Fall 11-14-2022

Statistical Methods for Differential Gene Expression Analysis under the Case-Cohort Design

LIDONG WANG

Lidong Wang

Candidate

Mathematics & Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Yan Lu, Chairperson

Huining Kang, Advisor

Guoyi Zhang

Fletcher Christensen

Statistical Methods for Differential Gene Expression Analysis under the Case-Cohort Design

by

Lidong Wang

B.E., University of Science and Technology Beijing, 2002
M.E., Biomedical Engineering, Beijing University of Technology, 2006
M.S., Computer Science, University of New Mexico, 2013

DISSERTATION

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosphy
Statistics

The University of New Mexico

Albuquerque, New Mexico

December, 2022

Acknowledgments

I would like to extend my sincere thanks to my committee members, Drs Yan Lu (Chair), Huining Kang, Fletcher Christensen, and Guoyi Zhang, for their helpful comments and suggestions. I am particularly grateful to my mentor and advisor, Dr. Kang, for his guidance, advisement, and help.

This dissertation was supported by the UNM Comprehensive Cancer Center (UNMCCC) and the Center for Advanced Research Computing (CARC). The research topics were motivated initially by the cancer-related genomic studies I conducted with Dr. Kang at the UNMCCC. To complete this dissertation, I have conducted extensive high-performance computing using the computational resources provided by CARC.

I would also like to thank my family for giving me emotional support.

Statistical Methods for Differential Gene Expression Analysis under the Case-Cohort Design

by

Lidong Wang

B.E., University of Science and Technology Beijing, 2002

M.E., Biomedical Engineering, Beijing University of Technology, 2006

M.S., Computer Science, University of New Mexico, 2013

PhD., Statistics, University of New Mexico, 2022

Abstract

Differential gene expression analysis has the potential to discover candidate biomarkers, therapeutic targets, and gene signatures, which are critical for the prevention and treatment of diseases. Survival analyses have been used for differentially expressed genes (DEGs) identification for high-throughput gene datasets, in which genomic features (genes) are associated with survival outcomes, usually survival time of individuals. However, unbalanced samples in rare diseases generally have a high cost on data collection if using a large sample and a low power if using a small sample. How to save money when using an unaffordable sample is a practical question. The case-cohort (CCH) study design can blend the economy of case-control studies with the advantages of cohort studies. But it has not been seen in the medical research literature where high-throughput genomic data were involved. This dissertation developed statistical methods for analyzing the high-throughput gene expression data under the CCH design.

It is straightforward to use the hypothesis testing methods such as the Likelihood Ratio test, Wald test, and score test based on the Cox Proportional Hazard (PH) model to identify DEGs associated with survival outcomes given a full cohort (random sample). But in a typical genomic study, thousands of hypothesis tests must be performed simultaneously, and a score test is usually preferred. It does not need to fit the Cox PH model iteratively; hence, it can save computing time and avoid potential convergence issues. Combining the advantage of the CCH study design and score test, we developed a score test under the CCH design to identify DEGs associated with survival outcomes. We provided asymptotic distribution theory and inferential procedures for the test. We also verified the validity of the inferential procedure in finite samples through simulation studies.

Another popular approach to DEG identification is the permutation-based score test. It is a non-parametric method, and when it is used for survival outcome-related DEG analysis, the strong PH and probability distribution assumptions do not need to be a concern. One advantage of this method is that it estimates the false discovery rate (FDR) directly from the permutation procedure, which takes into account the correlation among the genomic features (genes). However, it cannot be directly applied to the data from a CCH study design because a CCH sample is not a random sample. We developed a procedure to reconstruct a full cohort from a CCH sample and then perform the permutation-based score test on the reconstructed full cohort to identify the DEGs associated with survival outcomes. To illustrate the performance of our proposed method, we evaluated our testing procedures and compared our methods with other existing approaches in terms of the FDR and the power through the simulation study and the application to the real datasets from two cancer-related genomic studies.

Contents

List of Figures

List of Figures

List of Tables

Glossary

α	The level of significance is a real number between 0 and 1.
p	The number of genes.
n	The sample size.
H_0	The null hypothesis of a statistical test.
H_α	The alternative hypothesis of a statistical test.
β	The hazard ratio parameters.
$\hat{\beta}$	The estimator of hazard ratio parameters at (partial) maximum likelihood.
β_0	The value of hazard ratio parameters at H_0.
$\tilde{\beta}$	The estimator of hazard ratio parameters at maximum pseudo-likelihood.
β_*	A value between the maximum pseudolikelihood estimator $\tilde{\beta}$ and true value β.
T_j	The true survival time of the jth patient.
c_j	The censored time of the jth patient.

t_j	The $min(T_j, C_j)$.
x_{ij}	The gene expression for gene i and patient j.
$h(t\|x_j)$	The hazard rate in the proportional hazards (PH) model.
$h_0(t)$	The baseline hazard is the hazard rate in the case when all predictors are zero.
$L_p(\beta)$	The partial likelihood function of a survival data sample.
$l_p(\beta)$	The natural logarithm of the partial likelihood function.
$\lambda(x)$	The test statistic of a (partial) maximum likelihood test.
W_T	The test statistic of a Wald test.
$J(\theta)$	The partial likelihood information matrix.
$J(\theta_0)$	The real value of the partial likelihood information matrix at H_0.
$\hat{J}(\theta_0)$	The estimated value of the partial likelihood information matrix at H_0.
$U(\theta)$	The score process of a score test is related to a survival data sample.
$U(\theta_0)$	The real score processing value of a score test at H_0.
$\hat{U}(\theta_0)$	The estimated score processing value of a score test at H_0.
$S(\theta_0)$	The real score value of a score test at H_0.
$\hat{S}(\theta_0)$	The estimated score value of a score test at H_0.
S_i	In the permutation test, the score for gene i.
$S_{obs(i)}$	The observed score for gene i without permutation.

r_i	In the permutation test, the score process for gene i.
s_i	In the permutation test, the standard deviation.
rc_i	In the CCH based permutation test, the score process for gene i.
sc_i	In the CCH based permutation test, the standard deviation.
s_0	In the permutation test, the exchangeability factor.
$\tilde{\mathcal{L}}(\beta)$	The pseudo-likelihood function of a survival data sample.
δ_j	An indicator in the pseudo-likelihood function, 0 if there is no event, and 1 if there is an event.
ω_j	The weight for individual i in the pseudo-likelihood function.
S	The subcohort of a survival data sample.
$R(t_j)$	The risk set on time t_j in the pseudo-likelihood function.
$var(.)$	The function of variance.
$\tilde{V}(\beta)$	The estimator of variance matrix prentice used.
Σ	A variance matrix related to the first derivative of the log pseudo-likelihood function.
Δ	A variance matrix related to the first derivative of the log pseudo-likelihood function.
$\tilde{\Sigma}(\tilde{\beta})$	The estimator of the variance matrix Σ at $\tilde{\beta}$, which is related to the first derivative of the log pseudo-likelihood function.
$\tilde{\triangle}(\tilde{\beta})$	The estimator of the variance matrix Δ at $\tilde{\beta}$, which is related to the first derivative of the log pseudo-likelihood function.

$\tilde{\Sigma}(\beta_0)$	The estimated value of the variance matrix Σ at H_0, which is related to the first derivative of the log pseudo-likelihood function.
$\tilde{\triangle}(\beta_0)$	The estimated value of the variance matrix Δ at H_0, which is related to the first derivative of the log pseudo-likelihood function.
χ^2_k	Central chi-squared distribution with k degrees of freedom.
λ_j	The hazard process of the Cox proportional hazards model.
λ_0	A fixed function under the proportional hazards assumption.
$Y_j(t)$	An indicator function. At time t, if the jth patient/observation is "at risk" for observable failure, $Y_j(t) = 1$. Otherwise, $Y_j(t) = 0$.
$S(t\|x)$	The survival function of the Cox proportional hazards model.
$F(t\|x)$	The distribution function of the Cox proportional hazards model.
$\mathcal{R}$	The set of true DEGs in the dataset.
$\mathcal{F}$	The sets of genes are called significant by the full cohort analyses.
$\mathcal{C}$	The sets of genes are called significant by the subcohort-based CCH analyses.
Δ_{cutoff}	The value can be chosen to control FDR multiple comparison data or type I error for single genes.
$\rightarrow_D$	Convergence in distribution.
$\rightarrow_P$	Convergence in probability.

Chapter 1

Introduction

Differential gene expression analysis is one focal point of pharmaceutical and clinical research, which identifies, among thousands of genes, those associated with certain medical conditions in a population ([13], [18], and [60]). By studying the DEGs, it is possible to discover candidate biomarkers, therapeutic targets, and gene signatures, which are critical for the prevention and treatment of diseases ([64], [22], and [22]).

Survival analyses have been used for DEG identification for high-throughput gene datasets (for example, [44], [32], and [5]), in which genomic features (genes) are associated with survival outcome, usually survival time of individuals. Identifying DEGs usually involves analyzing a high-throughput dataset, which includes p genes, n patients, and a $p \times n$ gene expression matrix (Shown in Figure 2.1). For example, there are more than 20,000 genes in human beings. In a general procedure, we perform a statistical test on each gene in the dataset separately and obtain their p-values. Then we correct the p-values, for example, with the Benjamini-Hochberg method to adjust for the multiple tests. Genes with an adjusted p-value less than a given significant level passed the significant test and are called significant genes, which would be considered the potential biomarkers of interest and undergo further

biomedical investigations.

Under the proportional hazard assumption [14], hypothesis testing methods, for example, the partial likelihood ratio test (PLRT), Wald test, and score test can be used to identify DEGs in a full cohort dataset. The null hypothesis, H_0, is that the gene under test is not a DEG, while the alternative hypothesis, H_α, is that the gene under test is a DEG. In a NONDEG, the survival time has no relationship with the gene expression value, while for a DEG, the survival time is associated with the gene expression value. In other words, we want to test whether the log of hazard ratio parameter $\beta = 0$. The PLRT is related to partial maximum likelihood estimators (PMLEs). Its test statistic needs to estimate parameter β at PMLE in two parameter spaces to get the superior of the numerator and the denominator. For Wald tests, it needs to estimate parameter β at PMLE and use the estimated value to approximate the expected partial likelihood information matrix to build its test statistic. So, for both LRT and Wald tests, we need to iteratively fit the Cox-ph model to estimate β at PMLE, which is computationally expensive, and convergence is not guaranteed. Compared to them, the score test has been widely used in genomic studies because we do not need to estimate β at (P)MLE under the null hypothesis [42].

For rare diseases ([2] [1]), the number of events is few, and the number of censored cases is enormous. This unbalanced sample usually has a high cost on data collection if using a large sample and a low power if using a small sample ([17], [8], and [46]). Pragmatically, processing a disproportionately high number of controls makes little sense. How to save money when you have to use an unaffordable sample is a practical question. The case-cohort (CCH) design [39] is an observational study design that blends the economy of case-control studies with the advantages of cohort studies ([20], and [38]). It is one of the survival analysis methods, where cases are defined as observations that had an event, and controls are right-censored observations. CCH designs consider a random sample of the full cohort, called a sub-cohort. At the time

of analysis, cases outside the sub-cohort will be added to the sample to form the CCH. The CCH design can be far more efficient than a full cohort analysis, mainly when dealing with rare outcomes. So the CCH design can save substantial time and money for studies with expected low incidence rates. However, since data of a CCH design are not a random sample, special measures have to be taken in the analysis to adjust for the sampling.

CCH analysis methods have been developed ([39] [49] [16] [6]) and three of them were compared in the context of fitting models [35]. These methods use different weighting schemes for controls and cases inside or outside of the subcohort. For example, Barlow's method appears to be the most natural approach as the weights are proportional to the subcohort size. If the subcohort is 10% of the full cohort, Barlow's method weighs controls in the subcohort as if they are worth 10 people. Since all cases are included in this design, cases outside the subcohort are given a weight of 1. Subcohort cases are treated in two ways. Before their events, they are weighted with a factor of 10 (just like the controls), but at the time of their event, they are treated like the cases outside the subcohort, with the weight of 1 individual.

Improving DEGs' identification under CCH has medical and biological values. The "Score process" of CCH was proved with asymptotic normality [49], and the related Wald test was proposed based on the distribution. To find out a good estimation of parameters, it needs to iteratively fit the Cox-ph model, and convergence is not guaranteed. One alternative method is to use a score test. To prove this procedure is theoretically valid. We need to estimate the covariance matrix of CCH asymptotic chi-square distribution at $\beta_0 = 0$. Furthermore, we need to derive the first derivative of the log Pseudo-likelihood Function and acquire the value of "Score Process". Then, we need to find a test statistic with certain distribution to calculate the p-value under the null hypothesis, $\beta = 0$.

In multiple testing comparisons, as the number of genes or features, p is large,

it usually brings low power for the whole procedure. Hence, we often use the false discovery rate (FDR) proposed by Benjamini and Hochberg rather than the family-wise error rate (FWER) to adjust for the multiple comparisons. A caveat is that the BH method assumes that all the test are independent while the expressions of many genes are correlated with each other. A permutation-based approach has been proposed to estimate the FDR directly through the permutation procedure that takes into account of the correlation among the genes. Proposed by Fisher [20] and Pitman [38], this method has been applied on biostatistics and quantized data analysis ([59] and [11]). A permutation test adopts a non-parametric statistic that obtains the p-value from the sample-specific permutation distribution of that statistic rather than from the theoretical distribution derived from the parametric assumption. Permutation procedures were applied to estimate false discovery rate (FDR) ([58] and [56]). When permutation tests are used on DEGs' identification, strong semi-parametric assumption and probability distribution assumption for p-value do not need to be concerned. Under the null hypothesis of genes (they are not DEGs), their FDR is obtained by calculating the average of a large number of possible FDR values calculated from related rearrangements of the observed data. However, the basic assumption of a permutation test is the "exchangeability" ([20] and [38]), which requires a survival dataset to be a random sample. Therefore, the permutation test can not be directly applied to a CCH sample because the CCH sample is not random. We want to find a new procedure to identify DEGs, which can save the benefits of both the CCH and permutation tests. We propose to reconstruct a whole cohort based on the given CCH and use the entire reconstructed cohort to do a permutation test. As the controls outside the CCH sample are missing at random and patients in the subcohort are a random sample, we used the expression values in the subcohort to impute the missing expression values of controls outside of the CCH sample.

This dissertation is organized as follows. In Chapter 2, we gave a background

review of tests and methods for DEG identification. In Chapter 3, we proposed a CCH-based score test, in which $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ were estimated at $\beta_0 = 0$ under null hypothesis, rather than at $\tilde{\beta}$. We built the test statistic of the test, which asymptotically follows a Chi-square distribution. Furthermore, we conducted simulation studies to evaluate the Type-I error and power for a single gene and to evaluate the false discovery rate (FDR) and power for high-throughput data. Besides, we applied the CCH-based score test to a real dataset to measure the consistency between the proposed CCH-based test and the full cohort analysis method. In Chapter 4, we proposed a CCH-based permutation test. We rebuilt the whole cohorts by imputing the missing data with resampling with replacement and used them to identify DEGs by permutation tests. We performed simulation studies to evaluate our methods and to compare our methods with some existing test procedures. We also used real data to compare the performance consistency of full-cohort and reconstructed full cohort to illustrate our methods. Finally, we gave conclusions and further research works in Chapter 5.

Chapter 2

Background

Recently there has been an increased focus on precision medicine ([23] [62] [33]), with researchers seeking to discover biomarkers that can inform if an individual is more or less likely to suffer harmful health outcomes or whether they will be receptive to treatment ([34] [29] [57]). High-throughput gene expression profiles have allowed the discovery of potential biomarkers ([65] [61] [36]).

Differential gene expression analysis can determine which genes are expressed at different levels between conditions by performing statistical analysis. Through DEG analysis, we can understand the biological differences between healthy and diseased states. For example, we can find upregulated genes and downregulated genes between events and censored cases and use them to discover potential biomarkers.

Parametric and Non-parametric DEG identification methods have been proposed for studying the changes in gene or transcript expressions under different conditions (e.g. control vs infected). DESeq [3], DESeq2 [32] and edgeR [44], use a negative binomial distribution to model RNA-Seq read counts for assessing differential expression. Limma [53] uses linear models based on the empirical Bayes method to identify DEGs. Fisher's exact test is non-parametric in the analysis of contingency tables,

which does not assume data across samples are based on the theoretical probability distribution.

2.1 DEG Analysis on Survival Data

DEG analysis can be applied to survival data. In survival data, a full cohort is usually a random sample from the population. As shown in Figure 2.1, the data structure includes three tables, clinical covariate, gene expression matrix, and annotation of genomic features. Clinical covariate has sample annotations, such as survival information for each patient. For patient j, time t_j is the length of time until the occurrence of an event of interest. However, we may not observe the time completely. That is, some patients may be censored. $\delta_j = 0$ means censored, and $\delta_j = 1$ means that the patient have experienced the event of interest. In our study, we will assume that censoring is non-informative [41]. That is, censoring should not be related to the probability of an event occurring.

The gene expression matrix has gene expression for all n patients and p genes (Shown in figure 2.1). It is a $p \times n$ matrix. We use x_{ij} to denote the gene expression related with ith gene and jth patient. So, the observed data in the full cohort/sample is (t_j, δ_j, x_{ij}) for patient j, where $t_j = min(T_j, C_j)$. T_j is the true survival time and C_j is the censored time. The table of Annotation of genomic features has Gene ID, Gene Symbol, Gene descriptions, etc.

When studying the DEGs in survival data, we want to know whether gene expression is associated with survival times. More specifically, we want to identify whether the survival times either decrease or increase with the increase of gene expression. For a gene, when survival times decrease with increasing its expression, we call it down-graded. Reversely, we call it upgraded. Both down-graded and upgraded genes are DEGs.

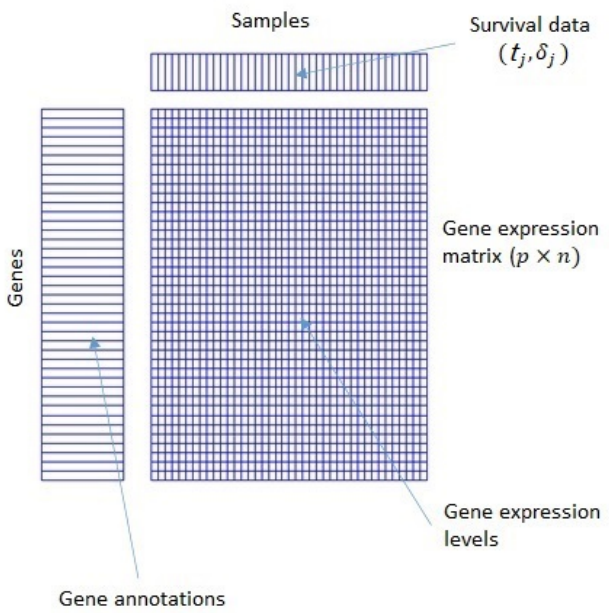


Figure 2.1: Dataset for a study of gene expressions associated with survival outcome.

2.1.1 Proportional Hazards Model

The proportional hazards (PH) model, proposed by Cox [14], is one of the most popular survival models in survival analysis. The hazard function is

$$h(t|x_j) = h_0(t)exp(\beta^T x_j), \tag{2.1}$$

where $x_j = (x_{1j}, x_{2j}, ..., x_{pj})^T$ is a vector of predictors of the jth sample unit, and $\beta^T = (\beta_1, \beta_2, ..., \beta_p)$ is a vector of unknown coefficients that we want to estimate. The factor $h_0(t)$ is called the baseline hazard, which is the hazard rate in the case when all predictors are zero. It requires no particular form for survival time. In other words, the baseline hazard is unspecified. In medical research, x_j usually consists of clinical and demographical covariates including the gene expression values.

For any two sets of predictors, x_i and x_{i*}, the hazard ratio (HR) is constant over

time.

$$h(t|x_j)/h(t|x_{j*}) = h_0(t)exp(\beta^T x_j)/h_0(t)exp(\beta^T x_{j*}) = exp(\beta^T x_j)/exp(\beta^T x_{j*}). \tag{2.2}$$

It is why the model is called proportional hazard. Non-informative censoring and proportional hazard are the critical assumptions in the Cox model.

Suppose there is a full cohort with size n, which is a random sample with t_j, δ_j, and x_{ij} ($j = 1, 2, ... n$), and their assumed probability distributions depend on some unknown parameter $\beta = (\beta_1, \beta_2, ...\beta_p)$. Then the partial likelihood function of the sample is called $L_p(\beta)$.

$$L_p(\beta) = \prod_{j=1}^{m} \frac{exp(\beta^T x_j)}{\sum_{i \in R_k} exp(\beta^T x_i)}, \tag{2.3}$$

where m is the number of unique event times, and R_k is the set of subjects at risk just before time t_j. Suppose there is exactly one event at each event time. The maximum likelihood estimation of the unknown parameter β would be the value that maximizes the likelihood function $L_p(\beta)$ based on the data we observed. If the likelihood function is differentiable in β_i, possible candidates for the MLE are the values of $(\beta_1, \beta_2, ...\beta_p)$ that solve

$$\frac{\partial}{\partial \beta_i} L_p(\beta) = 0. \tag{2.4}$$

To simplify the calculation, we usually use the natural logarithm of the likelihood function, which is below.

$$l_p(\beta) = log(L_p(\beta)) = log(\prod_{j=1}^{m} \frac{exp(\beta^T x_j)}{\sum_{i \in R_k} exp(\beta^T x_i)}). \tag{2.5}$$

The counting process formulation replaces the pair of variables (T_j, C_j) with the pair of functions $(N_j(t), Y_j(t))$, where $N_j(t)$ is the number of observed events in $[0, t]$

for individual j and $Y_j(t)$ is an indicator, 1 if individual j is under observation and at risk at time t, 0 otherwise. The two symbols system are the same in nature.

The partial likelihood function described by the counting process is below.

$$L_p(\beta) = \prod_{j=1}^{n} \prod_{t\geq 0} \left(\frac{Y_j(t) r_j(\beta,t)}{\sum_k Y_k(t) r_k(\beta,t)}\right)^{dN_j(t)}, \tag{2.6}$$

where $r_j(\beta, t)$ is the risk score for subject j with $r_j(\beta, t) = exp[X_j(t)\beta]$, and $dN_j(t)$ is a shorthand that allows mixed continuous and discrete processes to be handled by a single notation. As counting processes are purely jumping processes, $dN_j(t)$ is the number of events occurring precisely at t for subject j.

The log partial likelihood can be written as below.

$$l_p(\beta) = \sum_{j=1}^{n} \int_0^{\infty} [Y_j(t) X_j(t)\beta - log(\sum_k Y_k(t) r_k(\beta, t))] dN_j(t). \tag{2.7}$$

As the log operator would not change the maximum of $L_p(\beta)$, we can use $l_p(\beta)$ to find the maximum likelihood estimation of β. If the log-likelihood function is differentiable in β_i, possible candidates for the MLE are the values of $(\beta_1, \beta_2, ...\beta_p)$ that solve

$$\frac{\partial}{\partial \beta_i} l_p(\beta) = 0. \tag{2.8}$$

The maximum likelihood estimator is asymptotically normal. It follows

$$J(\beta_0)^{1/2}(\hat{\beta} - \beta_0) \rightarrow_D N(0, I), \tag{2.9}$$

where $J(\beta)$ is defined as the partial likelihood information. For one covariate case, it is easy to present as $J(\beta) = -\frac{\partial^2}{\partial \beta^2} l_p(\beta)$. And the extension is straightforward. Cox [15] and others ([43] [19] [27] [45]) have shown that this partial likelihood can be treated as an ordinary likelihood to derive valid (partial) MLEs of β. We introduce below the concepts of four hypothesis methods, the partial likelihood ratio test, Wald test, score test, and permutation test, which have been used on the partial likelihood of the full cohort.

2.1.2 Partial Likelihood Ratio Tests

The partial likelihood ratio test (PLRT) of hypothesis testing is related to the aforementioned partial maximum likelihood estimators (PMLEs), and it is performed by estimating two models and comparing the fit of one model to the fit of the other based on the ratio of their partial likelihoods. Suppose we have a statistical model with parameter space B. Consider a hypothesis testing problem in which the null and the alternative hypotheses are $H_0 : \beta \in B_0$ and $H_\alpha : \beta \in B_0^c$, where $B_0 \cup B_0^c = B$. The definition of the likelihood ratio test statistic for hypothesis testing is

$$\lambda(x) = \frac{sup_{B_0} L_p(\beta)}{sup_B L_p(\beta)}. \tag{2.10}$$

For the denominator, we can think of doing the maximization over the entire parameter space B, and for the numerator, we can think of doing the maximization of a subset of B. If the PMLE of the former is $\hat{\beta}$ and the PMLE of the latter is $\hat{\beta}_0$, we can write the form of the likelihood ratio test statistic as

$$\lambda(x) = \frac{L_p(\hat{\beta}_0)}{L_p(\hat{\beta})}. \tag{2.11}$$

For computational simplicity, the partial likelihood-ratio test statistic is expressed as a difference between the log-likelihoods.

$$\lambda_{LR}(x) = -2ln[\frac{L_p(\hat{\beta}_0)}{L_p(\hat{\beta})}] = -2[l_p(\beta_0) - l_p(\beta)], \tag{2.12}$$

where $\lambda_{LR}(x)$ converges asymptotically to a chi-square distribution with degrees of freedom equal to the difference in the dimensionality of β and β_0.

Let $\lambda_{LR}(x)_{obs}$ denote the observed value of the static calculated from the data, and the p-value is

$$p - value = P(\lambda_{LR}(x) \geq \lambda_{LR}(x)_{obs} | H_0). \tag{2.13}$$

2.1.3 Wald Test

Wald test (Wald, 1943) tests the significance of particular explanatory variables in a statistical model. Consider a hypothesis testing problem in which the null and the alternative hypotheses are $H_0 : \beta = \beta_0$ and $H_\alpha : \beta \neq \beta_0$. The definition of the Wald tests statistic for hypothesis testing is

$$W_T = J(\hat{\beta})[\hat{\beta} - \beta_0]^2, \tag{2.14}$$

where $J(\beta)$ is the partial likelihood information matrix. We need to estimate β at MLE to estimate it. ($J(\hat{\beta})$ is $J(\beta)$'s estimation at the MLE). The test statistics measure the weighted distance between the unrestricted estimate and its hypothesized value under the null hypothesis, where the weight is expected the partial likelihood information matrix or the inverse of the variance of the estimate. Compared with the likelihood-ratio test, it only requires estimating the unrestricted model, which can decrease the computational cost. However, as derived from a Taylor expansion, the Wald test is not invariant to equivalent but different nonlinear expressions of the null hypothesis, as they may lead to nontrivial differences in the corresponding Taylor coefficients.

