t)}{n_l(\gamma_l+\beta_l(1-\gamma_l))} d\int_0^t \frac{dN_l(s)}{Y_l(s)}$$

$$+\sum_{l=1}^{L} v_l \frac{2\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} \frac{Y_l(t)}{n_l(\gamma_l+\beta_l(1-\gamma_l))}\int_0^t \frac{dN_l(s)}{Y_l(s)} d\int_0^t \frac{dN_l(s)}{Y_l(s)}$$

$$+\sum_{l=1}^{L} v_l \frac{\beta_l\gamma_l(1-\gamma_l)}{\gamma_l+\beta_l(1-\gamma_l)}[-\frac{(1-p_l)}{p_l}+\frac{(1-p_l)(1-q_l)}{q_l}]\int_0^{\Gamma} \frac{Y_l(t)}{n_l(\gamma_l+\beta_l(1-\gamma_l))}(\int_0^t \frac{dN_l(s)}{Y_l(s)})^2 d\int_0^t \frac{dN_l(s)}{Y_l(s)}$$

$$= \sum_{l=1}^{L} v_l \left\{ \frac{\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{1 + \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} \int_0^{\Gamma} dN_l(t) \right\}$$

$$+ \sum_{l=1}^{L} v_l \left\{ \frac{2\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} \int_0^{\Gamma} \int_0^{t} \frac{dN_l(s)}{Y_l(s)} dN_l(t) \right\}$$

$$+ \sum_{l=1}^{L} v_l \left\{ \frac{\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{-\frac{1-p_l}{p_l} + \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} \int_0^{\Gamma} \{\int_0^{t} \frac{dN_l(s)}{Y_l(s)}\}^2 dN_l(t) \right\}.$$

We order the failures from the smallest to the largest and assume no two failures are tied to each other, then $\int_0^{\Gamma} dN_l(t) \approx D_l$, the total failures in stratum l, and $\int_0^{\Gamma} \int_0^{t} \frac{dN_l(s)}{Y_l(s)} dN_l(t) \approx \sum_{k_l=1}^{D_l} \sum_{k_l' \le k_l} \frac{1}{n_{k_l'}}$, where n_{k_l} represents the risk set size for k^{th} failure in stratum l. Similarly, we obtain $\int_0^{\Gamma} \{\int_0^{t} \frac{dN_l(s)}{Y_l(s)}\}^2 dN_l(t) \approx \sum_{k_l=1}^{D_l} \{\sum_{k_l' \le k_l} \frac{1}{n_{k_l'}}\}^2$. Thus, the above equation can be written as

$$\sigma_T^2 = \sum_{l=1}^{L} \frac{n_l}{n} \left\{ \frac{\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{1 + \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} D_l \right\}$$

$$+ \sum_{l=1}^{L} \frac{n_l}{n} \left\{ \frac{2\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} \sum_{k_l=1}^{D_l} \sum_{k_l' \le k_l} \frac{1}{n_{k_l'}} \right\}$$

$$+ \sum_{l=1}^{L} \frac{n_l}{n} \left\{ \frac{\beta_l \gamma_l (1-\gamma_l)}{\gamma_l + \beta_l (1-\gamma_l)} \{-\frac{1-p_l}{p_l} + \frac{(1-p_l)(1-q_l)}{q_l}\} \frac{1}{n_l(\gamma_l + \beta_l(1-\gamma_l))} \sum_{k_l=1}^{D_l} \{\sum_{k_l' \le k_l} \frac{1}{n_{k_l'}}\}^2 \right\}$$

$$= \frac{1}{n} \sum_{l=1}^{L} \left\{ \frac{\beta_l \gamma_l (1-\gamma_l)}{(\gamma_l + \beta_l (1-\gamma_l))^2} \{\frac{1}{p_l} - (\frac{(1-p_l)}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l})\} D_l \right\}$$

$$+ \frac{1}{n} \sum_{l=1}^{L} \left\{ \frac{2\beta_l \gamma_l (1-\gamma_l)}{(\gamma_l + \beta_l (1-\gamma_l))^2} \{\frac{1-p_l}{p_l} - \frac{(1-p_l)(1-q_l)}{q_l}\} \sum_{k_l=1}^{D_l} \sum_{k_l' \le k_l} \frac{1}{n_{k_l'}} \right\}$$

$$-\frac{1}{n}\sum_{l=1}^{L}\left\{\frac{\beta_l\gamma_l(1-\gamma_l)}{(\gamma_l+\beta_l(1-\gamma_l))^2}\{\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}\}\sum_{k_l=1}^{D_l}\{\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2\right\}$$

$$=\frac{1}{n}\sum_{l=1}^{L}\left\{\frac{\beta_l\gamma_l(1-\gamma_l)}{(\gamma_l+(1-\gamma_l)\beta_l)^2}[\frac{D_l}{p_l}-\{\{\frac{1-p_l}{p_l}-\frac{(1-p_l)(1-q_l)}{q_l}\}\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2\}]\right\},$$

which is the same as the equation (15).

C. Other Derivation

Simplification of: $\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2$:

Let $D_l / n_l = p_{Dl}$, then

$$\sum_{k_l=1}^{D_l}\{1-\sum_{k_l'\le k_l}\frac{1}{n_{k_l'}}\}^2$$

$$= \sum_{k_l=1}^{D_l}\{1-\sum_{k_l'=0}^{k-1}\frac{1}{n_l-k_l^{'}}\}^2$$

$$\approx \sum_{k_l=1}^{D_l}\{1-\int_0^{k_l/n_l}\frac{1}{1-x}dx\}^2$$

$$\approx n_l\int_0^{D_l/n_l}\{1-\int_0^{y}\frac{1}{1-x}dx\}^2dy$$

$$= n_l\int_0^{p_{Dl}}\{1+\log(1-y)\}^2dy$$

$$= -n_l\int_0^{\log(1-p_{Dl})}\{1+z\}^2e^z dz$$

$$= -n_l[z^2+1]e^z\ |_0^{\log(1-p_{Dl})}$$

$$= n_l\left[p_{Dl}-(1-p_{Dl})\{\log(1-p_{Dl})\}^2\right].$$