The Wald test can test the significance of multiple variables simultaneously, which has been implemented in the procedure corresponds to backward elimination in multiple regression. In the method, residuals are computed for the current model. Then the least significant parameter is removed from the model. Repeat to calculate residuals and remove the least significant parameter until a set is obtained that they are sufficiently unimportant to be eliminated.

$$W_T = \frac{[\hat{\beta}-\beta_0]^2}{Var(\hat{\beta})}. \tag{2.15}$$

The formula for the test statistic of a single parameter β is as above, where W_T follows an asymptotic χ_1^2 distribution and $Var(\hat{\beta})$ can be estimated from data. Let

W_{obs} denote the observed value of the static calculated from the data, and the p-value is

$$p - value = P(W_T \geq W_{obs}|H_0). \tag{2.16}$$

2.1.4 Score Test

The score test of hypothesis testing is related to the partial likelihood function. A score is the gradient of the likelihood function, evaluated at the hypothesized parameter value under the null hypothesis and used to assess constraints on statistical parameters [42]. Consider a hypothesis testing problem in which the null and the alternative hypotheses are $H_0 : \beta = \beta_0$ and $H_\alpha : \beta \neq \beta_0$. Let L_p be the partial likelihood function which depends on a univariate parameter β, and x be the data. The score $U(\beta)$ is defined as

$$U(\beta) = \frac{\partial}{\partial\beta} log L_p(\beta). \tag{2.17}$$

The partial likelihood information is

$$J(\beta) = -\frac{\partial^2}{\partial\beta^2} l_p(\beta). \tag{2.18}$$

The definition of the score test statistic for the hypothesis testing is below, which converges in distribution to a Chi-square distribution.

$$S(\beta_0) = U^T(\beta_0) J^{-1}(\beta_0) U(\beta_0) \sim {\chi^2}_k. \tag{2.19}$$

Under H_0, $U(\beta_0)$ and $J(\beta_0)$ are estimable if they are unknown.

$$\hat{S}(\beta_0) = \hat{U}^T(\beta_0) \hat{J}^{-1}(\beta_0) \hat{U}(\beta_0). \tag{2.20}$$

Let $\hat{S}(\beta_0)_{obs}$ denote the observed value of the static calculated from the data, and the p-value is

$$p - value = P(\hat{S}(\beta_0) \geq \hat{S}(\beta_0)_{obs}|H_0). \tag{2.21}$$

The score test only requires the computation of the restricted estimator under H_0. It checks whether the data violate a restriction on a model estimated by maximum likelihood. If the restricted estimator is near the maximum of the likelihood function, the score should not differ from zero by more than a sampling error. It is the main advantage of the score test over the Wald test and the likelihood-ratio test.

After discussing the three tests, we can see that the Score test fits just the restricted model, the Wald test fits just the unrestricted model, and the Likelihood Ratio test fits both models. Both the Wald and the Lagrange multiplier (or score) tests are asymptotically equivalent to the LR test in large sample sizes and, in finite samples, the three will generally come to the same conclusion, although their test statistics is maybe somewhat different. The advantage of the Wald and Lagrange multiplier (or score) tests is that they approximate the LR test but require only one model to be estimated.

2.1.5 Permutation Test

The basic assumption of a permutation test is the "exchangeability" ([20] and [38]). A permutation test calculates all possible values of the test statistic under possible rearrangements of the observed data. Under the hypothesis, observations are exchangeable if they are independent, identically distributed (i.i.d.), or jointly normal with equal covariances. If not, the joint distribution of the observations is variant after permutation. For example, the joint distribution of a set of normally distributed random variables is invariant under permutations of the variable subscripts if, in its covariance matrix, all diagonal elements have the same value σ^2 and all the off-diagonal elements have the same value.

Preserving transforms, asymptotic exchangeability, partial exchangeability, and weak exchangeability have been studied, which enlarged the application of the per-

mutation test. A set of observations (random variables) X will be said to be transformable exchangeable if there exists a transformation (measurable transformation) T, such that TX is exchangeable [12]. A set of random variables is weakly exchangeable if their joint distribution is invariant for a subset of permutations. For a sequence of discrete random variables that represent the outcomes of a finite Markov Chain, if the transition matrix of the Markov Chain is such that $P_{ij} = P_{ji}$ for all i and j, the sequence is of variables is partially exchangeable [63].

Tusher, Tibshirani, and Chu proposed "Significance Analysis of Microarrays" (SAM), which is a statistical technique for finding significant genes in a set of microarray experiments [58]. It uses permutation tests to identify differentially expressed genes (the expression of any genes is significantly related to the response). The response variable may be a grouping like untreated, treated (either unpaired or paired), a multiclass grouping (like breast cancer, lymphoma, colon cancer), a quantitative variable (like blood pressure), or a possibly censored survival time. Besides, SAM can be used on the studies of identifying exonic splicing enhancers, genetic dissection of transcriptional regulation, and finding binding sites of transcriptional regulators.

For simple linear regression questions, suppose there are n pairs of observations of a random variable Y with fixed values of a variable X. We want to test the null hypothesis of no (linear) relationship between Y and X. For example, the linear model of $Y = \mu + \beta X + \epsilon$, the null hypothesis is that the slope $\beta = 0$. If the null hypothesis is true, the n observations of Y could have been observed in any order with respect to the n fixed values of X. In other words, there are $n!$ unique possible permutations. The only assumption of a related test is that the observations Y are exchangeable under a null hypothesis [4].

Censored survival data have the form $(t_j, \delta_j | x_{ij})$, where i is the index for genes and j is the index for individuals. The response (time, status) pair, like (10,1) or

(20,0), is conditional on the gene expression. The first number is survival time, and the second is status (1=died, 0=censored). SAM considers that each pair (t_j, δ_j) is a responsible variable and each x_{ij} is an explanatory variable [24]. Possible ties in the survival times are handled by Breslow's method [9]. Under null hypothesis (the gene understudying is a NONDEG), the gene expression and the (t_j, δ_j) pairs should have no relationship. For each gene i, SAM assumes the pairs (t_j, δ_j) are exchangeable under the null hypothesis. Cox model uses partial likelihood, which involves only the ranks of the survival times, making the model semiparametric. SAM uses its score statistic on the permutation test. Under the null hypothesis (the gene understudying is a NONDEG), the gene expression and the (t_j, δ_j) pairs should have no relationship, which assures the Type-I error is correct. Under the alternative hypothesis (the gene understudying is a DEG), gene expression should associated with the (t_j, δ_j) pairs, assuring power.

SAM uses a modified score test statistic for a permutation test on a full cohort, in which each gene will be scored [10]. The definition of the score S_i for gene i is:

$$S_i = \frac{r_i}{s_i + s_0}, \tag{2.22}$$

where r_i is a score, s_i is a standard deviation and s_0 is an exchangeability factor to put the penalty to the genes with overall low expression values. For censored survival data, r_i is defined as

$$r_i = \sum_{k=1}^{K} (x_{ik}^* - d_k \bar{x}_{ik}). \tag{2.23}$$

And s_i is defined as

$$s_i = [\sum_{k=1}^{K} ((\frac{d_k}{m_k}) \sum_{j \in R_k} (x_{ij} - \bar{x}_{ik}))^2]^{1/2}, \tag{2.24}$$

where x_{ij} is the expression value of gene i for patient j. k be the indices of the K unique death times z_1, z_2, ... z_K, and R_1, R_2, ... R_K be the indices of the observations at risk at these unique death times, , that is $R_k = \{i : t_i \geq z_k\}$. Let $m_k = \#in R_k$. Let d_k be the number of deaths at time z_k. $x_{ik}^* = \sum_{t_j = z_k} x_{ij}$ and $\bar{x}_{ik} = \sum_{j \in R_k} \frac{x_{ij}}{m_k}$.

In SAM, FDR is defined as:

$$FDR \equiv E(\frac{V}{R}) \equiv \frac{\frac{1}{N}\sum_{j=1}^{N} \#(S^{(j)}_{perm(i)} > \triangle)}{\#(S_{obs(i)} > \triangle)}, \tag{2.25}$$

where R is the number of genes that are called a DEG and V is the number of genes that are incorrectly called a DEG. The decision rule for ith gene to be a DEG: if $S_{(i)} > \triangle$, it is; otherwise, it is not. The number of DEGs is related to the cutoff Δ. The denominator is the number of DEGs found with non-permuted data and related scores. In the project, it is the number of DEGs found from a full cohort or a full reconstructed cohort without permutation. The numerator is the average number of DEGs from N time permutations. For example, from the first permutation, we find $a1$ DEGs. And from the second permutation, we find $a2$ DEGs. Then the average number of DEGs from 2-time permutations is $(a1 + a2)/2$. When the number of permutations, N, is large enough, $\frac{\frac{1}{N}\sum_{j=1}^{N} \#(S^{(j)}_{perm(i)} > \triangle)}{\#(S_{obs(i)} > \triangle)}$ converges to $E(\frac{V}{R})$ by large number theory.

Observed	N Permutations			
$S_{obs(1)}$	$S^{(1)}_{perm(1)}$	$S^{(2)}_{perm(1)}$	$\cdots$	$S^{(N)}_{perm(1)}$
$S_{obs(2)}$	$S^{(1)}_{perm(2)}$	$S^{(2)}_{perm(2)}$	$\cdots$	$S^{(N)}_{perm(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_{obs(p)}$	$S^{(1)}_{perm(p)}$	$S^{(2)}_{perm(p)}$	$\cdots$	$S^{(N)}_{perm(p)}$
$\#(S_{obs(i)} \geq \Delta)$	$\#(S^{(1)}_{perm(i)} \geq \Delta)$		$\cdots$	$\#(S^{(N)}_{perm(i)} \geq \Delta)$

Figure 2.2: Plot of permutation test.

After permutations, valid Δ_{cutoff} can be chosen to control FDR for high-throughput data or type I error for single genes. For example, if we want FDR to be less than 0.05, we can solve related Δ_{cutoff} from the function's reverse function (4.1).

We can calculate $S_{obs(i)}$ with score function [10] for each gene in a full cohort or rebuild cohort. Using Δ_{cutoff} as the cutoff, the decision rule for ith gene to be a

DEG: if $S_{(i)} > \Delta_{cutoff}$, it is; otherwise, it is not.

When we apply SAM to a whole cohort, we need to choose the value of a tuning parameter "delta," which is related to the significant level. For example, there is a list of delta in Table 2.1. "M-FP" is the median false positive. "90P-FP" is 90th percent false positive. "M-FDR" is the median FDR. "90P-FDR" is the 90th percentage FDR. If we want to use the delta related to "M-FDR" 0.05, we should choose 0.7206676641. Then we use it to identify the list of DEGs.

Figure 2 is a standard output from the manual of SAM [10], where genes have positive scores or negative scores. The positive scores are related to up-regulated genes, and the negative scores are related to down-regulated genes. Most of the genes are shown linearly in the middle of the plot and are not DEGs, and DEGs are labeled at the left and right tails. Positive significant genes are DEGs in the up-regulated gene (labeled in red), and significant negative genes are DEGs in the down-regulated gene (labeled in green).

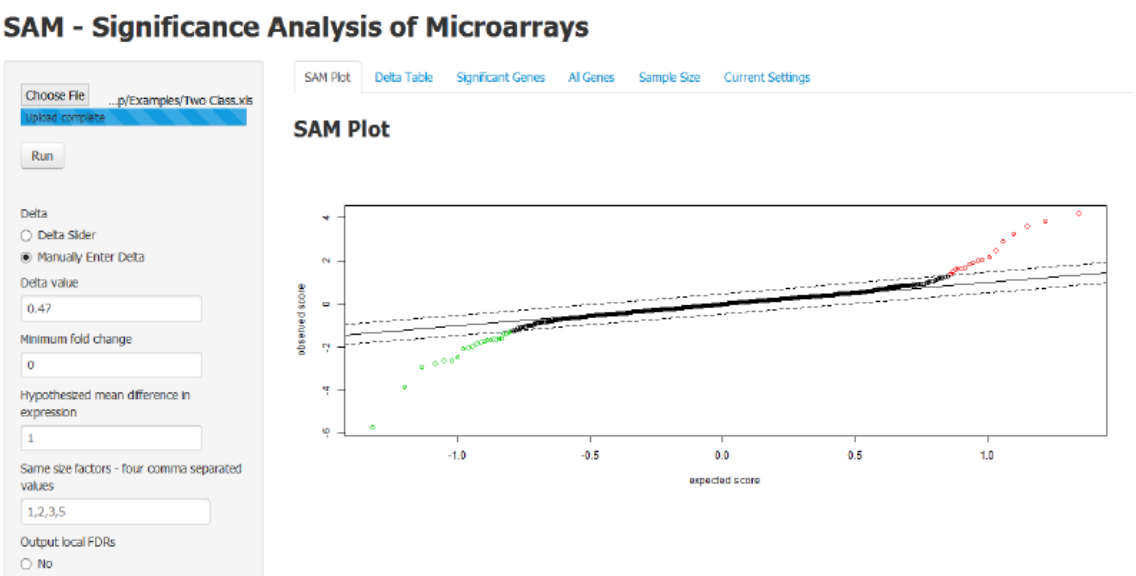


Figure 2: SAM result

Table 2.1: Delta Table of 300 Permutation Test with package SAMR.

	delta	M-FP	90P-FP	called	M-FDR	90P-FDR	cutlo	cuthi
1	0.0000000000	2.04e+04	2.33e+04	43493	0.47	0.54	-4.29e-01	0.33
2	0.0008569176	2.04e+04	2.33e+04	43469	0.47	0.54	-4.31e-01	0.33
3	0.0034276702	2.03e+04	2.33e+04	43404	0.47	0.54	-4.36e-01	0.33
4	0.0077122580	2.02e+04	2.32e+04	43331	0.47	0.54	-4.43e-01	0.33
5	0.0137106809	2.01e+04	2.31e+04	43204	0.46	0.53	-4.55e-01	0.33
6	0.0214229389	1.99e+04	2.29e+04	43005	0.46	0.53	-4.71e-01	0.33
7	0.0308490320	1.96e+04	2.27e+04	42785	0.46	0.53	-4.90e-01	0.33
8	0.0419889602	1.93e+04	2.24e+04	42521	0.45	0.53	-5.13e-01	0.33
9	0.0548427235	1.89e+04	2.20e+04	42195	0.45	0.52	-5.40e-01	0.33
10	0.0694103220	1.85e+04	2.171e+04	41892	0.44	0.52	-5.68e-01	0.33
11	0.0856917555	1.81e+04	2.13e+04	41526	0.44	0.51	-6.01e-01	0.33
12	0.1036870242	1.77e+04	2.10e+04	41170	0.43	0.51	-6.35e-01	0.33
13	0.1233961280	1.72e+04	2.06e+04	40713	0.42	0.51	-6.76e-01	0.33
14	0.1448190669	1.67e+04	2.03e+04	40277	0.42	0.50	-7.18e-01	0.33
15	0.1679558409	1.62e+04	1.98e+04	39826	0.41	0.50	-7.63e-01	0.33
16	0.1928064500	1.56e+04	1.92e+04	39226	0.40	0.49	-8.17e-01	0.33
17	0.2193708942	1.51e+04	1.88e+04	38758	0.39	0.49	-8.68e-01	0.33
18	0.2476491735	1.47e+04	1.84e+04	38184	0.38	0.48	-9.26e-01	0.33
19	0.2776412879	1.42e+04	1.79e+04	37596	0.38	0.48	-9.87e-01	0.33
20	0.3093472375	1.37e+04	1.76e+04	37047	0.37	0.48	-1.05e+00	0.33
21	0.3427670221	1.27e+04	1.66e+04	35745	0.36	0.46	-1.12e+00	0.37
22	0.3779006419	1.08e+04	1.48e+04	33265	0.32	0.45	-1.19e+00	0.47
23	0.4147480968	9.08e+03	1.294e+04	30921	0.29	0.42	-1.26e+00	0.58
24	0.4533093868	7.11e+03	1.08e+04	27709	0.26	0.39	-1.34e+00	0.72
25	0.4935845119	5.49e+03	8.82e+03	24619	0.22	0.36	-1.44e+00	0.85
26	0.5355734721	3.81e+03	6.66e+03	20719	0.18	0.32	-1.53e+00	1.03
27	0.5792762674	2.55e+03	4.85e+03	17124	0.15	0.28	-1.62e+00	1.22
28	0.6246928979	1.43e+03	3.11e+03	12877	0.11	0.24	-1.72e+00	1.47
29	0.6718233634	6.56e+02	1.75e+03	8680	0.076	0.20	-1.82e+00	1.80
30	0.7206676641	2.26e+02	7.30e+02	4485	0.050	0.16	-1.93e+00	2.43
31	0.7712257998	1.24e+02	4.56e+02	3360	0.037	0.14	-2.06e+00	2.73
32	0.8234977707	7.34e+01	3.00e+02	2698	0.027	0.11	-2.17e+00	2.95
33	0.8774835767	4.12e+01	2.02e+02	2167	0.019	0.093	-2.30e+00	3.11
34	0.9331832178	1.97e+01	1.13e+02	1549	0.013	0.073	-2.47e+00	3.24
35	0.9905966940	1.01e+01	6.57e+01	1124	0.0090	0.058	-2.63e+00	3.39
36	1.0497240053	5.37e+00	3.65e+01	840	0.0064	0.043	-2.79e+00	3.49
...	...	...	...	...	...	...	...	...
48	1.8929308798	0.00e+00	0.00e+00	5	0.00	0.00	-1.00e+10	5.20
49	1.9743380476	0.00e+00	0.00e+00	5	0.00	0.00	-1.00e+10	5.20
50	2.0574590504	0.00e+00	0.00e+00	2	0.00	0.00	-1.00e+10	5.49

2.2 Case Cohort Study Design

Case-Cohort study design is a kind of advanced Case-Control design, which keeps the properties of both Case-Control Study and Cohort Study ([30] [26]). Case-control study design identifies cases and controls and looks back to see if these cases' characteristics differ from controls [51]. Cases are individuals who develop the disease or outcome, and controls are individuals without the disease and outcome. The measure of association for a case-control study is typically an odds ratio. The case-control approach allows for the study of rare diseases [21]. They are efficient for rare diseases or diseases with a long latency period between exposure and disease manifestation, and they provide a much cheaper and quicker study of risk factors.

Cohort studies are longitudinal studies that begin with persons who do not have but are at risk of developing a disease or outcome ([54] [50]). The studies must follow the individuals over a period (usually several years). During follow-up, some people in the cohort will be exposed to a specific risk factor or characteristic until some develop a disease. We then explore the impact of this factor or feature. The measure of association is a relative risk, attributable risk, or depicted with survival analysis. For example, the British Doctors Study identifies the link between smoking and lung cancer. A cohort study may not be economical when the cohort size is extremely large and the disease event rate is low, especially for rare diseases.

The case-cohort (CCH) study design is a prospective observational study design that blends the economy of case-control studies with the philosophical soundness of cohort studies because not all members of the parent cohort require diagnostic testing ([52] [28]). For example, when dealing with rare outcomes or diseases, there is a disproportionately high number of controls. Similar to the idea of case-control study designs, it makes little sense to process all controls. In a CCH, controls are randomly selected from the parent cohort, forming a subcohort (a random sample

of the full (parent) cohort). At the time of analysis, all cases outside the subcohort will be added to the sample. In other words, a CCH sample consists of all cases of the full cohort, but only the controls in the subcohort. So CCH design can save time and money on the controls out of subcohort. CCH design will also reduce selection bias, as cases and controls are sampled from the same population. A CCH analysis is best suited to data that is cheap to collect but expensive to analyze or process. Taking a blood or tissue sample is relatively quick and easy, but fully genotyping an individual from such a sample requires considerably greater resources.

Building a likelihood function is indispensable to identify DEGs in data of a CCH study design. Similar to the partial likelihood function of a full cohort (equation 2.3), a weighted likelihood was proposed to account for the sampling scheme in a CCH study (equation 2.26) [39]. For survival time 1 to n, δ_j is 0 if there is no event and one if there is an event. ω_j is the weight for individual i, and the summation in the denominator only includes individuals at risk who are also in the subcohort (S is the subcohort and $R(t_j)$ is the risk set on time t_j). Individual j, however, can either be a case from inside or outside of the subcohort. x_j is its gene expression. Based on this weighted likelihood, methods to do significance testing and interval estimation for $\hat{\beta}$ were proposed.

$$\tilde{\mathcal{L}}(\beta) = \prod_{j=1}^{n} \left\{ \frac{e^{x_j\beta}}{\omega_j e^{x_j\beta} + \sum\limits_{\substack{i \neq j \\ i \in R(t_j) \cap S}} \omega_i e^{x_i\beta}} \right\}^{\delta_j} . \tag{2.26}$$

Under Case-Cohort study design (CCH design), we want to propose novel hypothesis testing methods for the identification of differentially expressed genes for survival data. This chapter will review related fundamental theories and their applications in identifying DEGs.

2.2.1 Prentice's CCH method

Epidemiologic cohort studies and disease prevention trials can be costly because they typically require the follow-up of several thousand subjects for many years. Synthetic case-control designs had been used to reduce the number of subjects in those studies and trials. However, some issues remain. Prentice proposed a case-cohort design to address the problems in these case-control designs [39]. For example, since the control may serve as a member of the comparison group for different cases, it is inefficient to align each selected control subject to its matched case in a synthetic case-control design. CCH selects a random sample, or a stratified random sample, of the entire cohort, constituting the comparison set of cases in the range of failure times.

For binary response, the maximum likelihood estimate of the odds ratio is

$$\hat{\lambda} = d_1(n_0 - d_0)d_0(n_1 - d_1)^{-1}, \tag{2.27}$$

where d_0 and d_1 are the number of failures, and n_0 and n_1 are the number of subjects, corresponding to the dependence of failure probability on the presence, $z = 1$, or absence, $z = 0$, respectively.

For time to response data, the relative risk parameter β can be estimated using case-cohort data, considering maximizing the function 2.26. The maximum pseudo-likelihood estimator $\hat{\beta}$ is defined by $U(\hat{\beta}) = 0$, where

$$U(\beta) = \frac{\partial log\tilde{L}(\beta)}{\partial \beta}, \tag{2.28}$$

and

$$var\{U(\beta)\} = \sum_{j=1}^{n}[var\{U_j(\beta)\} + 2\sum_{k|t_k<t_j} cov\{U_k(\beta), U_j(\beta)\}]. \tag{2.29}$$

Besides, for β_* between the maximum pseudolikelihood estimator $\hat{\beta}$ and true

value β, the Taylor expansion about the true β evaluated at $\hat{\beta}$ gives

$$n^{-1/2}U(\beta) = n^{-1}I(\beta_*)n^{1/2}(\hat{\beta} - \beta). \tag{2.30}$$

With equation 2.30, a hypothesis wald test was built, in which we need to iteratively fit PH model to get $\hat{\beta}$. And we need to compute estimators of the variance of the prentice score process because the actual values of $var\{U(\beta)\}$ can not be computed and need to be estimated. The estimator of variance matrix prentice used is below.

$$\tilde{V}(\beta) = \sum_{j=1}^{n} \delta_j \{v_{jj} + 2\Delta(t_j) \sum_{k|t_k<t_j} \delta_k v_{kj}\}. \tag{2.31}$$

2.2.2 Self-Prentice's CCH method

In Prentice's CCH method, the martingale method was applied. When certain generating $\sigma - algebras$ were not nested, the martingale convergence results were not sufficient. In self and Prentice [49], they developed another asymptotic distribution theory for the case-cohort maximum pseudolikelihood estimator, $\tilde{\beta}$, and related quantities using a combination of martingale and finite population convergence results.

$$\tilde{\beta} \rightarrow_p \beta_0. \tag{2.32}$$

$$n^{-1/2}\tilde{U}(\beta_0, 1) \rightarrow_D N(0, \Sigma + \Delta). \tag{2.33}$$

$\tilde{\beta}$ is a point estimation of β_0. Consistency of $\tilde{\beta}$ to β_0 and asymptotic normality of the score statistic $\tilde{U}(\beta_0, 1)$ were proved. With equations 2.32 and 2.33, asymptotic normality of $\tilde{\beta}$ was also proved. Conditions were listed to ensure the asymptotic distribution theories.

$$n^{1/2}(\tilde{\beta} - \beta_0) \rightarrow_D N(0, \Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}). \tag{2.34}$$

With equation 2.34, a hypothesis Wald test was built, in which we need to iteratively fit the PH model to get $\tilde{\beta}$. And we need to compute estimators of Σ and Δ, because the real values of Σ and Δ can not be computed and need to be estimated. The estimators they used are in equations 2.35 and 2.36.

$$\tilde{\Sigma}(\tilde{\beta}) = \frac{1}{n}\int_0^1 \tilde{V}(\tilde{\beta}, t)\, d\bar{N}(t). \tag{2.35}$$

$$\tilde{\triangle}(\tilde{\beta}) = \frac{1}{n^2}\int_0^1 \int_0^1 \tilde{G}(\tilde{\beta}, x, w)\, d\bar{N}(x)\, d\bar{N}(w). \tag{2.36}$$

2.2.3 Lin-Ying's CCH method

D. Y. LIN and Z. YING [16] proposed an approximate partial likelihood estimator (APLE) to the true value β_0, which is consistent and asymptotically normal under regularity conditions.

The approximate partial-likelihood score function can be written in equation 2.37, which is a function of the sum over the uncensored failure times of the observed value of $Z_i(X_i)$ minus its "estimated" conditional expectation. The APLE $\tilde{\beta}$ is the root to the $\tilde{U}(\beta) = 0$.

$$\tilde{U}(\beta) = \sum_{i=1}^{n} \Delta_i H_i(X_i)\{Z_i(X_i) - E(\beta, X_i)\}. \tag{2.37}$$

Under the assumptions that $A(\beta_0)$ is nonsingular and that $\tilde{U}(\beta) = 0$ has a unique root, $n^{1/2}(\tilde{\beta} - \beta_0)$ is asymptotically normal with mean 0 and covariance matrix $A^{-1}(\beta_0)B(\beta_0)A^{-1}(\beta_0)'$, which can be estimated as $A_n^{-1}(\tilde{\beta_0})B_n(\tilde{\beta_0})A_n^{-1}(\tilde{\beta_0})'$, where all variables/matrix were explictly defined in [16].

$$n^{1/2}(\tilde{\beta} - \beta_0) \rightarrow_D N(0, A^{-1}(\beta_0)B(\beta_0)A^{-1}(\beta_0)'). \tag{2.38}$$

Their variance estimator is a Jackknife estimator and is much easier to calculate than the estimators of Prentice [39] and Self and Prentice [49]. Furthermore, if there

are multiple subcohort augmentations, its form will keep unchanged. Furthermost, incomplete covariate measurements on the cases are allowed in the estimator.

2.2.4 Barlow's CCH method

Barlow [6] proposed a weighted CCH analysis. Tables 1 and 2 (from Barlow's paper) describe the basic idea. For a full cohort with n subjects, there are m failure and $n - m$ censor. When sampling fraction, the percentage of subcohort to full cohort, is α, the expected cell frequencies for each cell are listed in Table 1. In Barlow's method, the weights are proportional to the subcohort size. For example, if the sampling fraction is 0.2, Barlow's method weighs controls in the subcohort as 5. Since all cases are included in this design, cases outside the subcohort are given a weight of 1. Subcohort cases are treated in two ways.—Before their event, they are weighted with a factor of 5, just like the controls. But at the time of their event, they are treated like cases outside the subcohort, with the weight of 1. So his method is sensitive to the sampling fraction.

TABLE 1. Expected cell frequencies in a case-cohort design for a cohort of size n with m failures and sampling fraction α

	Failure	Censored	Total
Subcohort	Cell 1: $\alpha\, m$	Cell 3: $\alpha\,(n\text{-}m)$	$\alpha\, n$
Not subcohort	Cell 2: $(1\text{-}\alpha)\, m$	Cell 4: $(1\text{-}\alpha)\,(n\text{-}m)$	$(1\text{-}\alpha)\, n$
Total	m	$n\text{-}m$	n

He also compared his method with prentice and self-prentice methods. Table 2 (from Barlow's paper) describes the weighting schemes for the three ways. For prentice's approach, only weight "Case outside subcohort before failure" to 0 and weight all rest situations to 1. And for the Self and Prentice method, only weight "Case outside subcohort before failure" and "Case outside subcohort at failure" to 0 and weight all rest situations to 1.

TABLE 2. Denominator weights in the pseudolikelihood for cases ad controls by method

Outcome type and timing	Prentice [2]	Self and Prentice [7]	Barlow [8]
Case outside subcohort before failure	0	0	0
Case outside subcohort at failure	1	0	1
Case in subcohort before failure	1	1	$1/\alpha$
Case in subcohort at failure	1	1	1
Subcohort control	1	1	$1/\alpha$

When a case outside of sub-cohort suddenly appears at its own failure time, this may cause a correlation if this case was not included in the earlier failure time [6]. Barlow used a simple jackknife variance estimate on the estimation of true value β to consider the correlation. The equation is listed below. $\Delta(\hat{\beta}_j)$ is the change of $\hat{\beta}$ if the jth individual is deleted, and it is a p-dimensional vector for p covariates.

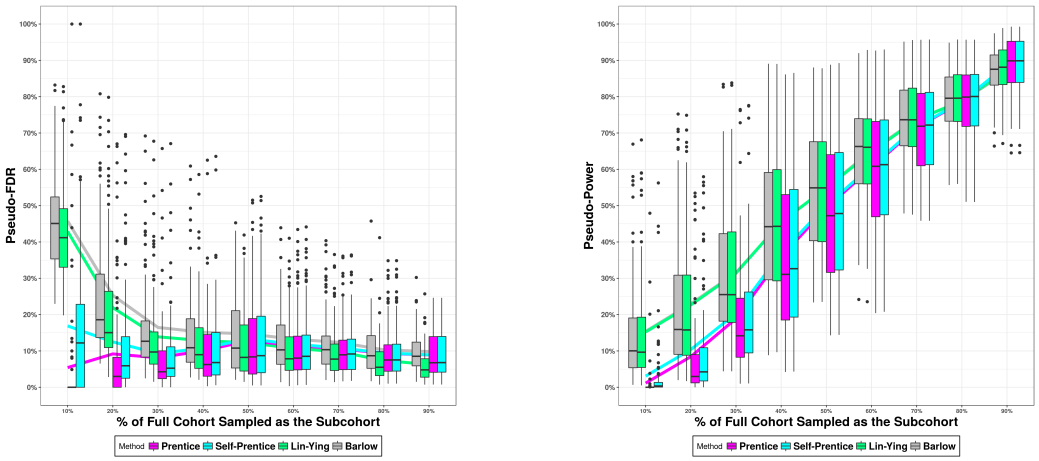
$$Var(\hat{\beta}) = \sum_i (\Delta(\hat{\beta}_j))_{p*1} (\Delta(\hat{\beta}_j)^T)_{1*p}. \tag{2.39}$$

2.2.5 CCH designs on high-throughput genomic studies

Onland [35] compared three CCH analysis methods, Prentice, Self and Prentice, and Barlow, in the context of fitting models. As discussed above, these methods use different weighting schemes for controls and cases inside or outside the subcohort. The CCH designs can be used in high-throughput genomic studies. But compared with full cohort study designs, the related research is rarely reported.

In John Carl Pesko 's Ph.D. dissertation [37], the performance of four CCH methods (Prentice, Self-Prentice, Lin-Ying, and Barlow) were compared on Genomic Data, which is high-throughput as the number of features is usually far greater than the number of observations. The purpose is to identify differentially expressed genes (DEGs) from the actual and simulated datasets with the four methods and explore their difference. True DEGs are known in simulation studies. So he can estimate the power of a method with the proportion of DEGs found to be significant and the

FDR with the ratio of significant genes that are not DEGs.

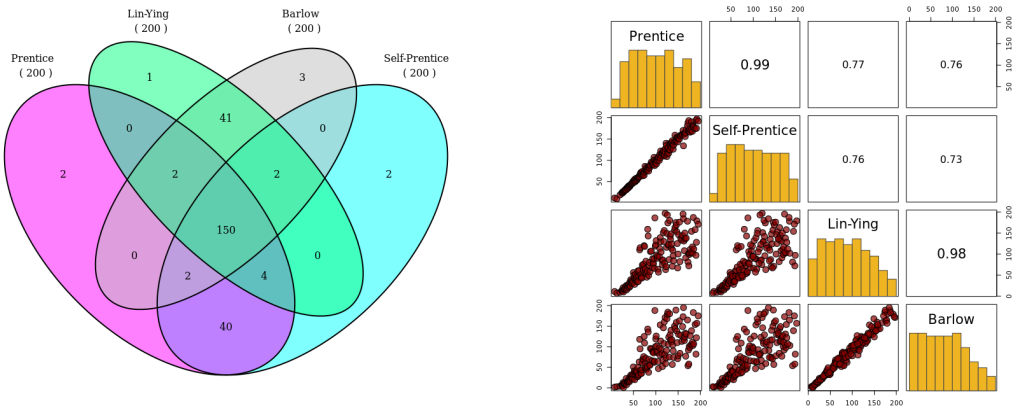


(a) Pseudo-FDR (b) Pseudo-power

Figure 2.3: Box plots displaying the pseudo-FDR and Pseudo-power achieved by each CCH method for the childhood leukemia dataset. Each box represents the 5-number summary for pseudo-FDR and Pseudo-power over 100 samples for a given level of π, with black dots indicating outliers. The smooth lines represent the mean pseudo-FDR for each method. (The images are from John's dissertation)

For real data, true DEGs are not known, but it is possible to investigate how well a CCH analysis captures the results from a complete cohort analysis. To do this, two measures, pseudo-FDR, and high pseudo-power were defined. Pseudo-FDR is the proportion of significant genes in a CCH analysis that are not also detected by the full cohort analysis, while pseudo-power is the proportion of significant genes from the full cohort analysis that are detected by a CCH analysis. For a CCH method to be considered an adequate substitute for complete cohort analysis, it should have low pseudo-FDR and high pseudo-power.

The left sub-figure of Figure 2.3 shows the pseudo-FDR achieved by each CCH method. Lower pseudo-FDR indicates good performance. The right sub-figure of Figure 2.3 displays the pseudo-power achieved by each CCH method. Higher pseudo-power means good performance.



(a) venn (b) scatter

Figure 2.4: Left-subfigure: the Venn diagram of the CCH method agreement for the childhood leukemia dataset considers the top 200 genes identified by each method. Counts in overlapping regions indicate the number of the top 200 genes agreed upon by the approaches involved in the overlap. Right-subfigure: Scatter plot matrix of gene ranking by method for the simulated data. Genes are ranked from 1 to 200 in order from most to least significant, based on their adjusted p-value. The lower triangle displays the bivariate scatter plots, while the upper triangle shows the corresponding correlation coefficient (Pearson's r). (The images are from John's dissertation)

Results show that the performance of the four methods differs in small subcohorts. For example, Barlow and Lin & Ying demonstrate much higher pseudo-FDR than the other two methods, and Self-Prentice exhibited a large amount of variability at $\pi = 10\%$. As π increases, the performance of the four methods becomes similar. For example, mean pseudo-FDR tends to decrease, and the distribution of pseudo-FDR for each method approaches a similar shape, particularly when $\pi \geq 40\%$. For pseudo-power, they have a similar pattern.

In addition to FDR and power, two approaches were used to consider method concordance. The Venn diagram (Left-subfigure of Figure 2.4) displayed how many genes identified as significant across methods are the same. The overlap indicates the number of shared genes among the top 200 most significant genes identified by

each way. As at least 150 out of 200 genes are the same, the results show high concordance among the four methods. And a scatter plot matrix (Right-subfigure of Figure 2.4) compares how similarly the top genes are ranked in each way. A correlation coefficient of 1 would indicate perfect agreement between two methods, and 0 would indicate definitive agreement between two approaches. The results show a high correlation among the four methods.

Chapter 3

A CCH-Based Score Test

Improving DEGs' identification under CCH has medical and biological values. We want to combine the advantages of the CCH study design and score test in differential gene expression analysis of rare diseases.

Self and Prentice [49] proved the asymptotic normality of the "CCH score process", $\tilde{U}(\beta_0, 1)$, in equation 3.1, which is defined as the first derivative of the log Pseudolikelihood function. The variance matrix $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ are positive definite matrix. However, the specific form of related functions to compute $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ can not be explicitly defined. For example, we do not know the specific form of functions (Refer to Appendix A). $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ can not be computed and need to be estimated. Prentice and Self estimated $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ with $\tilde{\Sigma}(\tilde{\beta})$ and $\tilde{\Delta}(\tilde{\beta})$ at $\tilde{\beta}$. To acquire $\tilde{\beta}$, need to fit the Coxph model iteratively. Convergence and computational cost are potential risks.

In this chapter, a CCH-based score test was proposed, in which the covariance matrix $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ in the asymptotic chi-square distribution of CCH were estimated at β_0 rather than at $\tilde{\beta}$. The Cox PH model requires $\beta_0 = 0$ for NONDEGs, which makes that survival time has no association with gene expression data. Fur-

thermore, the "Score Process" was acquired by deriving the first derivative of the log Pseudo-likelihood function. Then, a test statistic with Chi-squared distribution was built to calculate the p-value under the null hypothesis of the proposed score test, $\beta = 0$, which is equivalent to the null hypothesis that the tested gene is a NON-DEG. Simulation studies were conducted to study the Type-I error and power for a single gene, and false discovery rate (FDR) and power for high-throughput data. Besides, real datasets were applied to measure the consistency between the proposed CCH-based test and the full cohort analysis method.

3.1 Asymptotical Distribution of the CCH Score

The "CCH score process", $\tilde{U}(\beta_0, 1)$, asymptotically follows a normal distribution.

$$n^{-1/2}\tilde{U}(\beta_0, 1) \rightarrow_D N(0, \Sigma(\beta_0) + \Delta(\beta_0)). \tag{3.1}$$

$\Sigma(\beta_0)$ and $\Delta(\beta_0)$ are positive definite. With singular value decomposition, we get

$$(\Sigma(\beta_0) + \Delta(\beta_0))^{(-1/2)} n^{-1/2}\tilde{U}(\beta_0, 1) \rightarrow_D N(0, I). \tag{3.2}$$

By squaring $(\Sigma(\beta_0)+\Delta(\beta_0))^{(-1/2)} n^{-1/2}\tilde{U}(\beta_0, 1)$, a CCH-based score test statistic can be found, which has an asymptotic chi-square distribution.

$$n^{-1}\tilde{U}(\beta_0, 1)^T(\Sigma(\beta_0) + \Delta(\beta_0))^{-1}\tilde{U}(\beta_0, 1) \rightarrow_D \chi^2_k. \tag{3.3}$$

When β is a single variable, the equation 3.3 has a simple form.

$$n^{-1}\frac{\tilde{U}(\beta_0,1)^2}{\Sigma(\beta_0)+\Delta(\beta_0)} \rightarrow_D \chi^2_1. \tag{3.4}$$

When β is a single variable and under the null hypothesis, $\beta = 0$, a Z-test or a chi-square test is equivalent to identifying DEGs. When we use a Z-test, the gene is up-regulated if the statistic has a positive value, and the gene is down-regulated

if the statistic has a negative value. However, different datasets have different gene expression values; some have magnitude gene expressions and large covariance. If divided by their variance, a positive value, their test statistics are "standardized," which is an advantage of using chi-square distribution.

3.2 A Proposed Chi-Square Score Test

To propose a Chi-Square Score Test, the variance matrix Σ and Δ need to be estimated, and $\tilde{U}(\beta_0, 1)$ need to be calculated. They will be shown step by step in the next subsections.

3.2.1 Estimating the variance matrix of the asymptotic Chi-Square distribution

The Cox proportional hazards model for the hazard process yields

$$\lambda_j(t) = Y_j(t)\lambda_0(t)r\{\beta_0' Z_j(t)\}, \tag{3.5}$$

where λ_0 is a fixed function under the proportional hazards assumption, and $Y_j(t)$ is an indicator function. At time t, if the jth patient/observation is "at risk" for observable failure, $Y_j(t) = 1$. Otherwise, $Y_j(t) = 0$. Consistent with the method of simulating the data for the proportional hazards model (Refer to Section 3.3), $r(x) = exp(x)$, $r^{(1)}(x) = dr(x)/dx$ and $r^{(2)}(x) = dr^{(1)}(x)/dx$. For each unique event/failure time t,

$$r\{\beta_0' Z_l(t)\} = exp(\beta_0' Z_l(t)), \tag{3.6}$$

$$r^{(1)}\{\beta_0' Z_l(t)\} = dr\{\beta_0' Z_l(t)\}/d\{\beta_0' Z_l(t)\} = d(exp(\beta_0' Z_l(t)))/d\{\beta_0' Z_l(t)\} = exp(\beta_0' Z_l(t)),$$

(3.7)

and

$$r^{(2)}\{\beta_0' Z_l(t)\} = dr^{(1)}\{\beta_0' Z_l(t)\}/d\{\beta_0' Z_l(t)\} = d(exp(\beta_0' Z_l(t)))/d\{\beta_0' Z_l(t)\} = exp(\beta_0' Z_l(t)). \tag{3.8}$$

Under the null hypothesis, $\beta_0 = 0$, $r\{\beta_0' Z_l(t)\} = 1$ and $r^{(i)}\{\beta_0' Z_l(t)\} = 1$, $i = 1$ or 2. We also need function $u(x) = log(r(x)) = x$. Similarly, $u^{(1)}(x) = du(x)/dx = 1$ and $u^{(2)}(x) = du^{(1)}(x)/dx = 0$.

Self and Prentice [49] defined specific forms of functions to prove asymptotical normality of $\tilde{\beta}$ and to estimate the true value of β_0 by $\tilde{\beta}$, but they did not set up a score test with their theory. We can reference these functions to do hypothesis testing and propose a Chi-Square score test on the CCH data. However, we need to expand them at β_0, not at $\tilde{\beta}$, because β_0 is assumed to be known under null hypothesis, which is one advantage of the CCH-based Score Test.

For each unique event time t and each patient/observation l in a CCH, $X_l^{(i)}(\beta_0, t)$ and $\tilde{S}^{(i)}(\beta_0, t)$ are defined below, for $i = 0$, 1 and 2, respectively.

$$X_l^{(0)}(\beta_0, t) = r\{\beta_0' Z_l(t)\} = exp(\beta_0' Z_l(t)). \tag{3.9}$$

$$X_l^{(1)}(\beta_0, t) = Z_l(t) r^{(1)}\{\beta_0' Z_l(t)\} = Z_l(t) exp(\beta_0' Z_l(t)). \tag{3.10}$$

$$X_l^{(2)}(\beta_0, t) = Z_l(t)^{\otimes 2} u^{(1)}\{\beta_0' Z_l(t)\}^2 r\{\beta_0' Z_l(t)\} = Z_l(t)^{\otimes 2} exp(\beta_0' Z_l(t)), \tag{3.11}$$

where $z^{\otimes 2}$ denotes the $p * p$ matrix with (i, j) element $z_i z_j$ for any $z' = (z_1, z_2, ...z_p)$. When $\beta_0 = 0$, $X_l^{(0)}(0, t) = 1$, $X_l^{(1)}(0, t) = Z_l(t)$ and $X_l^{(2)}(0, t) = Z_l(t)^{\otimes 2}$.

$$\tilde{S}^{(i)}(\beta_0, t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) X_l^{(i)}(\beta_0, t), \tag{3.12}$$

where $\tilde{n}^{-1}$ is the number of patients/observations in the CCH, for $i = 0$, 1 and 2, respectively.

Under the null hypothesis, $\beta = 0$, the values of $\tilde{S}^{(i)}(0,t)$ can be calculated by the functions below.

$$\tilde{S}^{(0)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) X_l^{(0)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t). \tag{3.13}$$

$$\tilde{S}^{(1)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) X_l^{(1)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) Z_l(t). \tag{3.14}$$

$$\tilde{S}^{(2)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) X_l^{(2)}(0,t) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) Z_l(t)^{\otimes 2}. \tag{3.15}$$

And define

$$\tilde{E}(\beta_0, t) = \tilde{S}^{(1)}(\beta_0, t) / \tilde{S}^{(0)}(\beta_0, t), \tag{3.16}$$

$$\tilde{V}(\beta_0, t) = \tilde{S}^{(2)}(\beta_0, t) / \tilde{S}^{(0)}(\beta_0, t) - \tilde{E}(\beta_0, t)^{\otimes 2}, \tag{3.17}$$

and

$$\tilde{\Sigma}(\beta_0) = \tfrac{1}{n} \textstyle\int_0^1 \tilde{V}(\beta_0, t)\, d\bar{N}(t). \tag{3.18}$$

For $x, w \in [0,1]^2$, $\tilde{Q}^{(i)}(\beta_0, x, w)$, $\tilde{H}^{(i)}(\beta_0, x, w)$ and $\tilde{G}(\beta_0, x, w)$ are defined below for $i = 0$, 1 and 2, respectively. They are bounded on $\beta_0 \times [0,1]^2$.

$$\tilde{Q}^{(0)}(\beta_0, x, w) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) Y_l(w) X_l^{(0)}(\beta_0, x) X_l^{(0)}(\beta_0, w). \tag{3.19}$$

$$\tilde{Q}^{(1)}(\beta_0, x, w) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) Y_l(w) X_l^{(1)}(\beta_0, x) X_l^{(1)}(\beta_1, w). \tag{3.20}$$

$$\tilde{Q}^{(2)}(\beta_0, x, w) = \tilde{n}^{-1} \sum_{l \in \tilde{C}} Y_l(t) Y_l(w) X_l^{(0)}(\beta_0, x) X_l^{(1)}(\beta_1, w). \tag{3.21}$$

$$\tilde{H}^{(0)}(\beta_0, x, w) = \tilde{Q}^{(0)}(\beta_0, x, w) - \tilde{S}^{(0)}(\beta_0, x) \tilde{S}^{(0)}(\beta_0, w). \tag{3.22}$$

$$\tilde{H}^{(1)}(\beta_0, x, w) = \tilde{Q}^{(1)}(\beta_0, x, w) - \tilde{S}^{(1)}(\beta_0, x) \tilde{S}^{(1)}(\beta_0, w)'. \tag{3.23}$$

$$\tilde{H}^{(2)}(\beta_0, x, w) = \tilde{Q}^{(2)}(\beta_0, x, w) - \tilde{S}^{(0)}(\beta_0, x) \tilde{S}^{(1)}(\beta_0, w). \tag{3.24}$$

$$\begin{aligned}
\tilde{G}(\beta_0, x, w) = {} & \tfrac{1-\tilde{\alpha}}{\tilde{\alpha}} [\{\tilde{S}^{(0)}(\beta_0, x) \tilde{S}^{(0)}(\beta_0, w)\}^{-1} \tilde{H}^{(1)}(\beta_0, x, w) \\
& + \{\tilde{S}^{(0)}(\beta_0, x) \tilde{S}^{(0)}(\beta_0, w)\}^{-2} \tilde{S}^{(1)}(\beta_0, x) \tilde{S}^{(1)}(\beta_0, w)^T \tilde{H}^{(0)}(\beta_0, x, w) \\
& - \tilde{S}^{(0)}(\beta_0, x)^{-1} \tilde{S}^{(0)}(\beta_0, w)^{-2} \tilde{S}^{(1)}(\beta_0, w) \tilde{H}^{(2)}(\beta_0, w, x) \\
& - \tilde{S}^{(0)}(\beta_0, w)^{-1} \tilde{S}^{(0)}(\beta_0, x)^{-2} \tilde{S}^{(1)}(\beta_0, x) \tilde{H}^{(2)}(\beta_0, x, w)],
\end{aligned} \tag{3.25}$$

where $\tilde{\alpha} = (1 - eventRate) \times sr + eventRate$. The $eventRate$ is the event rate in the full cohort, and sr is the subfraction of the subcohort out of the full cohort.

$$\tilde{\triangle}(\beta_0) = \frac{1}{n^2} \int_0^1 \int_0^1 \tilde{G}(\beta_0, x, w)\, d\bar{N}(x)\, d\bar{N}(w). \tag{3.26}$$

3.2.2 Computing the "Score Process" from the Pseudo-likelihood function

A weighted likelihood function was proposed [39] to account for the sampling scheme in a CCH study.

$$L(\beta) = \prod_{j \in E} \frac{e^{x_j \beta}}{e^{x_j \beta} + \sum\limits_{\substack{k \in S \\ k \neq i}} Y_k e^{x_k \beta}}, \tag{3.27}$$

where E is the set of unique event time. In each unique event time j, x_j is the gene expression value of the case in the event at time j. If more than one case has an event at the same time, we compute the contribution of each of them separately and then multiply them together. The summation in the denominator only includes individuals at risk who are also in the subcohort. Y_k indicates whether patient k is in risk at event time j or not. If in risk, $Y_k = 1$. Otherwise, $Y_k = 0$. However, individuals related with event time j can either be a case from inside or outside of the subcohort.

$$logL(\beta) = \sum_{j \in E} [log e^{x_j \beta} - log(e^{x_j \beta} + \sum_{\substack{k \in S \\ k \neq j}} Y_k e^{x_k \beta})]. \tag{3.28}$$

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{j \in E} [\frac{x_j e^{x_j \beta}}{e^{x_j \beta}} - \frac{x_j e^{x_j \beta} + \sum\limits_{\substack{k \in S \\ k \neq j}} Y_k x_k e^{x_k \beta}}{e^{x_j \beta} + \sum\limits_{\substack{k \in S \\ k \neq j}} Y_k e^{x_k \beta}}]. \tag{3.29}$$

$$\tilde{U}(\beta,1) = \frac{\partial logL(\beta)}{\partial \beta} = \sum_{j\in E}[\frac{x_j e^{x_j\beta}}{e^{x_j\beta}} - \frac{x_j e^{x_j\beta} + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k x_k e^{x_k\beta}}{e^{x_j\beta} + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k e^{x_k\beta}}]. \tag{3.30}$$

When expanding the equation in β_0 and, under null hypothesis, $\beta = 0$, we have

$$\frac{\partial logL(\beta)}{\partial \beta}|_{\beta=0} = \sum_{j\in E}[x_j - \frac{x_j + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k x_k}{1 + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k}]. \tag{3.31}$$

By the definition of "Score" process, $\tilde{U}(\beta_0 = 0, 1)$ has the form of function 3.32, which can be looked as, summation of gene expression value of each case minus the related weighted average gene expression value.

$$\tilde{U}(\beta_0 = 0,1) = \frac{\partial logL(\beta)}{\partial \beta}|_{\beta=0} = \sum_{j\in E}[x_j - \frac{x_j + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k x_k}{1 + \sum\limits_{\substack{k\in S\\ k\neq j}} Y_k}]. \tag{3.32}$$

3.2.3 Setting up CCH score test statistic

$\Sigma(\beta_0)$ can be estimated by $\tilde{\Sigma}(\beta_0)$, and $\Delta(\beta_0)$ can be estimated by $\tilde{\Delta}(\beta_0)$ (See appendix A for details). The estimators can be calculated from equation 3.18 and equation 3.26, respectively. Then we can approximately write the score test statistic for CCH design as:

$$n^{-1}\tilde{U}(\beta_0,1)^T(\Sigma(\tilde{\beta_0}) + \tilde{\Delta}(\beta_0))^{-1}\tilde{U}(\beta_0,1) \rightarrow_D \chi^2_k. \tag{3.33}$$

When β is a single variable and under the null hypothesis, $\beta = \beta_0 = 0$, the equation 3.33 has a simple form (equation 3.34). $n^{-1}\frac{\tilde{U}(\beta_0=0,1)^2}{\tilde{\Sigma}(\beta_0=0)+\tilde{\Delta}(\beta_0=0)}$ is used as the test statistic in the proposed CCH based score test, which follows a χ^2_1 distribution. Where $\tilde{U}(\beta_0 = 0, 1)$ can be calucated from equation 3.32.

$$n^{-1}\frac{\tilde{U}(\beta_0=0,1)^2}{\tilde{\Sigma}(\beta_0=0)+\tilde{\Delta}(\beta_0=0)} \rightarrow_D \chi^2_1. \tag{3.34}$$

3.3 Simulation Study

3.3.1 Data Simulation

Bender et al. [7] discussed techniques to generate survival times for simulation studies regarding Cox proportional hazards models. The survival function of the Cox proportional hazards model is given by

$$S(t|x) = exp[-H_0(t)e^{x\beta}], \tag{3.35}$$

where $H_0(t)$ is the cumulative hazard function, and β is the vector of regression coefficients associated with the predictor covariates x. The distribution function of the Cox model is

$$F(t|x) = 1 - exp[-H_0(t)e^{x\beta}]. \tag{3.36}$$

As $F(t|x)$ follows a continuous $\mathcal{U}(0,1)$ distribution, a random variable $U = 1 - F(t|x) = exp[-H_0(t)e^{x\beta}]$ also follows a continuous $\mathcal{U}(0,1)$ distribution.

By the inverse Probability integral transform (PIT), survival time T can be written as a random variable:

$$T = S^{-1}(U|x) = H_0^{-1}\left(-\frac{log(U)}{e^{x\beta}}\right). \tag{3.37}$$

Using a Weibull distribution with scale λ and shape ρ, the inverse of the cumulative hazard function is below:

$$H_0^{-1}(t) = (\lambda^{-1}t)^{1/\rho}. \tag{3.38}$$

Based on the two formulas above, survival times can be generated with:

$$T = \left(-\frac{log(U)}{\lambda e^{x\beta}}\right)^{\frac{1}{\rho}}. \tag{3.39}$$

We can use equation 3.39 to generate survival times. What we need are to generate a set of realization u, choose a $\beta = \texttt{log(HR)}$, generate a set of x from $\texttt{Normal}(0, 1)$ distribution, and plug them into the equation. Then we can get a set of survival times.

Survival times can be used to generate one gene expression vector (the length of the vector is the size of all patients). For our simulations, the shape, scale, and censoring rate parameters were fixed at $\rho = 1$, $\lambda = 1$, and $\lambda_{\texttt{cens}} = 10$ to give an incidence rate near 10%. To generate gene expression of many DEGs, we generate survival times and censoring information using the method described above and rewrite the survival time equation in terms of x_{ij}. The expression level of gene i for individual j is:

$$x_{ij} = \frac{-log\left(\frac{-\lambda t_j^{\rho}}{log(U_j)}\right)}{\beta_i} + e_{ij}, \tag{3.40}$$

where the e_{ij}s are $\texttt{N}(0, 1)$ perturbations. We draw a set of perturbations to generate expression levels for each DEG. To ensure each expression level and survival time correlate around the original pair. We did a check. If the correlation between perturbation and survival time is between 0.8 and 1.2 times the correlation between original expression and survival time, we accept the perturbated expression values and add the vector to the dataset. Otherwise, we redo the perturbation. We can draw random numbers for the null genes since we only care that they are unrelated to survival.

We need to separate the controls and cases in a survival analysis study. For all patients, draw censoring times $C \sim Exp(\lambda_{cens})$, and compare them to their corresponding survival times. If censoring occurs after the survival time (censoring time is longer than the survival time) for an individual, he/she was observed to have experienced an event and is considered a case. Otherwise, he/she is in the control group.

Simulating single DEG or single NONDEG datasets

For simulated data, we need to know which genes are DEGs and which are NON-DEGs. For NONDEG genes, we picked $\beta = 0$ on equation 3.37, and we simulated 1000 NONDEG genes for each combination of event rate (0.05, 0.1, 0.15 and 0.2) and full cohort size (500, 1000, 1500 and 2000). Each gene was saved as a separate dataset.

For DEG genes, we picked $\beta = log(effectsize)$ on Equation 3.37, and "effect size" can pick values from 1.2 - 1.3 for a low correlation of gene expression value and survival time, up to 1.7 - 1.8 for a high correlation of expression value and survival time. We simulated 1000 DEG genes for each combination of event rate (0.05, 0.1, 0.15 and 0.2) and full cohort size (500, 1000, 1500 and 2000). Each gene was saved as a separate dataset.

Simulating high-throughput datasets

A high-throughput dataset includes DEGs and NONDEGs at the same time. Besides event rate, effect size, and subfraction for a single gene, the proportion of DEGs in the dataset should be considered. To consider randomness, 100 dataset with 5%, 10%, 15% or 20% proportion of DEGs were simulated for each combination of event rate (0.05, 0.1, 0.15 and 0.2) and effect size (1.2 to 1.3, 1.3 to 1.4, ... up to 1.7 to 1.8), respectively. Each dataset has 1000 patients and 2000 genes, and they were generated and saved independently. For example, at effect size 1.5 to 1.6, if proportion of DEGs is 20% and event rate is 0.1, a dataset should include 400 DEGs (2000 genes $\times$ 0.2 = 400 genes) and around 100 cases (1000 $\times$ 0.1 = 100).

Choosing event rate and fractions of subcohort to sample subcohorts from full-cohorts

CCH designs consider a random sample of the full cohort, called a subcohort. At the time of analysis, we add all cases outside the subcohort to the sample. In other words, a CCH sample consists of all cases (both in and out of the subcohort), but only the controls in the subcohort. For example, if the event rate is 10% (10% of patients are cases), there are 100 cases out of 1000 patients. Fraction of subcohort are 10%, 20%, 30%, 40%, 50%, 60% 70%, 80%, and 90%, respectively. If fraction is 20%, the expected number of patients in a sample is 280 (1000*0.2*0.9 + 100 = 280). Shown in Table 3.2).

Sub-cohort Fraction	**Full Cohort Size**			
	500	***1000***	***1500***	***2000***
0.1	73	145	218	290
0.2	120	240	360	480
0.3	168	335	503	670
0.4	215	430	645	860
0.5	263	525	788	1050
0.6	310	620	930	1240
0.7	358	715	1073	1430
0.8	405	810	1215	1620
0.9	453	905	1358	1810
1.0	500	1000	1500	2000

Table 3.1: The expected number of patients in a CCH for each sub-fraction. The event rate is 0.05, and the full cohort size is 500, 1000, 1500, and 2000, respectively. If Sub-cohort fraction is 20%, the expected number of patients in a sample is 240 ($1000 \times 0.2 \times 0.95 + 50 = 240$).

Sub-cohort Fraction	**Full Cohort Size**			
	500	***1000***	***1500***	***2000***
0.1	95	190	285	380
0.2	140	280	420	560
0.3	185	370	555	740
0.4	230	460	690	920
0.5	275	550	825	1100
0.6	320	640	960	1280
0.7	365	730	1095	1460
0.8	410	820	1230	1640
0.9	453	905	1358	1820
1.0	500	1000	1500	2000

Table 3.2: The expected number of patients in a CCH for each sub-fraction. The event rate is 0.1, and the full cohort size is 500, 1000, 1500, and 2000, respectively. . If Sub-cohort fraction is 20%, the expected number of patients in a sample is 280 $(1000 \times 0.2 \times 0.9 + 100 = 280)$.

In this study, we choose event rate 0.05, 0.10, 0.15 and 0.20, and full cohort size 500, 1000, 1500 and 2000. The expected number of patients in CCH's are listed in Table 3.1 - Table 3.4.

Sub-cohort Fraction	**Full Cohort Size**			
	500	***1000***	***1500***	***2000***
0.1	118	235	353	470
0.2	160	320	480	640
0.3	203	405	608	810
0.4	245	490	735	980
0.5	288	575	863	1150
0.6	330	660	990	1320
0.7	373	745	1118	1490
0.8	415	830	1245	1660
0.9	458	915	1373	1830
1.0	500	1000	1500	2000

Table 3.3: The expected number of patients in a CCH for each sub-fraction. The event rate is 0.15, and the full cohort size is 500, 1000, 1500, and 2000, respectively. If the Sub-cohort fraction is 20%, the expected number of patients in a sample is 320 $(1000 \times 0.2 \times 0.85 + 150 = 320)$.

Sub-cohort Fraction	**Full Cohort Size**			
	500	***1000***	***1500***	***2000***
0.1	140	280	420	560
0.2	180	360	540	720
0.3	220	440	660	880
0.4	260	520	780	1040
0.5	300	600	900	1200
0.6	340	680	1020	1360
0.7	380	760	1140	1520
0.8	420	840	1260	1680
0.9	460	920	1380	1840
1.0	500	1000	1500	2000

Table 3.4: The expected number of patients in a CCH for each sub-fraction. . The event rate is 0.2, and the full cohort size is 500, 1000, 1500, and 2000, respectively. If the Sub-cohort fraction is 20%, the expected number of patients in a sample is 360 ($1000 \times 0.2 \times 0.8 + 200 = 360$).

3.3.2 Type I error and power for single gene datasets

To validate our "CCH Score" method, its performance on type I error should be evaluated. To prove its efficiency, its power should be evaluated. The performance of Type I error and Power are listed in the figures below. Type I error is defined as the percentage of falsely identified NONDEGs out of all NONDEGs. For example, if 50 NONDEGs are falsely identified as DEGs out of all 1000 NONDEGs, the Type I error = 50/1000 = 0.05. For Power, we simulate 1000 DEG genes with a β larger than 0 for each parameter's combination on Equation 3.37. The survival time t will have a negative association relationship with gene expression value x. t will decreases if x increases given $-log(U)$ unchanged. The Power is defined as the percentage of correctly identified DEG genes out of all DEG genes. For example, if you identified 800 out of 1000 DEGs, the Power is 800/1000=0.80. Type I error and Power work for both full-cohort and Sub-cohort methods.

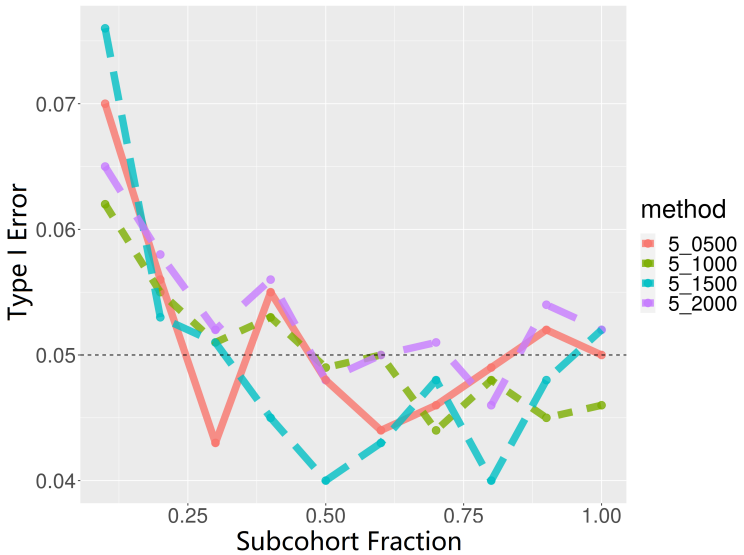


Figure 3.1: At case rate 0.05, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and full cohort 500, 1000, 1500, and 2000, respectively. For example, "5_0500" means at case rate 0.05 and full cohort size 500.

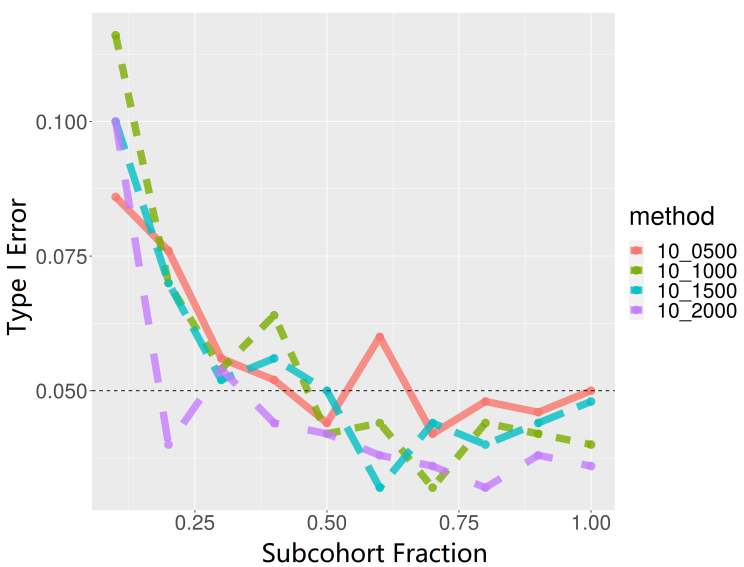


Figure 3.2: At case rate 0.10, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and full cohort 500, 1000, 1500, and 2000, respectively. For example, "10_1000" means at case rate 0.1 and full cohort size 1000.

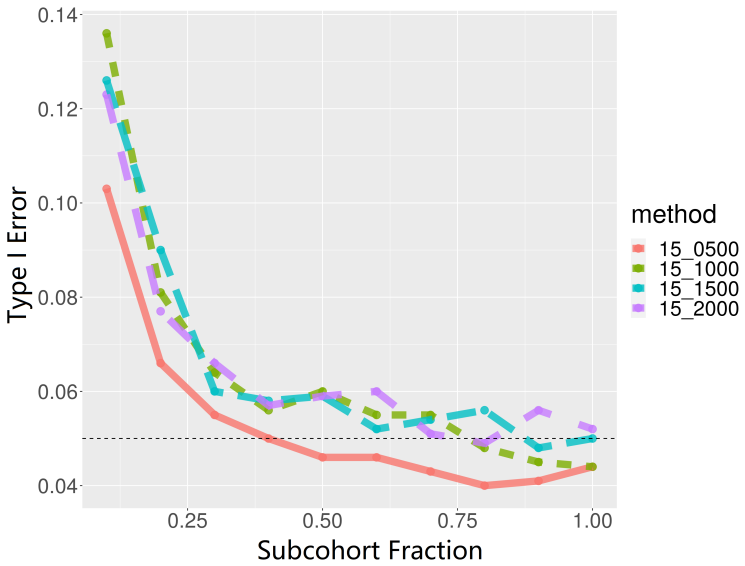


Figure 3.3: At case rate 0.15, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and full cohort 500, 1000, 1500, and 2000, respectively. For example, "15_1500" means at case rate 0.15 and full cohort size 1500.

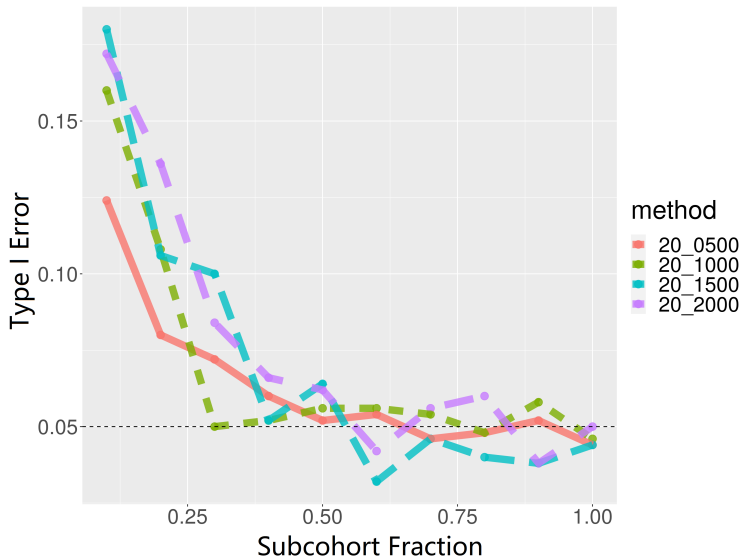


Figure 3.4: At case rate 0.20, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and full cohort 500, 1000, 1500, and 2000, respectively. For example, "20_0500" means at case rate 0.2 and full cohort size 500.

Figure 3.1 shows that, at case rate 0.05, Type I error of the "CCH Score" method on simulated single gene data with hazard ratio 1.5-1.6 will decrease quickly with the increase of subcohort fraction and keep around 0.05 for full cohort size 500, 1000, 1500 and 2000, respectively. Increasing or decreasing the full cohort size will not influence Type I error at a case rate of 0.05. For case rate 0.10, 0.15 and 0.2, Figure 3.2, Figure 3.3 and Figure 3.4 show similar results.

To evaluate Type I error for different full cohort sizes. Figure 3.5 shows that, at full cohort 500, Type I error of "CCH Score" method on simulated single gene data with hazard ratio 1.5-1.6 will decrease quickly with the increase of subcohort fraction and around 0.05 for case rate 0.05, 0.1, 0.15 and 0.2, respectively. For full cohort size 1000, 1500 and 2000, Figure 3.6, Figure 3.7 and Figure 3.8 show similar results.

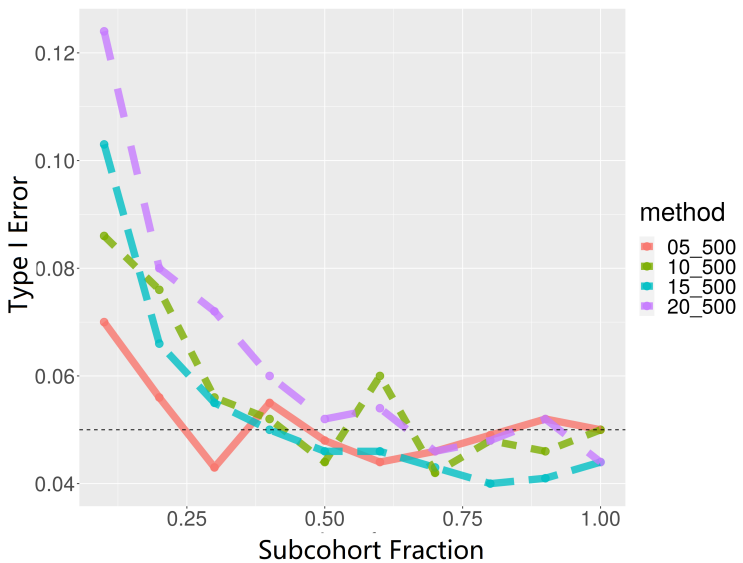


Figure 3.5: At full cohort 500, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15, and 0.2, respectively. For example, "20_0500" means at case rate 0.2 and full cohort size 500.

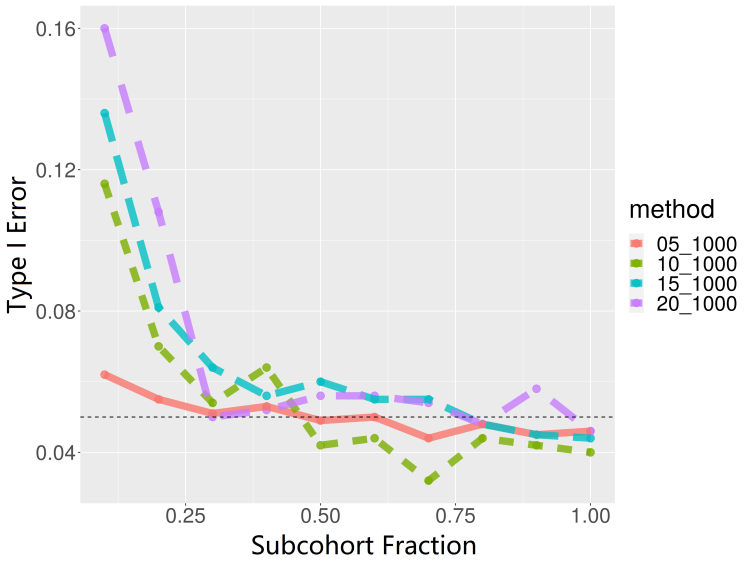


Figure 3.6: At full cohort 1000, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15, and 0.2, respectively. For example, "20_1000" means at case rate 0.2 and full cohort size 1000.

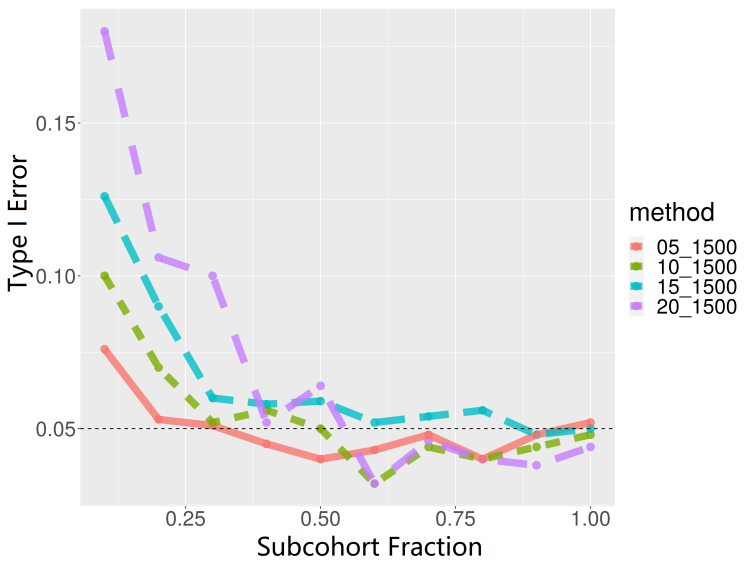


Figure 3.7: At full cohort 1500, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15, and 0.2, respectively. For example, "10_1500" means at case rate 0.1 and full cohort size 1500.

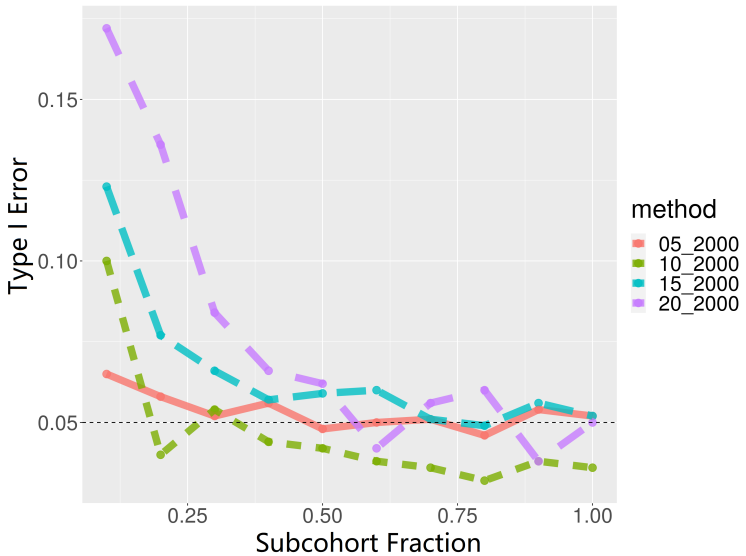


Figure 3.8: At full cohort 2000, Type I error of "CCH Score" method is a function of subcohort fraction on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15, and 0.2, respectively. For example, "20_2000" means at case rate 0.2 and full cohort size 2000.

In all, the Type I error of our "CCH Score" method is valid for all tested single gene data. From a small sample of 500 to a medium sample of 2000, Type I error works well at event rates 0.05, 0.1, 0.15, and 0.2, which are usually the event rates range of rare diseases.

Comparing power with other existing CCH based methods

We computed Power of "CCH Score" method for each combination of event rate (0.05, 0.10, 0.15 and 0.20) and full cohort size (500, 1000, 1500 and 2000) on subcohort fraction from 0.1, 0.2, 0.3 ... 1.0, and compared them with four existing methods that can also applied to the gene expression analysis under the CCH Design, "Prentice" [39], "SelfPrentice" [49], "LinYing" [16] and "Barlow" [6].

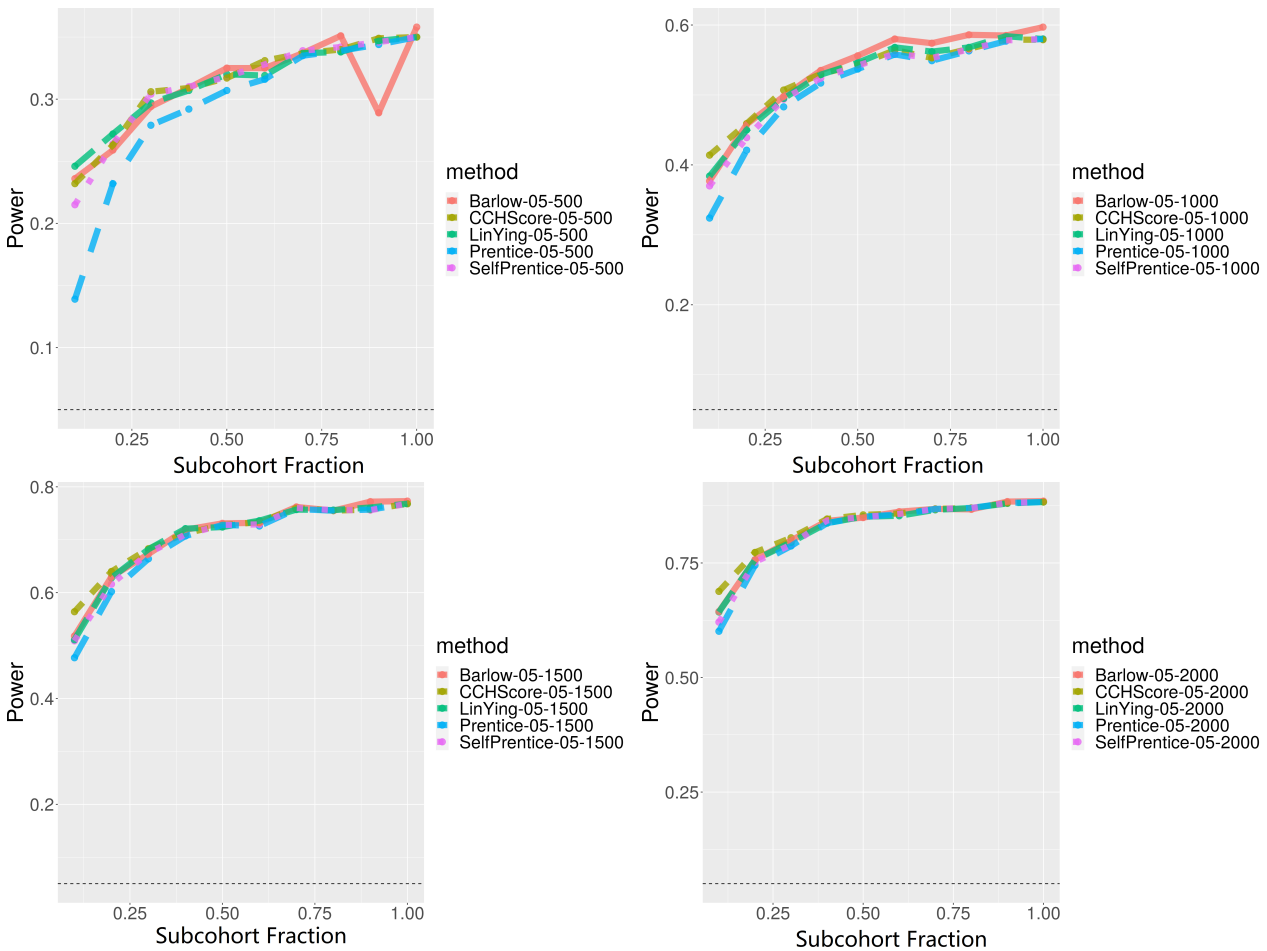


Figure 3.9: At case rate 0.05, comparing the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and full cohort size 500, 1000, 1500 and 2000, respectively. For example, "Barlow-05-1000" means at case rate 0.05 and full cohort size 1000, and with the "Barlow" method.

Figure 3.9 shows that, at case rate 0.05, the power of the "CCH Score" method on simulated single gene data with hazard ratio 1.5-1.6 will increase quickly with the increase of subcohort fraction and reach the almost same point at subcohort fraction = 1.0 (full cohort) as the same as other four methods for full cohort size 500, 1000, 1500 and 2000, respectively.

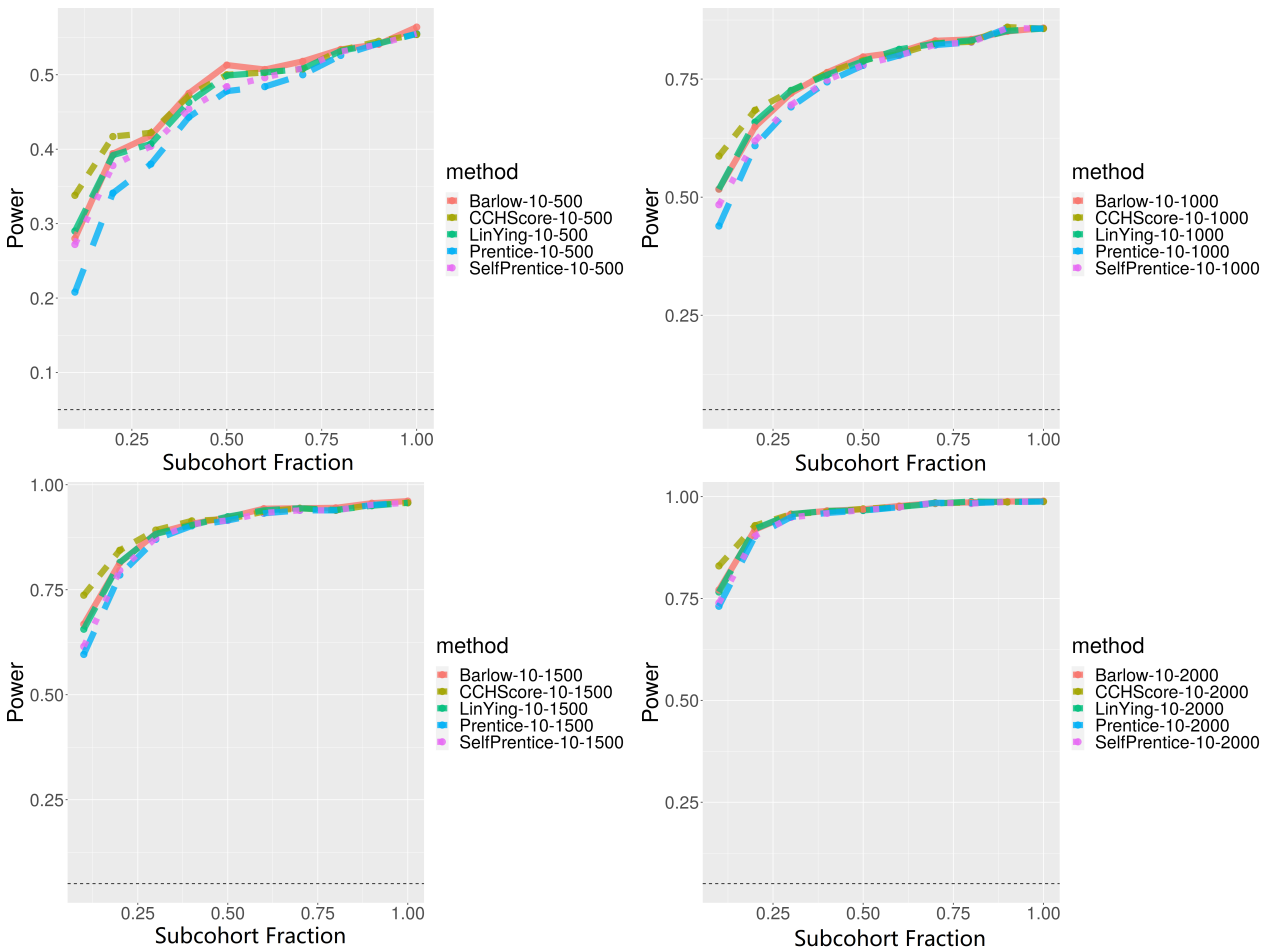


Figure 3.10: At case rate 0.10, comparing the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and full cohort size 500, 1000, 1500 and 2000, respectively. For example, "Barlow-10-1000" means at case rate 0.1, full cohort size 1000, and with the "Barlow" method.

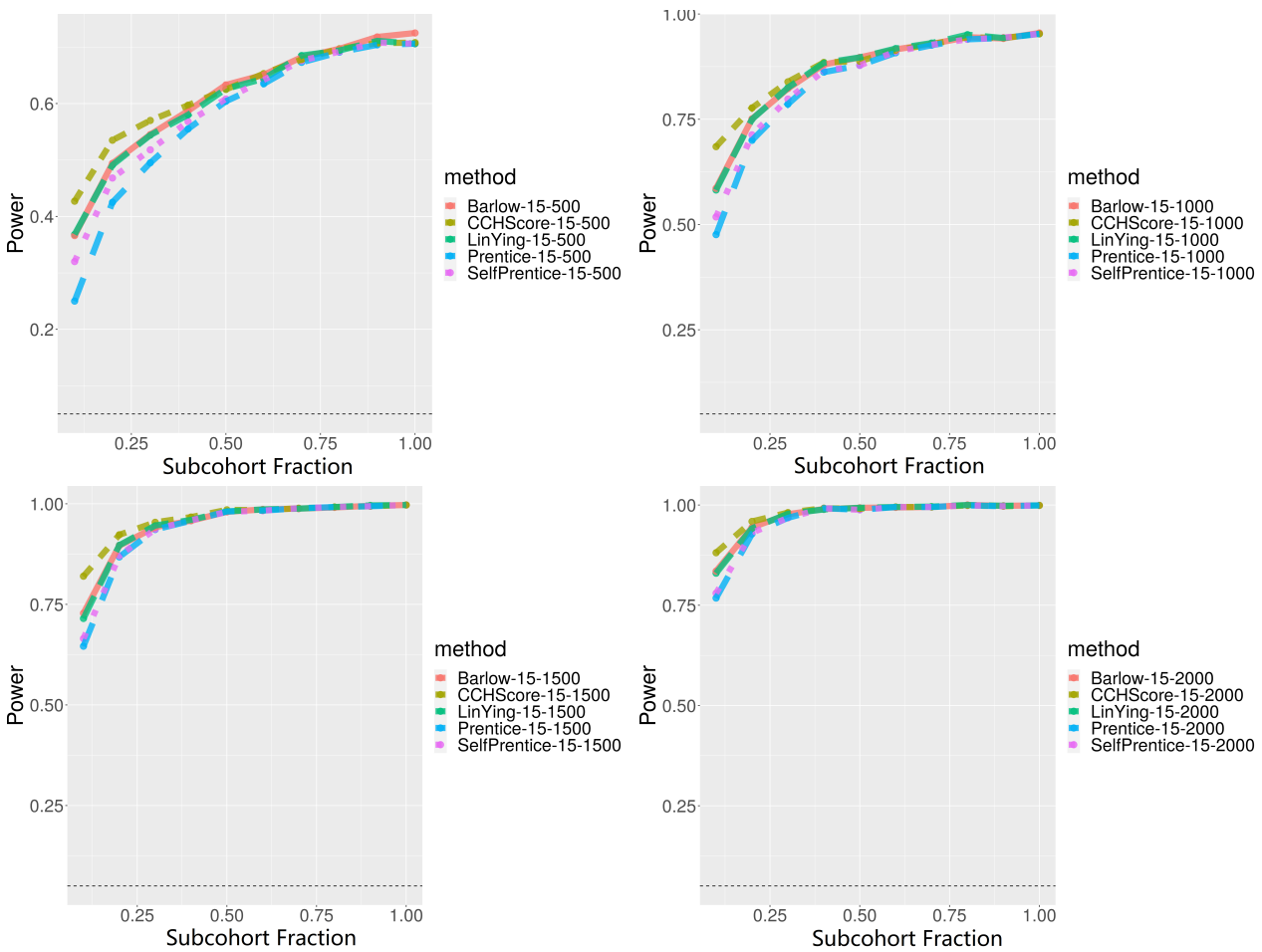


Figure 3.11: At case rate 0.15, comparing the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods methods on simulated single gene data with hazard ratio 1.5-1.6 and full cohort size 500, 1000, 1500 and 2000, respectively. For example, "Barlow-15-1000" means at case rate 0.15, full cohort 1000, and with the "Barlow" method.

The power of "CCH Score" method is consistent with the other four existing methods. Increasing or decreasing the full cohort size will not influence the consistency of the five methods at a case rate of 0.05. Besides, with the increase of the full cohort size from 500 to 2000, the power will increase. For case rate 0.10, 0.15 and 0.2, Figure 3.10, Figure 3.11 and Figure 3.12 show similar results.

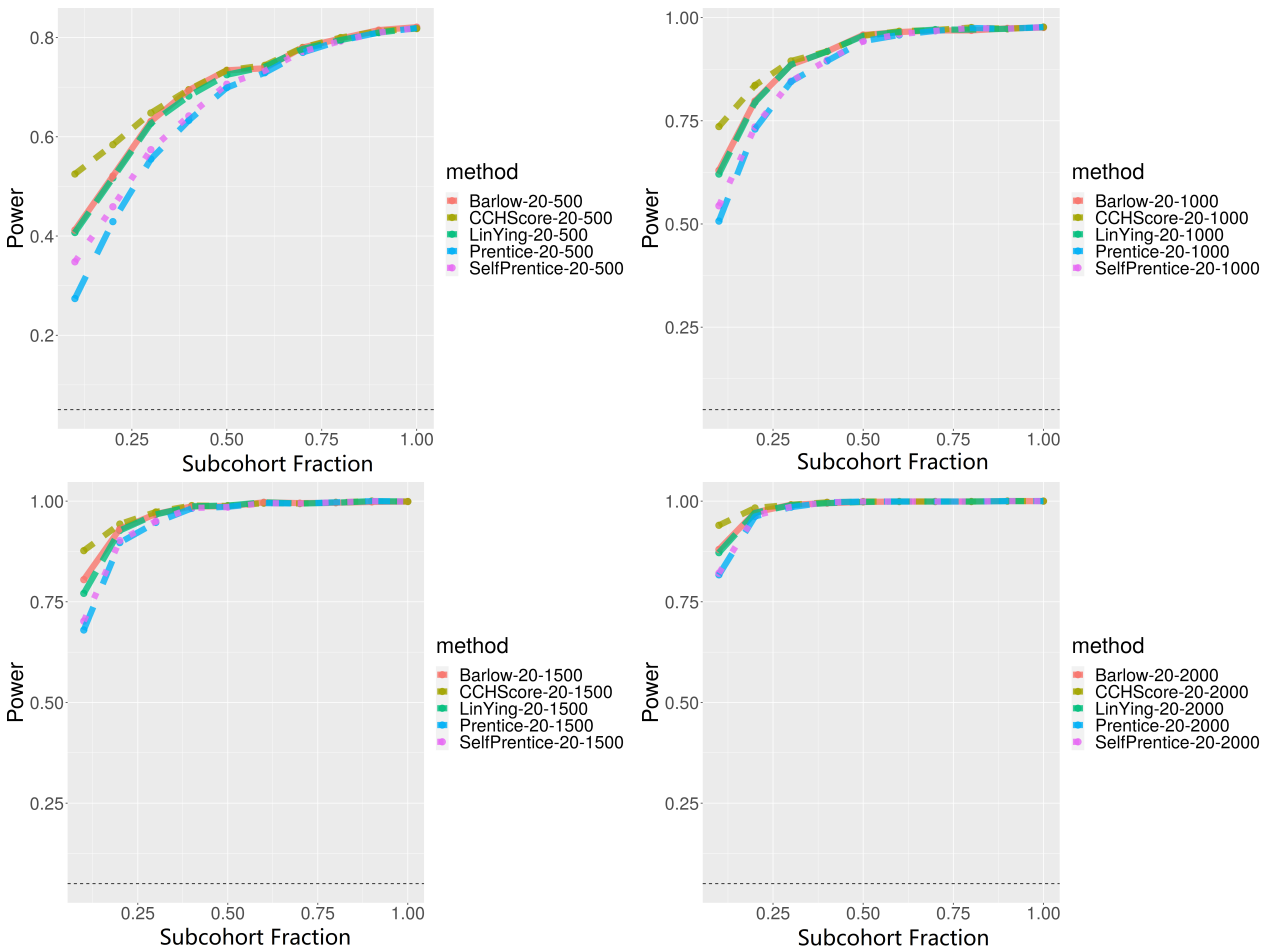


Figure 3.12: At case rate 0.20, comparing the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and full cohort size 500, 1000, 1500 and 2000, respectively. For example, "Barlow-20-1000" means at case rate 0.2, full cohort 1000, and with the "Barlow" method.

Figure 3.9 shows that at case rate 0.05 and full cohort size 500, the power of the "Barlow" method has an apparent drop down, as it fails to return a p-value 200 times out of 1000 simulation, which essentially decreases its power. "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods need to iteratively fit the PH model to estimate $\tilde{\beta}$ and have a risk to fail convergence. There are more examples in the next chapter. "CCH Score" method only involves matrix calculation, does not

need to fit the PH model to estimate β, and does not need to consider the convergence problem. Compared with other methods, it is robust and would not lose power from missing return p-values.

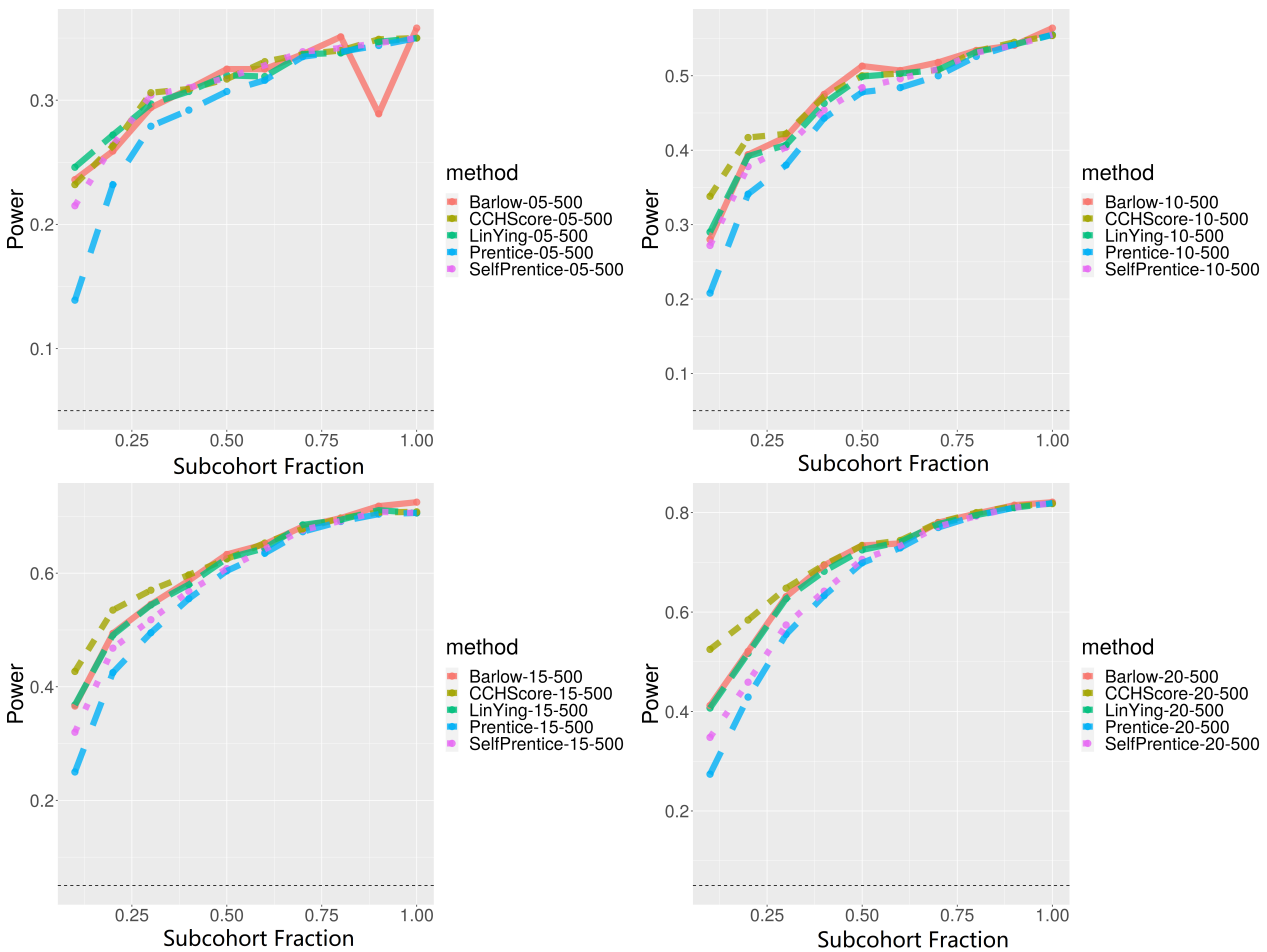


Figure 3.13: At full cohort size 500, compare the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15 and 0.2, respectively. For example, "Barlow-05-500" means at case rate 0.05, full cohort size 500, and with the "Barlow" method.

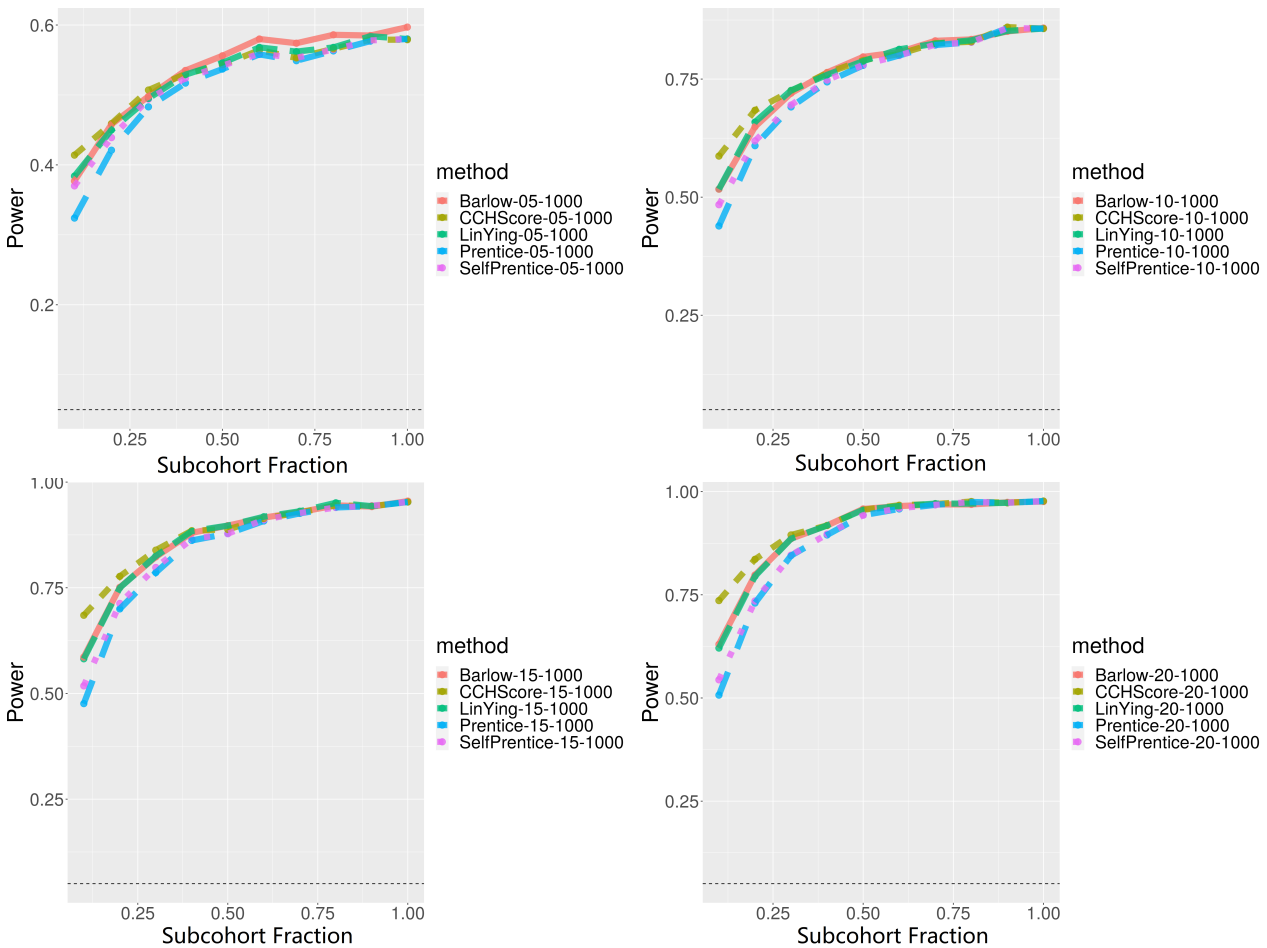


Figure 3.14: At full cohort size 1000, compare the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15 and 0.2, respectively. For example, "Barlow-10-1000" means at case rate 0.1, full cohort size 1000, and with the "Barlow" method.

Figure 3.13 shows that the results of simulated single gene data with hazard ratio 1.5-1.6 or full cohort size 500, 1000, 1500, and 2000, respectively. At full cohort size 500, the power of the "CCH Score" method will increase quickly with the increase of subcohort fraction and reach the almost same point at subcohort fraction = 1.0 (full cohort) as the same as the other four methods, "Prentice," "SelfPrentice," "LinYing," and "Barlow." The "CCH Score" performance is consistent well with the

other four existing methods. Increasing or decreasing the case rate will not influence the consistency of the five methods at a full cohort size of 500. Besides, with the increase of case rate from 0.05 to 0.2, the Power will increase. For full cohort size 1000, 1500, and 2000, Figure 3.14, Figure 3.15 and Figure 3.16 show similar results. In all, the Power of our "CCH Score" method is consistent well with existing mainstream methods. From a small sample size of 500 to a medium sample size of 2000, the Power works well at different event rates.

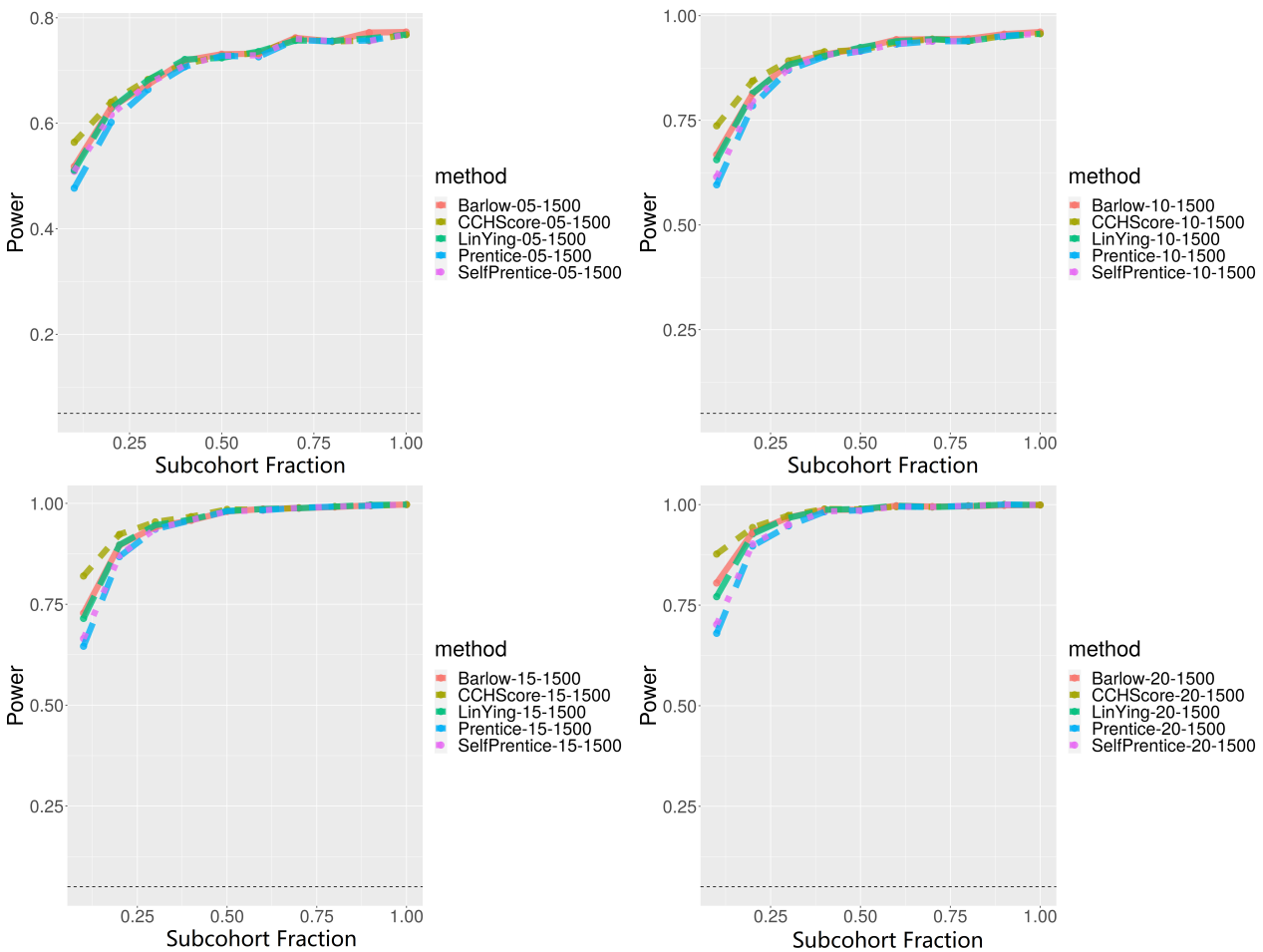


Figure 3.15: At full cohort size 1500, compare the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15 and 0.2, respectively. For example, "Barlow-05-1500" means at case rate 0.05, full cohort size 1500, and with the "Barlow" method.

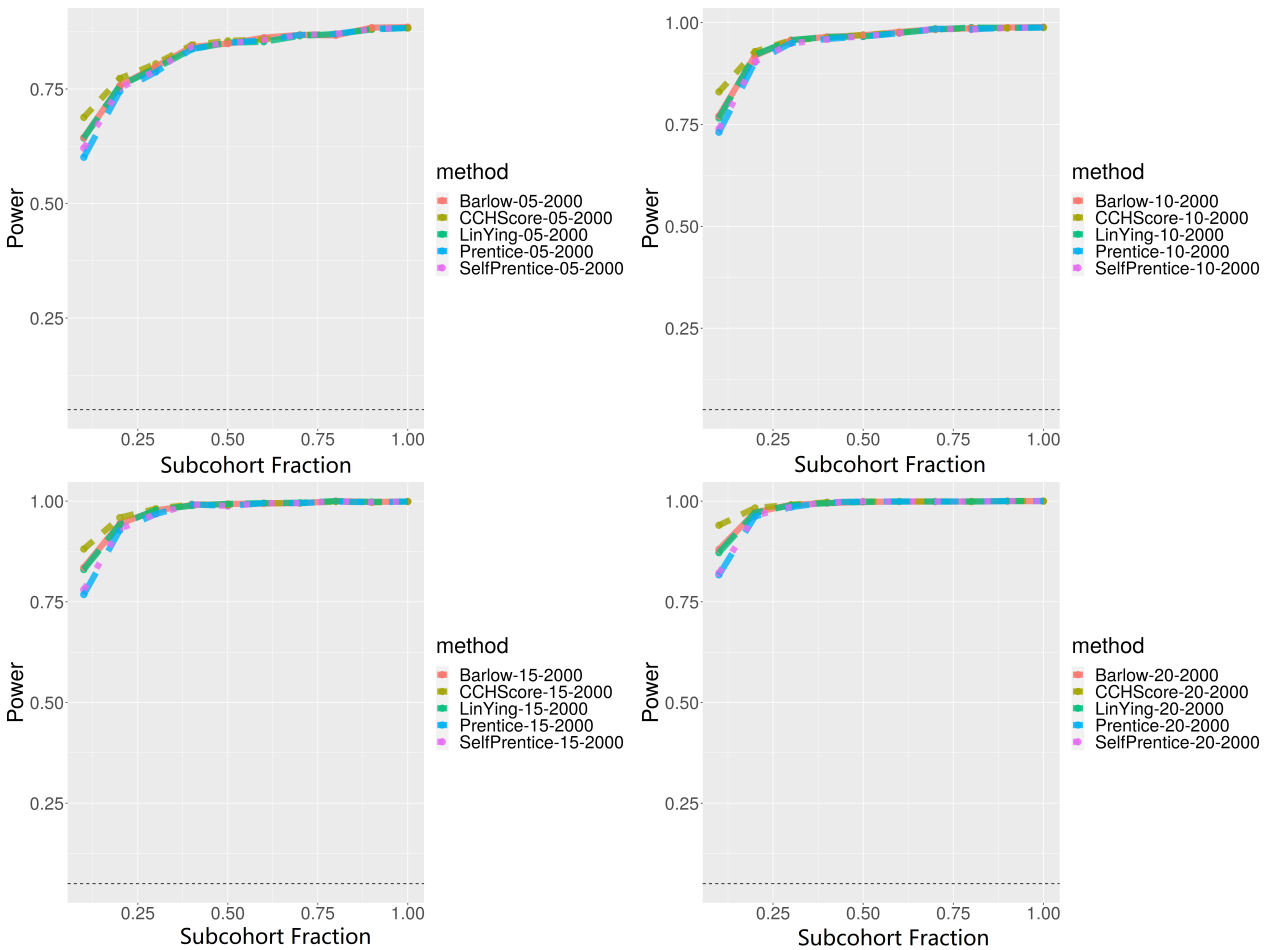


Figure 3.16: At full cohort size 2000, compare the power of the "CCH Score" method with those of "Prentice," "SelfPrentice," "LinYing," and "Barlow" methods on simulated single gene data with hazard ratio 1.5-1.6 and case rate 0.05, 0.1, 0.15 and 0.2, respectively. For example, "Barlow-05-2000" means at case rate 0.05, full cohort size 2000, and with the "Barlow" method.

To sum up, the Type I errors of the "CCH Score" method on single gene datasets are valid, which means the procedure is correct. Furthermore, the powers of the "CCH Score" method are comparable with other mainstream methods, because they have similar performance in all combinations of event rate and full cohort size. Besides, our "CCH Score" method is more robust than theirs because it does not iteratively fit the PH model to estimate $\tilde{\beta}$ and does not need to consider the conver-

gence problem. The "CCH Score test" method is computationally more robust than others.

3.3.3 False discovery rate and power on high-throughput dataset

After testing the performance of the CCH Score method on a single gene dataset, we want to test how good it is on the high-throughput dataset, which includes DEGs and NONDEGs simultaneously. For simplicity, we draw a set of perturbations to generate expression levels for DEGs (The details are in the Data Simulation section). For simulated data, we know which ones are DEGs and NONDEGs, and labeled them. Then, we used the false discovery rate (FDR) and power to measure the performance of CCH based Score test method on the simulated data.

$$\texttt{FDR} \equiv \frac{\#(\mathcal{C} \cap \mathcal{R}^c)}{\#(\mathcal{C})}, \tag{3.41}$$

where $\mathcal{C}$ is the set of DEGs identified by the "CCH Score test" method, and $\mathcal{R}$ is the set of true DEGs in the dataset. ^{c} indicates the complement of a set. The numerator is the number of NONDEGs that are falsely called significant by the CCH analysis., and the denominator is the total number of genes that are called significant (or a DEG) by the CCH analysis.

Power is defined as:

$$\texttt{Power} \equiv \frac{\#(\mathcal{C} \cap \mathcal{R})}{\#(\mathcal{R})}. \tag{3.42}$$

Power is the proportion of true DEGs that are called significant by a CCH analysis.

As shown in Figure 3.17 and 3.18, FDR of CCH Score method decreases quickly with the increase of subfraction and keep around 0.05. All five methods have similar FDR performance with the increase of subfraction.

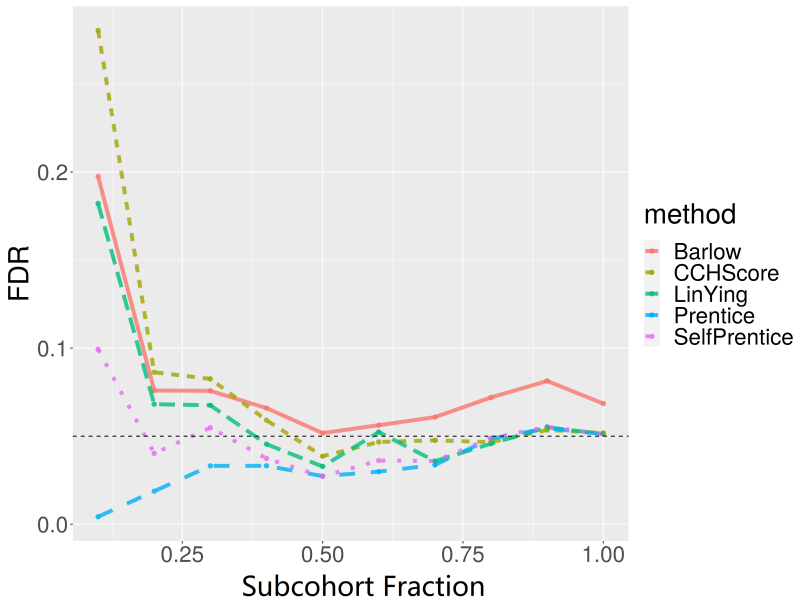


Figure 3.17: Comparing FDR of the "CCH Score" method with the other four existing methods on simulated multiple gene data with hazard ratio 1.4-1.5, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patients (full cohort).

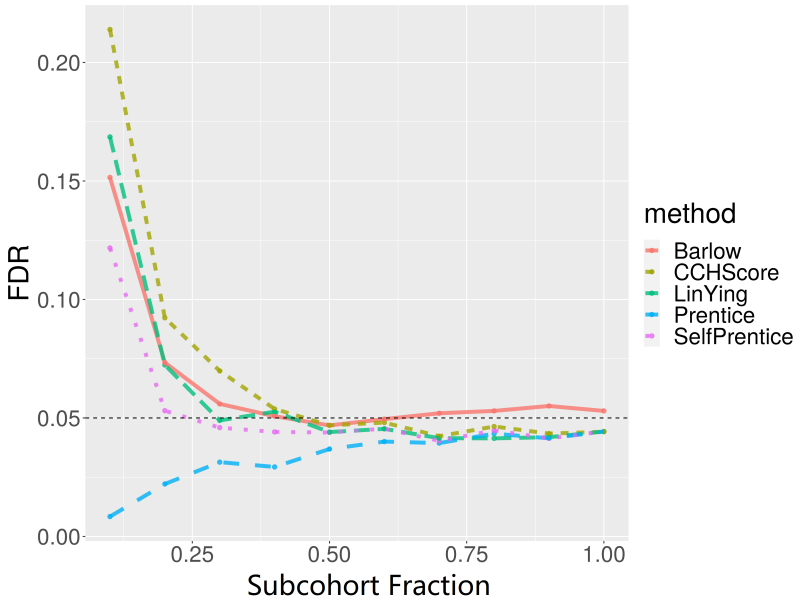


Figure 3.18: Comparing FDR of the "CCH Score" method with the other four existing methods on simulated multiple gene data with hazard ratio 1.5-1.6, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patients (full cohort).

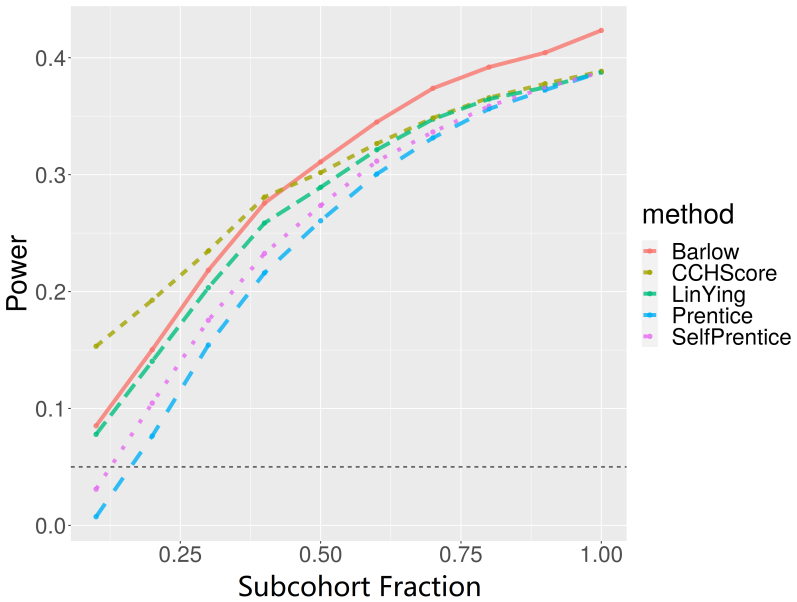


Figure 3.19: Comparing the power of the "CCH Score" method with the other four existing methods on simulated multiple gene data with hazard ratio 1.4-1.5, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patients (full cohort).

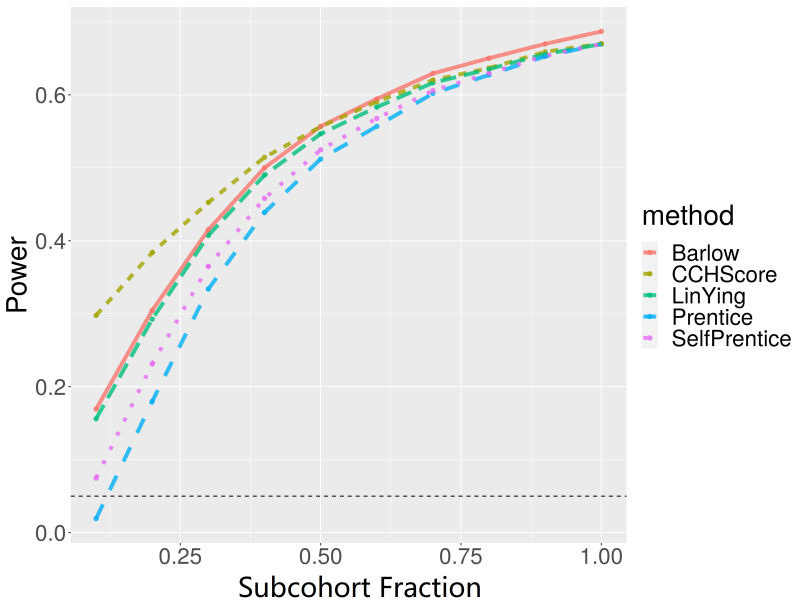


Figure 3.20: Comparing the power of the "CCH Score" method with the other four existing methods on simulated multiple gene data with hazard ratio 1.5-1.6, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patients (full cohort).

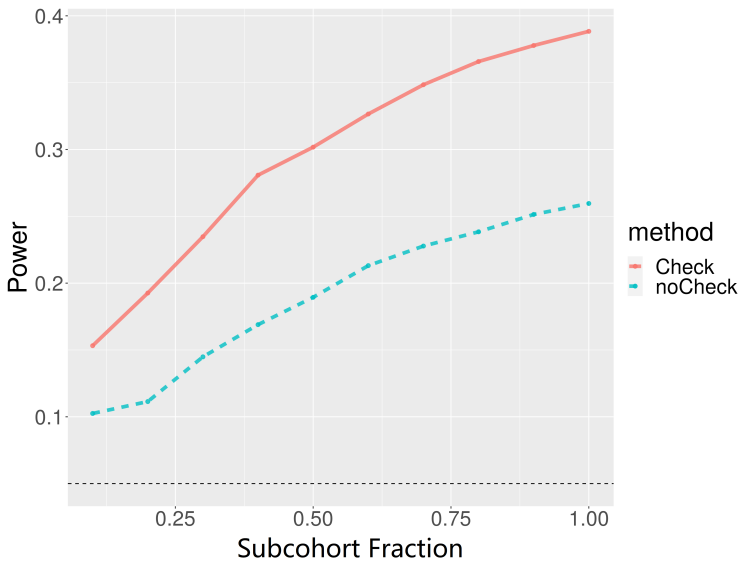


Figure 3.21: Comparing the power of "CCH Score" method on simulated multiple gene dataset with or without checking the correlation between genes' expression and survival time (hazard ratio 1.4-1.5, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patient (full cohort)).

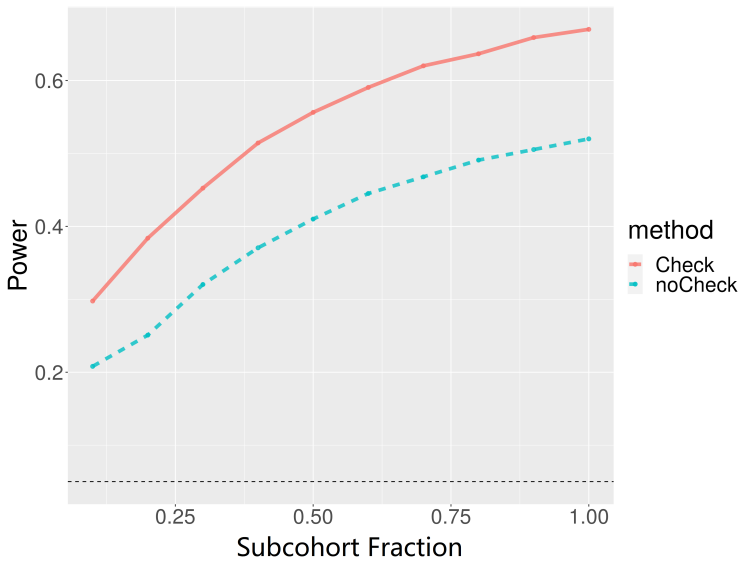


Figure 3.22: Comparing the power of "CCH Score" method on simulated multiple gene dataset with or without checking the correlation between genes' expression and survival time (hazard ratio 1.5-1.6, case rate 0.1, the proportion of DEG 10%, 2000 genes, and 1000 patient (full cohort)).

As shown in Figure 3.19 and 3.20, the Power of the "CCH Score test" method increases with the increase of subfraction, and all five methods have similar power performance in the subfraction.

When we simulate a high-throughput dataset, we should check whether, for all genes, the correlations between expression value and survival time are similar. To tell the difference with or without checking, we simulated datasets accordingly. Results show that "CCH Score" method has more power on the dataset with check than those without check (shown in Figure 3.21 and 3.22). The reason is that "DEGs" have scattered gene expression values in a not checking dataset, and some of them may be close to the random distribution of NONDEGs.

3.4 Application

Measure the Consistency between CCH and Full Cohort Analysis

We treat the full cohort analysis as "the truth" for real data problems. In other words, we treat the genes deemed significant by the full cohort analysis as "true" DEGs. So, we find a baseline to compare the CCH methods.

To evaluate the performance consistency of full-cohort and subcohort-based CCH methods, we define Type I-agreement and Type II-agreement, which are the same as "pseudo-FDR" and "pseudo-power" in John's dissertation [37]. Let $\mathcal{F}$ and $\mathcal{C}$ be the sets of genes called significant by the full cohort and subcohort-based CCH analyses, respectively. To determine the sets of significant genes for each method, we test genes one by one to get their p-values. P-values were adjusted using the BH procedure, and genes were called significant if their adjusted p-values were equal to or less than

0.05. We also define a function $\#(\cdot)$ to count the number of significant genes in the (). Type I-agreement is defined as:

$$\texttt{Type I-agreement} \equiv \frac{\#(\mathcal{C} \cap \mathcal{F}^{\mathsf{c}})}{\#(\mathcal{C})}, \tag{3.43}$$

where $^{\mathsf{c}}$ indicates the complement of a set. The numerator is the number of significant genes in a CCH analysis but are not detected by the full cohort analysis, and the denominator is the number of the significant genes in a CCH analysis.

Type II-agreement is defined as:

$$\texttt{Type II-agreement} \equiv \frac{\#(\mathcal{C} \cap \mathcal{F})}{\#(\mathcal{F})}. \tag{3.44}$$

Type II-agreement is the proportion of significant genes from the complete cohort analysis that are detected by a CCH analysis.

For a CCH method to be considered an adequate substitute for full cohort analysis, it should have low Type I-agreement and high Type II-agreement. We think the DEG found by the full cohort is "true DEG." For example, if you find 100 DEGs with the whole cohort, you see 80 "DEG" with a subcohort. 70 out of the 100 "true DEG," and the rest 10 are not. As the percentage of falsely discovered DEGs out of all genes found with a subcohort, the Type I-agreement is 10/(10+70)=12.5%. As the percentage of true DEGs found with a subcohort out of all "true DEGs" with the full cohort, the Type II-agreement is 70/100 = 70%.

Apply CCH Score method on the TCGA RNASeq Version 2 breast cancer dataset

The example features an analysis of the TCGA RNASeq Version 2 breast cancer data (BRCA). We downloaded the data on October 3, 2017 using the R/Bioconductor package RTCGAToolbox [48]. The dataset, pre-processed and normalized using the

RSEM algorithm [31], contains 1, 037 primary tumors, and the clinical data has 988 subjects with primary tumors. After removing 10 males, 71 cases whose follow-up times are less than or equal to 25 days, and 3 cases that didn't have RNA-Seq data, we left 904 subjects for our analysis. BRCA data has 16005 genes, and the whole genes were selected as a dataset. Of the 904 individuals in the RNA-seq dataset, only 115 had an event, giving an observed incidence rate of 12.7%. We choose 0.1, 0.2, ... and 0.9 as sub-cohort fraction. For the real dataset, we did simulations 100 times for each sub-cohort fraction.

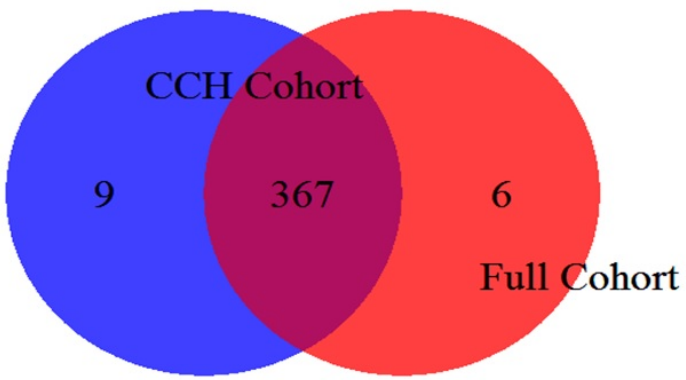


Figure 3.23: Type I_agreement and Type II_agreement on one time simulation with sampling fraction 0.9 for BRCA dataset.

The consistency between an original full cohort analysis method and the CCH-based score method was measured. As shown in Figure 3.23, for BRCA dataset, in one time simulation with sampling fraction 0.9, a full cohort analysis identified 373 DEGs, and a full reconstructed cohort identified 376 DEGs. They have 367 DEGs in common. The related "Type I_agreement" = 9/(367+9)= 2.4% and "Type II_agreement"=367/(367+6) = 98.4%.

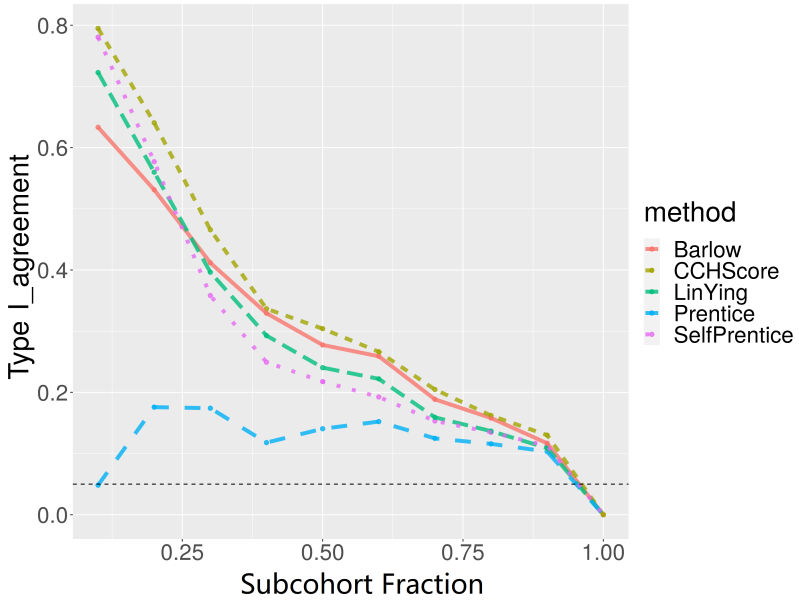


Figure 3.24: Type I_agreement on BRCA data for "Prentice," "SelfPrentice," "LinYing," and "Barlow" and "CCH Score" methods, respectively.

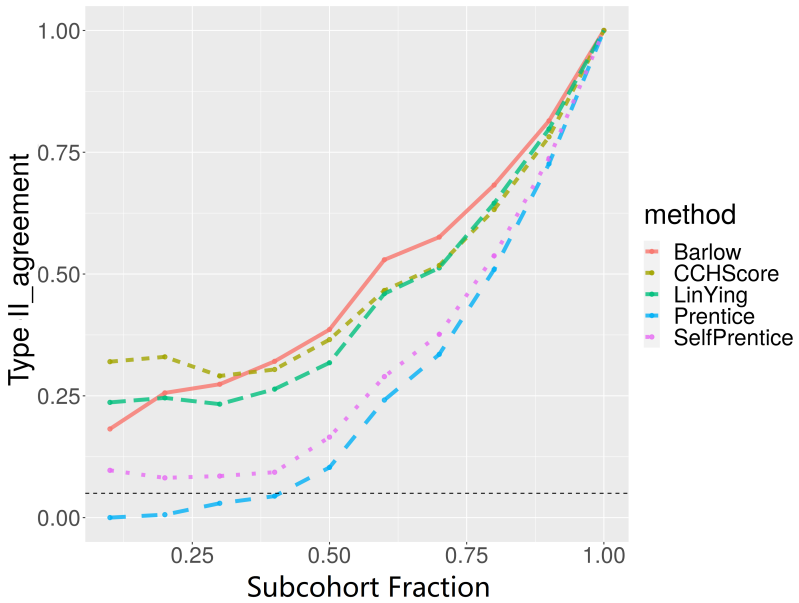


Figure 3.25: Type II_agreement on BRCA data for "Prentice," "SelfPrentice," "LinYing," and "Barlow" and "CCH Score" methods, respectively.

For a "CCHScore" method to be considered an adequate substitute for full co-

hort analysis, it should have to decrease "Type I_agreement" and increase "Type II_agreement" with the increasing of subcohort, although "Type I_agreement" has not to be below 5% and "Type II_agreement" has not to be above 95%, following the nominal 0.05 and 0.95 significance levels.

As shown in Figure 3.24, "Prentice" method has 0 values in low sub-cohort fraction, because it cannot find any DEG at low sub-cohort fraction (number of DEG = 0). It is very conservative. The other four methods have a high value at the low sub-cohort fraction, decreasing gradually with the increasing sub-cohort fraction. CCH-Score is between the highest Type I_agreement and the lowest Type I_agreement. CCHScore has comparative performance with others on "Type I_agreement."

As shown in Figure 3.25, all five methods have low values at low sub-cohort fractions. They increase gradually with the increase of sub-cohort fraction and reach almost the same value at sub-cohort fraction = 1 (Full cohort). "Prentice" method still has 0 values in low sub-cohort fraction, because it cannot find any DEG at the low sub-cohort fraction (number of DEG = 0). It is conservative. CCH Score method is between the highest Type I_agreement and the lowest Type I_agreement. CCH Score method has a comparative performance with other methods on "Type II_agreement."

Chapter 4

A Case-Cohort Design Based Permutation Test

Improving DEGs' identification has medical and biological values. One alternative method is to use a permutation test to calculate the false discovery rate (FDR), which is preferred by researchers who want a list of candidate features that contains a small number of false positives. Proposed by Fisher [20] and Pitman [38], this method has been applied on biostatistics and quantized Data Analysis ([59] and [11]). A permutation test is a non-parametric statistic that obtains the p-value from the sample-specific permutation distribution of that statistic rather than from the theoretical distribution derived from the parametric assumption. Permutation procedures were applied on estimating false discovery rate (FDR) ([58] and [56]). When permutation tests are used on DEGs' identification, strong semi-parametric assumptions and probability distribution assumptions for p-value do not need to be concerned. Under the null hypothesis of the permutation test (The genes under study are not DEGs), their FDR is obtained by averaging a large number of possible FDR values calculated from related rearrangements of the observed data.

The basic assumption of a permutation test is that the data of the responsible variable must be exchangeable under the null hypothesis ([20], and [38]). In a CCH of survival data, the values of the response variable, survival times, exist. However, the permutation test can not be directly applied to a CCH sample, because the CCH sample is not a random sample and the gene expression values of controls out of the subcohort are missing. We want to find a new procedure to identify DEGs, which can save the benefits of both the CCH and the permutation test. We propose reconstructing the entire cohort by imputing the missing gene expression values outside the subcohort and performing the permutation test on the re-constructed whole cohort. Random sampling with replacement will be used to reconstruct entire cohorts.

In this work, we proposed a permutation-based score test procedure and applied it to high-throughput gene differential expression analysis under CCH design. Firstly, we reconstructed the full cohorts by imputing the missing data with resampling with the replacement method. Then, we performed the CCH-based permutation test on the reconstructed full cohort to identify the DEGs associated with survival outcomes. To illustrate the usage and advantages of this method, we evaluated our testing procedures through simulation studies with datasets generated based on PH and non-PH models, respectively. Besides, we applied the proposed approach to the same RNA-sequencing data set (BRCA) and a microarray data set (ALL) from a leukemia study. Furthermost, we compared the proposed method with some existing CCH methods that can also be applied to the gene expression analysis under the CCH Design.

4.1 Rebuilding full cohorts from a CCH sample

The CCH study design is a prospective observational study design that blends the economy of case-control studies with the philosophical soundness of cohort studies.

CCH designs consider a random sample of the whole cohort, called a subcohort. At the time of analysis, add all cases outside the subcohort to the sample. In other words, a CCH sample consists of all cases (both in and out of the subcohort) but only the controls in the subcohort. A CCH analysis is best suitable for data that is cheap to collect but expensive to analyze or process.

Suppose there are p genes and n patients in a full cohort gene dataset. As shown by the right diagram at the right of Figure 4.1, a CCH gene dataset is a subset of the full cohort, and the subcohort was randomly chosen from the full cohort. Therefore, the gene expression values outside the CCH are missing at random. We impute these missing gene expression values by replacing them with the data obtained from sampling with replacement from the subcohort. In this way, we can form a new full cohort data set (as shown in the right diagram in Figure 4.1) for the permutation test.

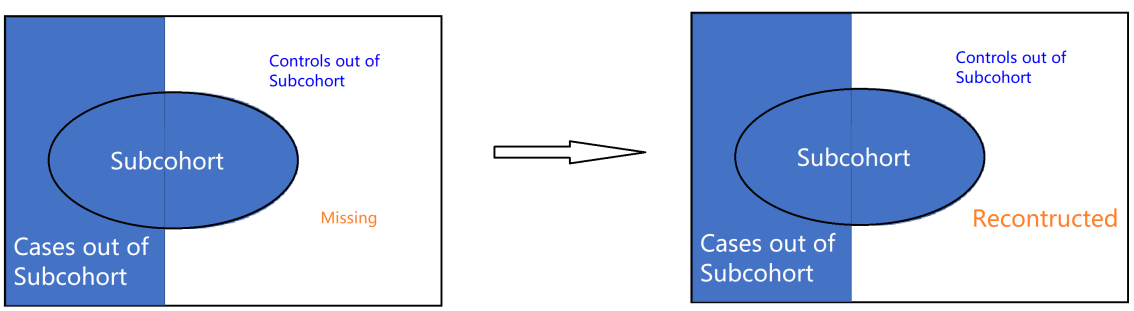


Figure 4.1: A CCH sample from a related full cohort.

More specifically, for gene i, the observed data in the full cohort/sample is (t_j, δ_j, x_{ij}) for patient j, where $t_j = min(T_j, C_j)$. T_j is the true survival time and C_j is the censored time (refer Chapter 2.1). As shown in figure 4.2, missing expression values are imputed by resampling with replacement to form a full cohort in this project. To explain the "reconstructed full cohort" method, we must consider the relationship between gene expression and survival times. In reconstructed pairs, survival times are unchanged because they come from the times of controls

in the whole original cohort, and missing gene expression values are resampled with replacement from a subcohort randomly. We put an expression value and a survival time together in a pair by chance. Under the null hypothesis, the expression value and time have no relationship in the original cohort, "$(T_j, C_j)|x_{ij} = (T_j, C_j)$". A subcohort is a random sample of the original cohort, and its expression value and time should also have no association. As the missing gene expressions are resampled from the subcohort randomly, they should also have no association with time, which will keep the type I error or FDR at the corresponding nominal levels.

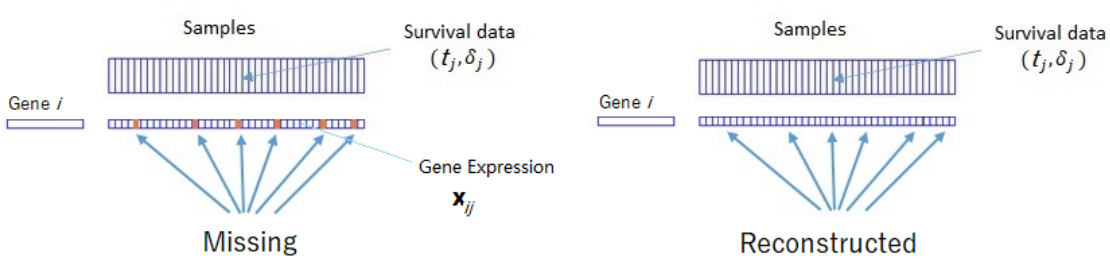


Figure 4.2: Imputing one gene expression from a CCH to a full cohort.

In computation, the pairs of expression value and time are created by chance. When they are associated, this will increase type I error. On the other hand, when they have no relationship, this will not impact type I error. When the sub-cohort fraction is large enough, the number of expression values resampled with replacement is small. So the number of pairs with an association is small. When most expression and time pairs have no relationship, the type I error should keep at the corresponding nominal levels in the permutation test. The power can be evaluated in simulation studies.

4.2 A Proposed Permutation Score Test

Scoring the reconstructed full cohort

Similar with the censored survival full cohort permutation tests, reconstructed full cohorts of censored survival data were scored in this project. The definition of the score S_i for gene i is:

$$S_i = \frac{rc_i}{sc_i + sc_0}, \tag{4.1}$$

where rc_i is the numerator of the score, sc_i is a standard deviation and sc_0 is an exchangeability factor. We use $sc_0 = 0$ for simplification. For censored survival data, rc_i is defined as

$$rc_i = \sum_{k=1}^{K}(x^*_{ik} - d_k \bar{x}_{ik}). \tag{4.2}$$

And sc_i is defined as

$$sc_i = [\sum_{k=1}^{K}((\frac{d_k}{m_k})\sum_{j \in R_k}(x_{ij} - \bar{x}_{ik}))^2]^{1/2}, \tag{4.3}$$

where x_{ij} is the expression value of gene i for patient j. x_{ij} is either from CCH or be imputed. As the same as the original full cohort analysis, k be the indices of the K unique death times z_1, z_2, ..., z_K, and R_1, R_2, ..., R_K be the indices of the observations at risk at these unique death times, that is $R_k = \{i : t_i \geq z_k\}$. Let $m_k = \#inR_k$. Let d_k be the number of deaths at time z_k. $x^*_{ik} = \sum_{t_j = z_k} x_{ij}$ and $\bar{x}_{ik} = \sum_{j \in R_k} \frac{x_{ij}}{m_k}$.

As shown in Appendix B, $rc_i = r_i + \sum_{k=1}^{K}(-d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k})$, where r_i is the "score process" of the original full cohort and is defined in equation 2.23. As $E(\sum_{k=1}^{K}(-d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k})) = 0$ under the assumption that e_{ij} has mean 0 and finite variance σ_e^2 for patient j outside of CCH, $E(rc_i) = r_i$. Under the null hypothesis,

the value of r_i is close to 0. The value of rc_i will also fluctuate around 0 under the large number theory. And under the alternative hypothesis, the absolute value of r_i is large. The value of rc_i will also fluctuate around the value of r_i under the large number theory. $Var(rc_i)$ is finite (equation B.7).

As $S_i = \frac{rc_i}{sc_i + sc_0}$ and $E({sc_i}^2) \geq {s_i}^2$ (shown in Appendix B), the expected score value of reconstructed full cohort method should be smaller or equal to that of original full cohort method, which may make the power of the former smaller or equal to the power of the latter. This relationship makes sense because the former only uses a part of the original data, while the latter method uses all the original data.

A Score Based Permutation Test

In this project, the null hypothesis of a permutation test is that the gene under testing is not a DEG. As shown in Figure 2.2, N times permutation were done for each gene. With the scores of each permuted reconstructed full cohort, we can identify DEGs by a permutation test. In DEGs, gene expression of n patients is related to survival time, and this relationship will be broken after permuting the gene expression. The difference in scores before and after permutations will be tested for each gene with FDR or p-value method.

Identify DEGs with FDR

We define FDR and identify DEGs as the same as that in SAM [58]. Please see 2.1.5 as a reference.

Identify DEGs with p-values

To determine the sets of significant genes for a full cohort, we test genes one by one to get their p-values by equation 4.4. The null hypothesis is that the gene is not a DEG. P-values were adjusted using the BH procedure, and genes were called significant if their adjusted p-values were equal to or less than 0.05.

$$p - value(i) \equiv Pr(S \geq S_{obs(i)}|H_0) \equiv \frac{\#(S^{(j)}_{perm(i)} > S_{obs(i)})}{N}. \tag{4.4}$$

Measuring the consistency between rebuild full cohorts and the original full cohort

In a simulation study, we simulated DEGs and non-DEGs and labeled them differently. Type I-error and power can be used to measure the performance of our method on a single gene, while FDR and power can be used to measure the performance of our approach on a high-throughput dataset.

For data from real problems, we treat the initial full cohort analysis as "the truth." We assume the genes deemed significant by the full cohort analysis as "true" DEGs. So we have a baseline to which we compare the CCH methods. To evaluate the performance consistency of full-cohort and subcohort-based CCH methods (reconstructed full cohort method) for real data, we define Type I-agreement and Type II-agreement as the same as those in Chapter 3 (Refer 3.4 of Chapter 3).

4.3 Simulation Studies

4.3.1 Simulation of proportional hazard model data

Data Simulation Method

The data simulation method is the same as the data simulation method of the PH model in Chapter 3. Please refer to that for details.

Comparing methods of rebuilding full cohorts

Suppose we have an original full cohort and a CCH from the original full cohort. We proposed and compared three reconstructing full cohort methods, "CCH," "subcohort," and "controls in CCH," and want to figure out which one has the best performance on power and type I error. As their names indicating, they reconstruct an entire cohort by resampling with replacement data from a CCH, a subcohort, or only the controls in the CCH, respectively. To compare them with the random sampling method, a "random sample" method randomly selects patients from the original cohort according to the expected number of patients in the CCH.

As shown in Figure 4.3, 4.4, 4.5 and 4.6, all reconstructing methods have power higher than that of "random sample" method. "CCH" method has the lowest (best) Type I error and FDR, but its power is the lowest among the reconstructing methods. "subcohort" method resamples in a random sample, while "controls in CCH" resamples in the controls of CCH, which is not a random sample. The latter's FDR is slightly higher than 0.05 in large sub-fraction of full cohort and decreases slowly, although its power is higher than that of the "subcohort" method. So, we choose the "subcohort" method as the best method.

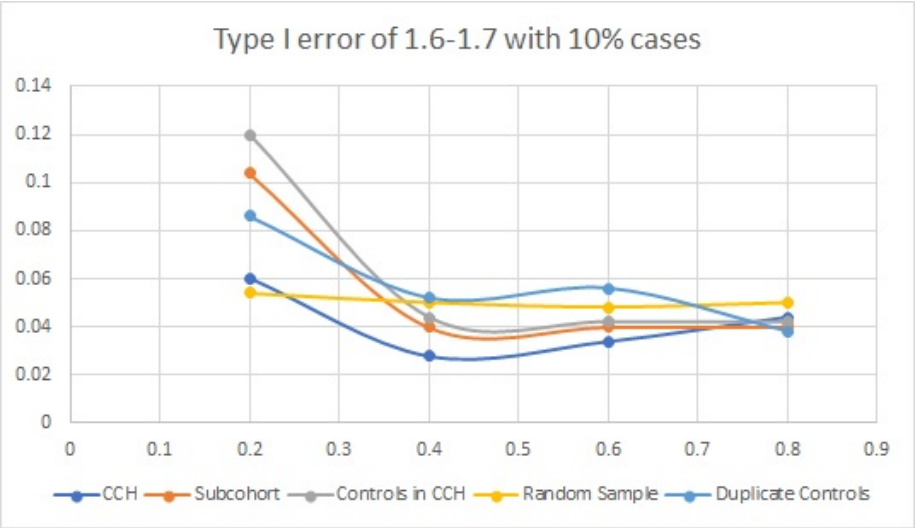


Figure 4.3: Type I error of reconstructing full cohort methods on simulated single gene data with hazard ratio 1.6-1.7, full cohort size 1000, and case rate 0.1.

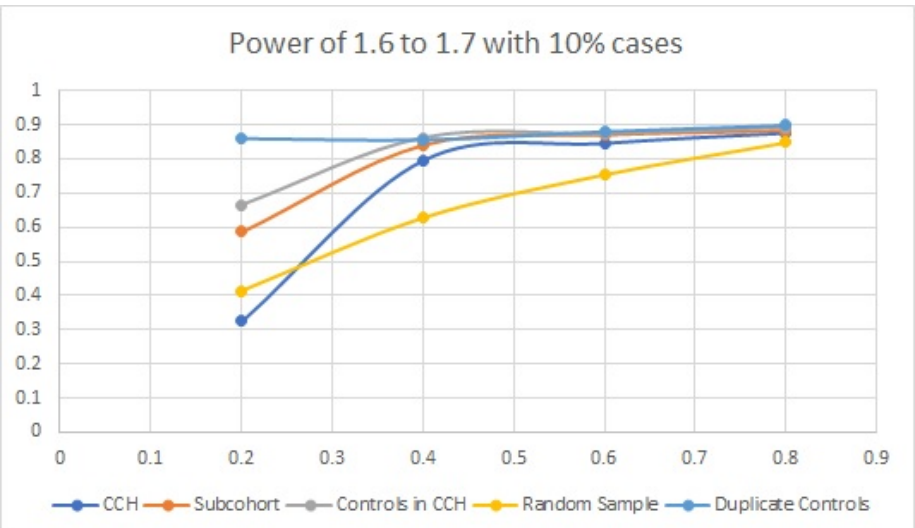


Figure 4.4: Power of reconstructing full cohort methods on simulated single gene data with hazard ratio 1.6-1.7, full cohort size 1000, and case rate 0.1.

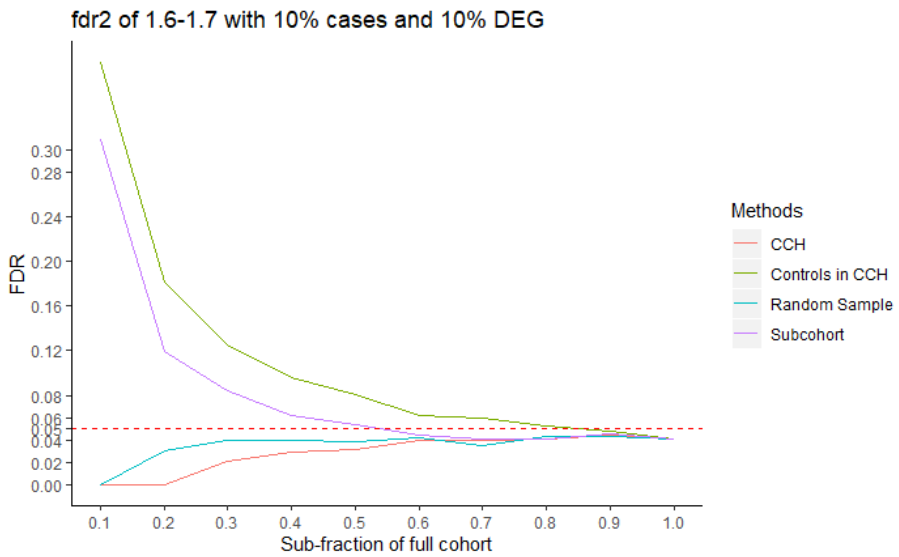


Figure 4.5: FDR of reconstructing full cohort methods on simulated high throughput gene data with hazard ratio 1.6-1.7, full cohort size 1000, genes 2000, case rate 0.1, and DEG 10%.

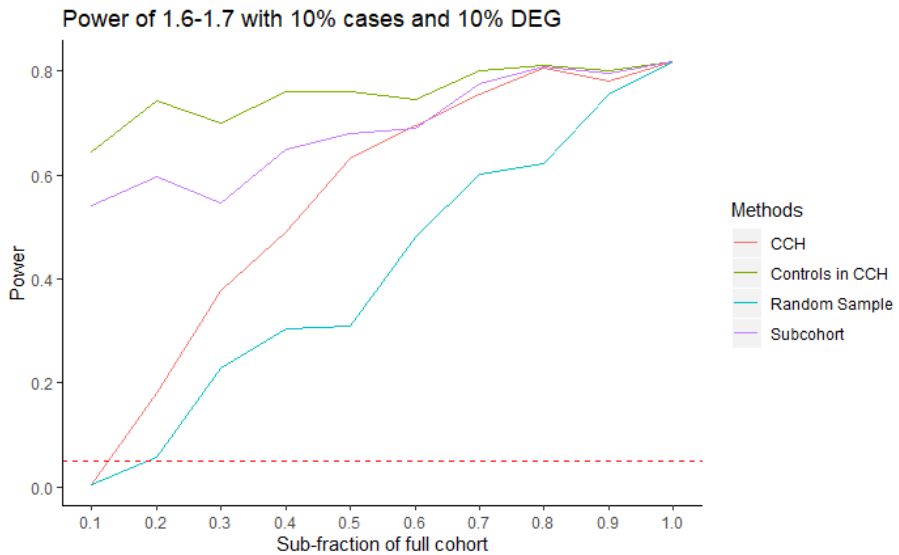


Figure 4.6: Power of reconstructing full cohort methods on simulated high throughput gene data with hazard ratio 1.6-1.7, full cohort size 1000, genes 2000, case rate 0.1, and DEG 10%.

Comparing "subcohort" method with existing CCH methods

There are four well know existing CCH methods, "Prentice," "SelfPrentice," "LinYing," and "Barlow." We want to compare the performance of "subcohort" with them.

As shown in Figure 4.7 and 4.9, the "subcohort" method has high Type I error and FDR at the low sub-fraction, but they decrease quickly and become valid when sub-fraction of the full cohort is large enough. Besides, when the subcohort is large enough and the percentage of cases in patients is not low, or the number of cases in a dataset is not very rare, the "subcohort" method has very similar performance to other existing methods. As shown in Figure 4.8 and 4.10, all five ways have comparable Power when the sub-fraction of a full cohort is large enough.

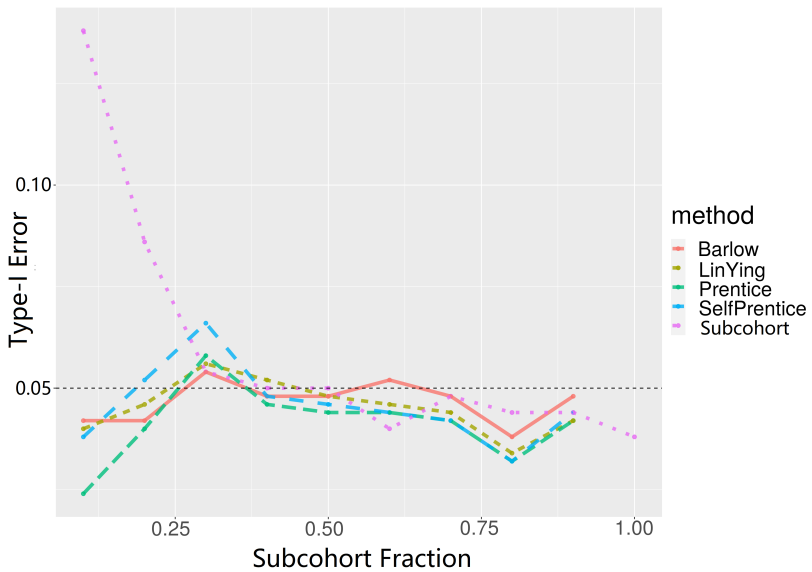


Figure 4.7: Type I error of subcohort and other existing four methods on simulated single gene data with hazard ratio 1.6-1.7, full cohort size 1000, and case rate 0.1.

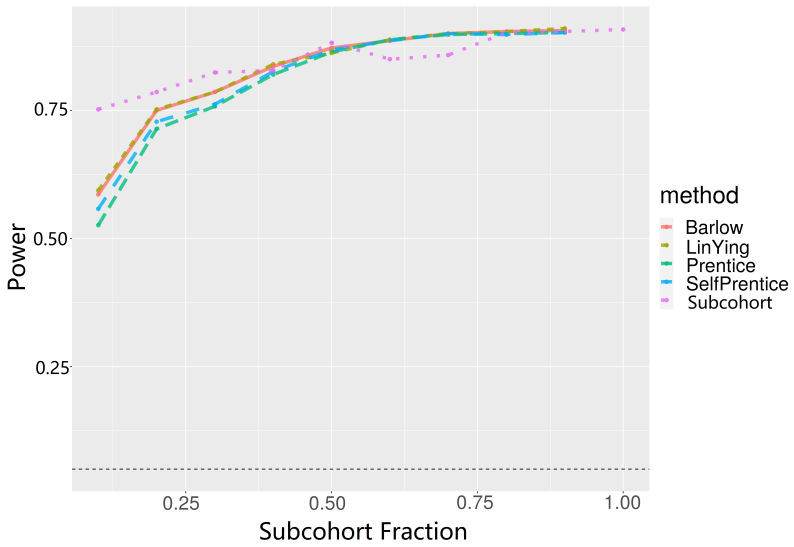


Figure 4.8: Power of subcohort and other existing four methods on simulated single gene data with hazard ratio 1.6-1.7, full cohort size 1000, and case rate 0.1.

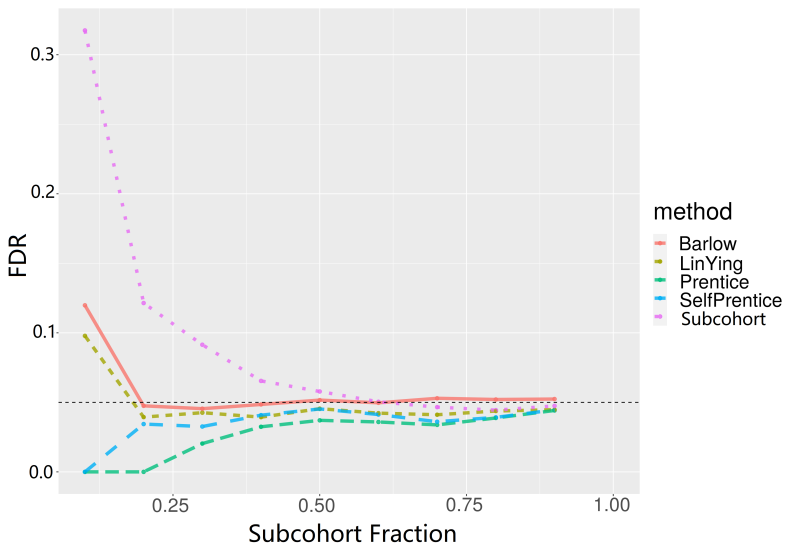


Figure 4.9: FDR of subcohort and other existing four methods on simulated multiple gene data with hazard ratio 1.6-1.7, full cohort size 1000, genes 2000, case rate 0.1 and DEG 10%.

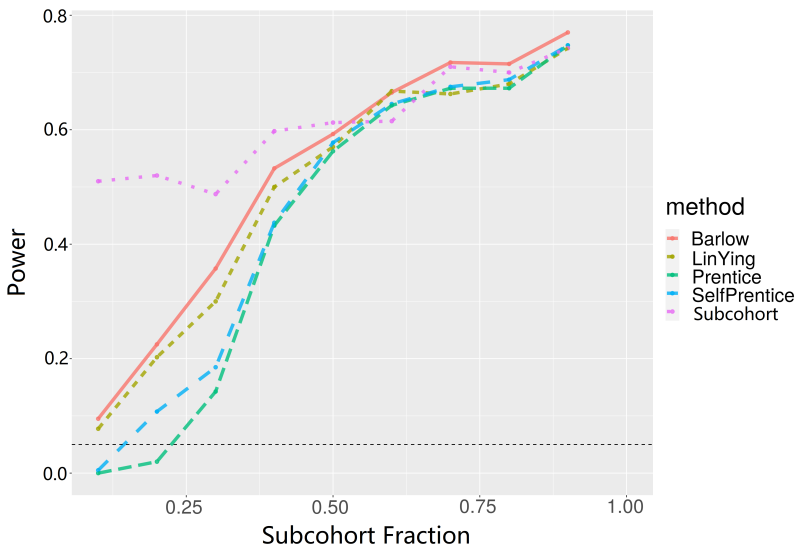


Figure 4.10: Power of subcohort and other existing four methods on simulated multiple gene data with hazard ratio 1.6-1.7, full cohort size 1000, genes 2000, case rate 0.1 and DEG 10%.

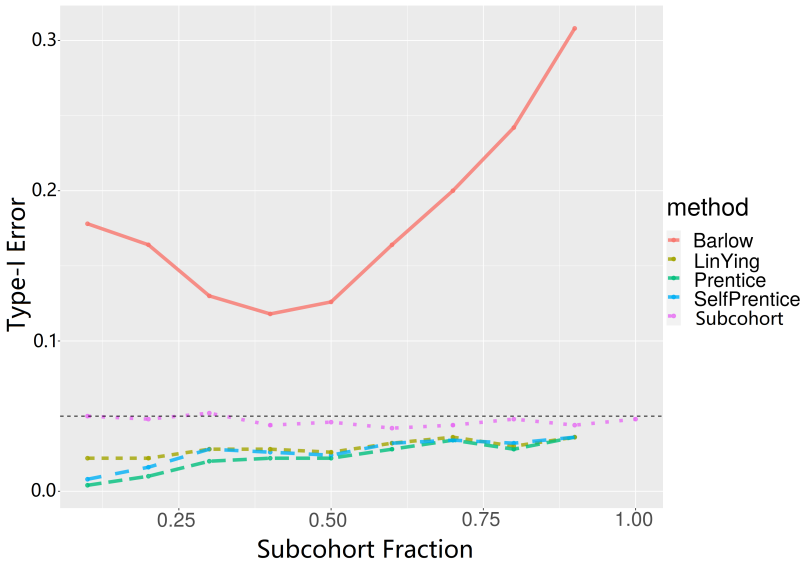


Figure 4.11: Type I error of subcohort and other existing four methods on simulated single gene data with hazard ratio 1.9-2.0, full cohort size 1000, and case rate 0.002.

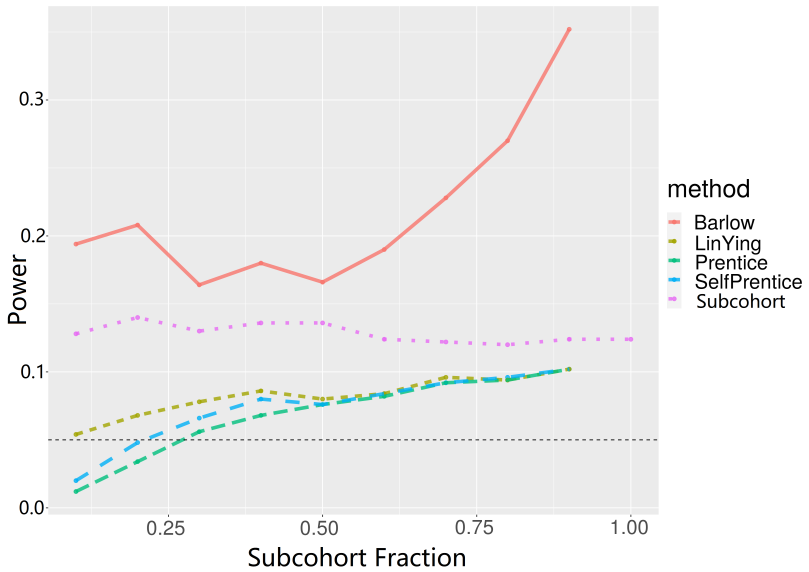


Figure 4.12: Power of subcohort and other existing four methods on simulated single gene data with hazard ratio 1.9-2.0, full cohort size 1000, and case rate 0.002.

However, for example, when there are 1000 patients and only 5% of them are cases, there are only 50 cases in the full cohort. The "Barlow" method begins to infinite Type I error. And when the percentage of cases decreases to 0.2%, the number of patient events is scarce, only 2 or 3. As shown in Figure 4.11 and Figure 4.12, "Prentice," "SelfPrentice," and "LinYing" methods can keep type-I error around 0.05, but they have low Power. Although the "Barlow" method has relatively high Power, it cannot stay type-I errors around 0.05, and this means this method is not correct in the situation. Compared with the existing four methods, our approach can keep Type I error at approximately 0.05 while having relatively high Power, which is an advantage of our "subcohort" method over the other four methods if people need to study rare event diseases.

4.3.2 Simulation 1 of non-proportional hazard data

In a proportional hazards model, the hazard ratio remains constant from the beginning to the end of the study. In practice, this does not occur for most medical interventions. Stensrud et al. [55] discussed three scenarios regarding proportional hazards. They are "no immediate effect," "immediate and delayed effects in opposite directions," and "variations in disease susceptibility" which were illustrated by three articles previously published in the Journal of the American Medical Association (JAMA). These scenarios showed that hazards are not proportional when the treatment effect changes over time. For example, the hazard ratio was 1.8 during the first year and 0.70 after five years of follow-up. The overall hazard ratio of 1.24 from a Cox proportional hazards model was a weighted average, which needs to be interpreted as a weighted average of the actual hazard ratios over the entire follow-up period.

When hazard rates are non-proportional, the power is lost for both log-rank & Cox PH tests. Log-rank is no longer the most powerful test, and the score test based on the Cox model is no longer the best partial-likelihood statistics. Stensrud [55] implied that statistical tests for non-proportional data are unnecessary because it varies over the follow-up period, and tests of proportional hazards yielding high P-values are probably underpowered. Besides, the related problems, such as an incorrect standard variance estimator will be reported when the statistical model includes covariates other than the treatment group indicator, or the magnitude of the Cox hazard ratio depends on the distribution of censoring, should be overcome by estimating valid 95% confidence intervals with resampling with replacement methods and by estimating an inverse probability weighted hazard ratio.

Data Simulation Method

$$x_{ij} = \frac{-log\left(\frac{-\lambda t_j^{\rho}}{log(u_j)}\right)}{\beta_i}. \tag{4.5}$$

$$x_{ij} = x_{ij} \times ratio. \tag{4.6}$$

$$x_{ij} = x_{ij} + e_{ij}. \tag{4.7}$$

The scenarios talked above have two different hazard ratios over the entire follow-up period. In other words, they are nonproportional hazard data. To simulate a dataset with two different hazard ratios, we can combine data from two parts to create a whole cohort. For example, we want to simulate the hazard ratio as 1.8 at the start and 1.2 after three years. We generate survival time first. Then use equation 4.5 to calculate single gene expression values. We change the values related with after three years by multiplying the ratio with equation 4.6, where the ratio is 1 for the first years and is 1.8/1.2 after that. For multiple genes, add random values for gene expression got from equation 4.6.

Type-I error and power for DEGs' identification in single gene datasets

For a dataset with two different hazard ratios, there is one hazard ratio for the first three years and another afterward. When the subcohort is large enough and the percentage of cases in patients is not low, or the number of cases in the dataset is not very rare, the six methods have a very similar performance of type-I error and power for DEGs' identification in single gene datasets of the first type of non-PH model.

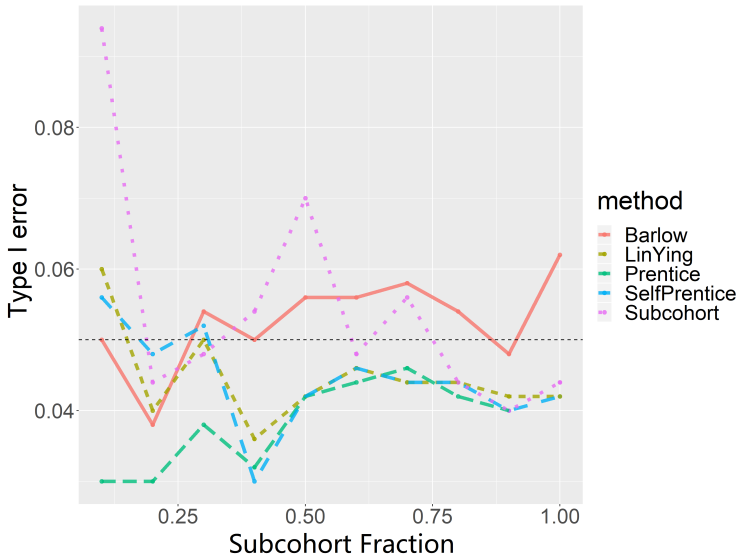


Figure 4.13: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.05. the hazard ratio is 1.8 at the start and 1.2 after three years.

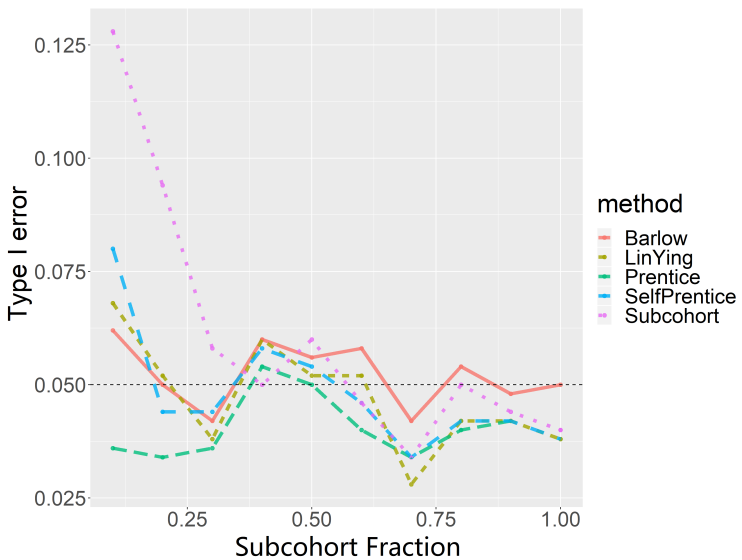


Figure 4.14: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.1. The hazard ratio is 1.8 at the start and 1.2 after three years.

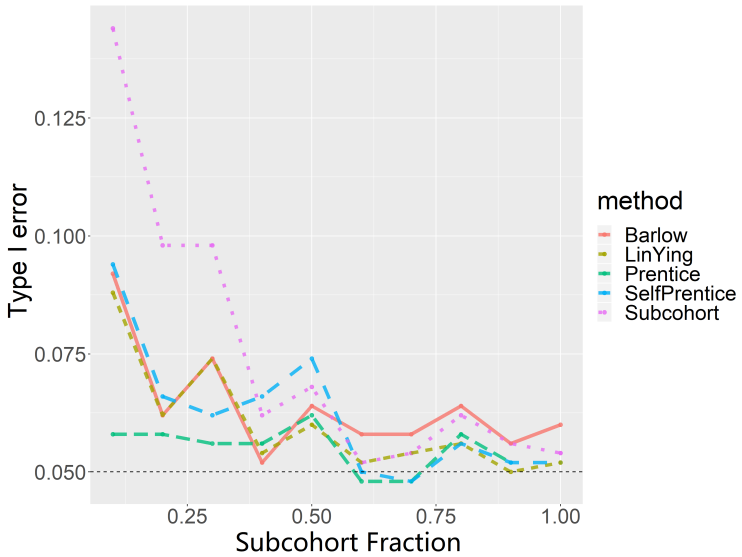


Figure 4.15: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.15. The hazard ratio is 1.8 at the start and 1.2 after three years.

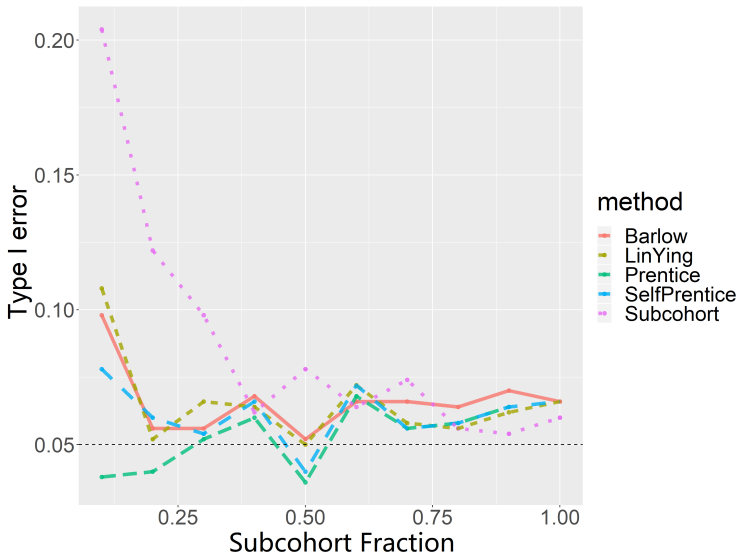


Figure 4.16: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.2. The hazard ratio is 1.8 at the start and 1.2 after three years.

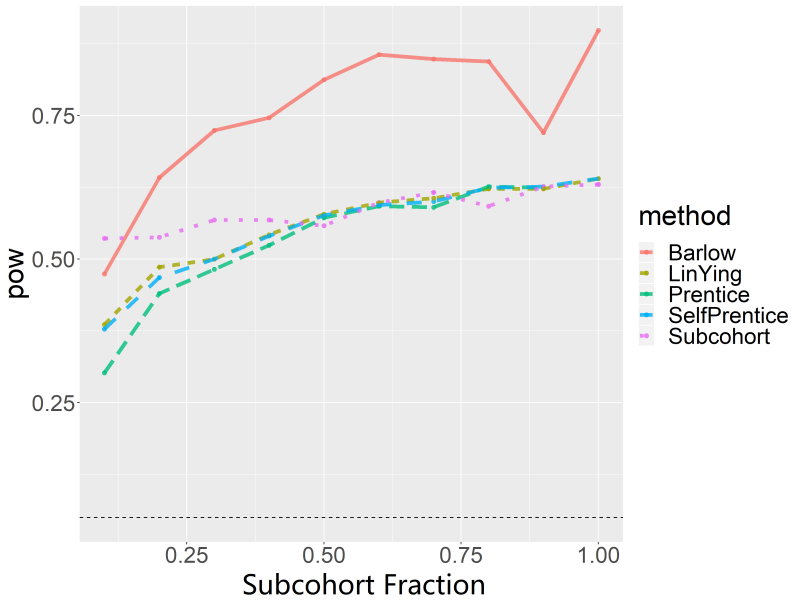


Figure 4.17: Power of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.05. The hazard ratio is 1.8 at the start and 1.2 after three years.

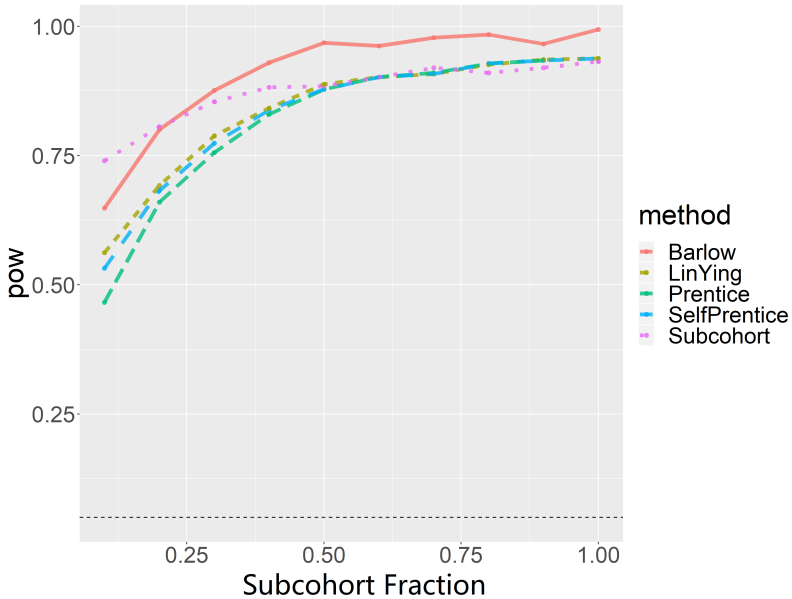


Figure 4.18: Power of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.1. The hazard ratio is 1.8 at the start and 1.2 after three years.

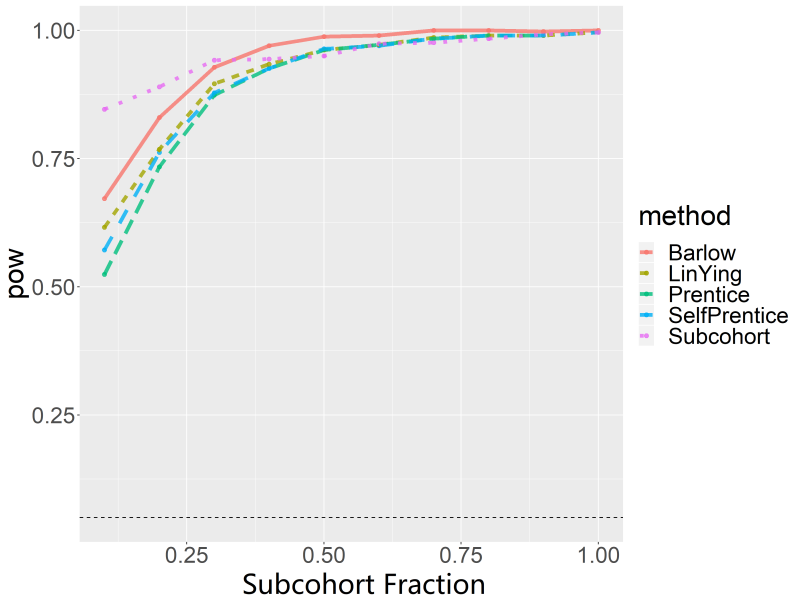


Figure 4.19: Power of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.15. The hazard ratio is 1.8 at the start and 1.2 after three years.

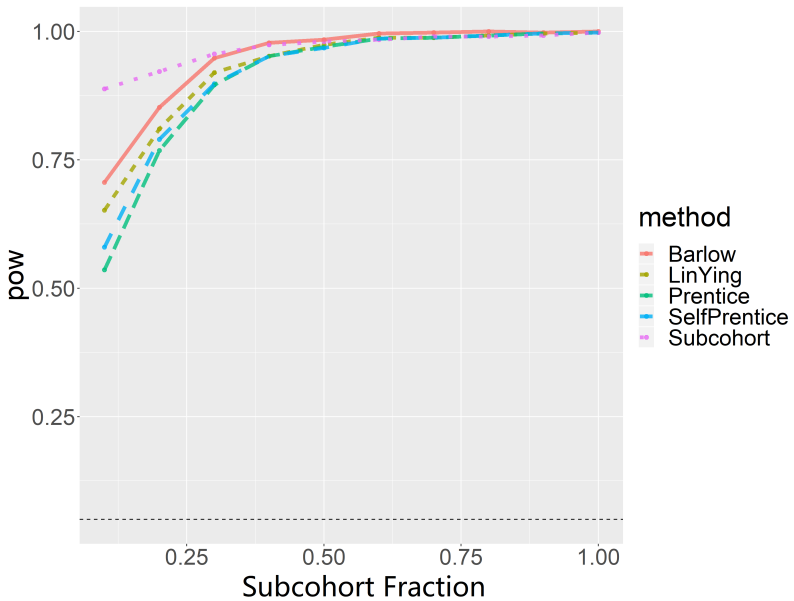


Figure 4.20: Power of subcohort and other existing four methods on simulated single gene data with full cohort size 500 and case rate 0.2. The hazard ratio is 1.8 at the start and 1.2 after three years.

As shown in Figure 4.13 to 4.16, the type I errors of five methods are comparable, although in case rate 0.05 and 0.1, the type I error of "Barlow" method is higher than other four methods. But when the case rate increases to 0.15 or 0.2, the type I error curves of the five methods become very similar in the middle and high subcohort fraction range.

The situation is the same when we compare the "power" of the five methods. We found that in the case of rates 0.05 and 0.1, the power of the "Barlow" method is higher than the other four methods. But when the case rate increases to 0.15 or 0.2, the power curves of five methods almost overlap in middle and high subcohort fraction range (Shown in Figure 4.17 to 4.20).

FDR and power for DEGs' identification in high-throughput gene datasets

High-throughput gene datasets include multiple genes. We used two methods to consider the multiple comparisons. The first used p-value BH-adjustment, and the second estimated FDR as the same as that of chapter 3.

For the first type of non-PH data, the six methods have similar performances of FDR on high-throughput datasets. When the subcohort is large enough and the percentage of cases in patients is not low, or the number of cases in a dataset is not very rare, the "subcohort" method performs very similarly to other existing methods. Although in case rate of 0.05, "Prentice" and "SelfPrentice" are better than the other four (Shown in figure 4.21). But when the case rate increases to 0.1, 0.15, or 0.2, the FDR curves of the six methods become very similar in the middle and high subcohort fraction range. As shown in Figure 4.22 to 4.24, their FDR is comparable.

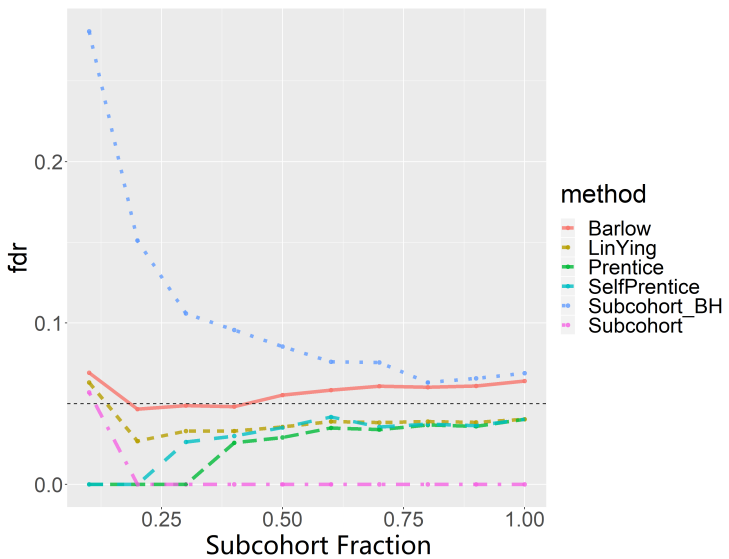


Figure 4.21: FDR of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.05. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset, and 10% are DEGs.

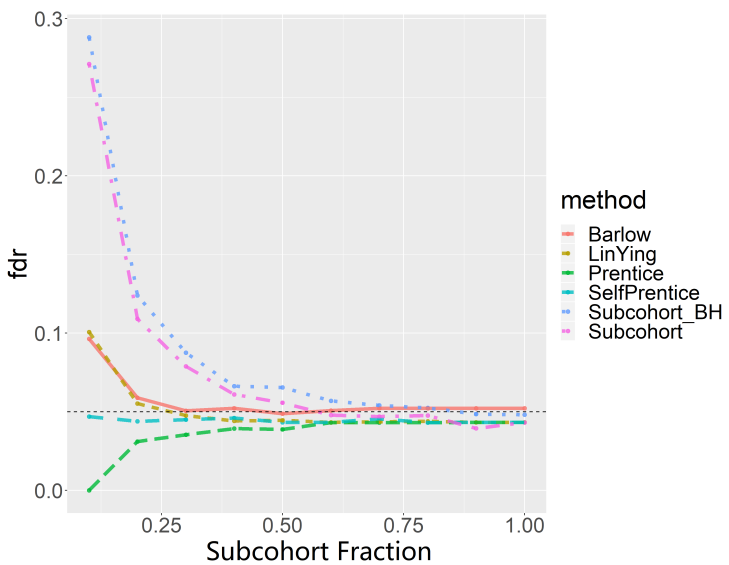


Figure 4.22: FDR of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.1. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset, and 10% are DEGs.

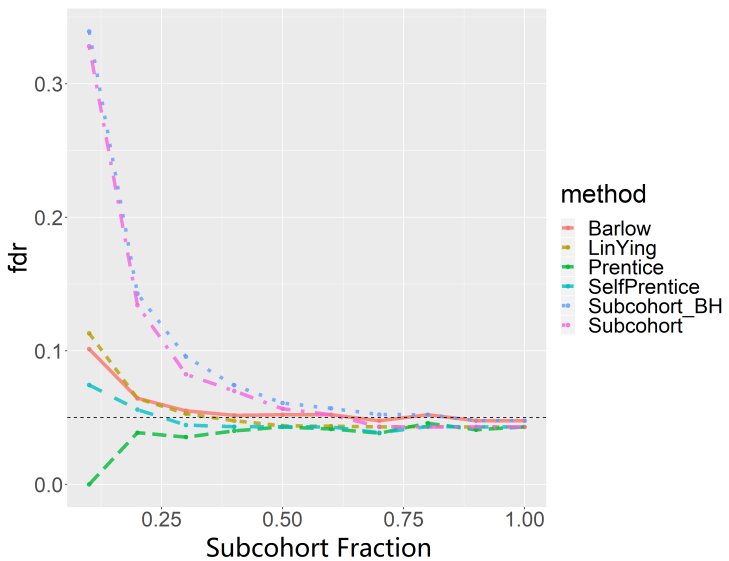


Figure 4.23: FDR of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.15. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset, and 10% are DEGs.

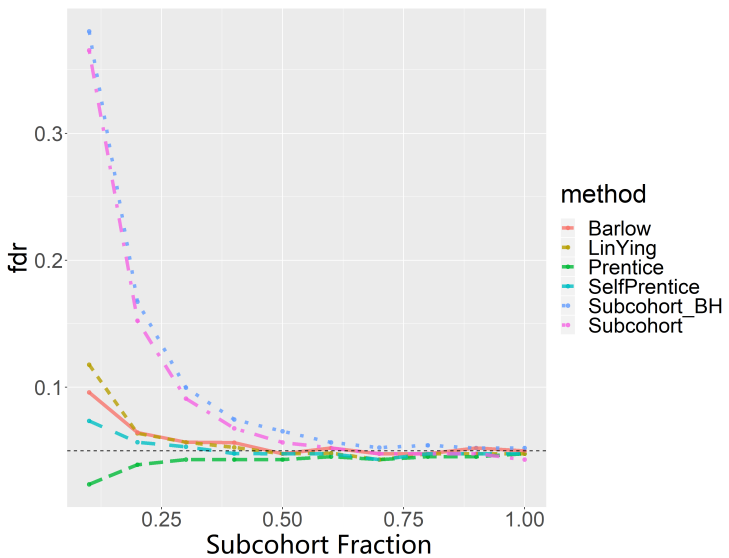


Figure 4.24: FDR of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.2. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset, and 10% are DEGs.

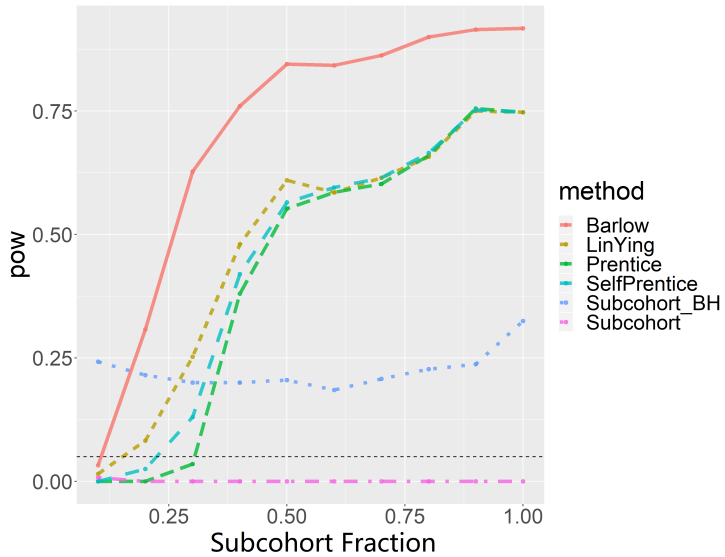


Figure 4.25: Power of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.05. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset and 10% are DEGs.

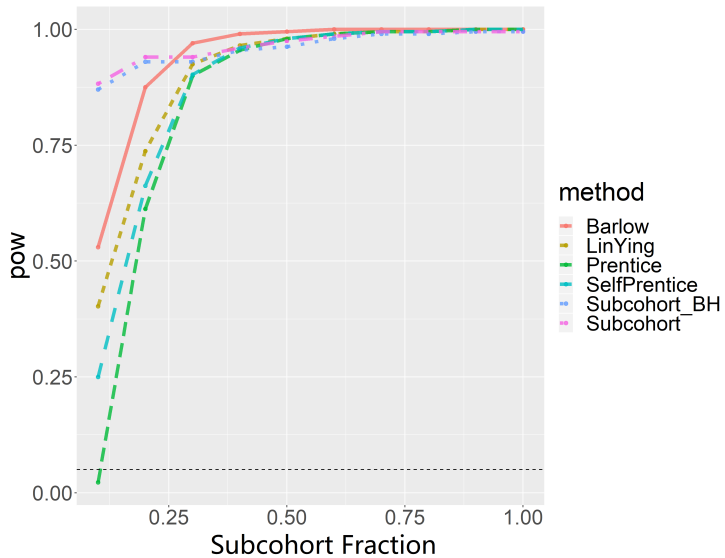


Figure 4.26: Power of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.1. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset and 10% are DEGs.

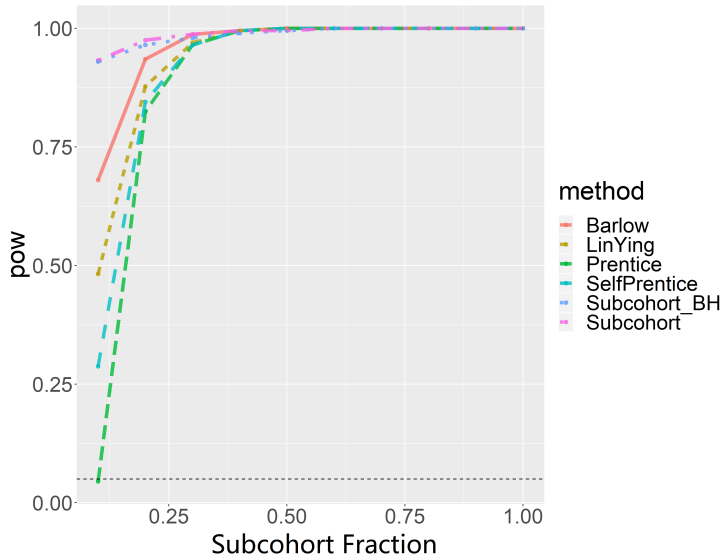


Figure 4.27: Power of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.15. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset and 10% are DEGs.

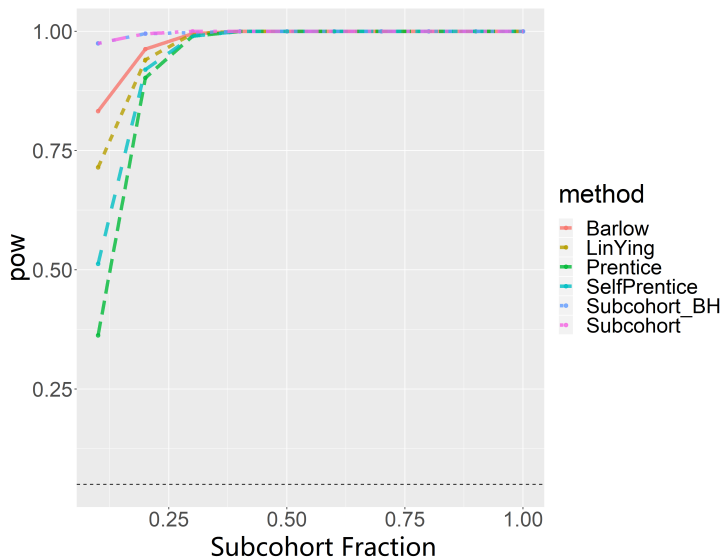


Figure 4.28: Power of subcohort, p-value BH-adjustment of subcohort, and other existing four methods on simulated multiple gene data with full cohort size 1000 and case rate 0.2. The hazard ratio is 1.8 at the start and 1.2 after three years. There are 2000 genes in each dataset and 10% are DEGs.

The situation is the same when we compare the "power" of the six methods. We found that in case rate 0.05, their powers are diverse (Shown in Figure 4.25. But when the case rate increases to 0.1, 0.15, or 0.2, the power curves of the six methods almost overlap in the middle and high subcohort fraction range (Shown in Figure 4.26 to 4.28).

4.3.3 Simulation 2 of non-proportional hazard data

Simulate survival times and genes' expression values

$$h(t|x, z(t)) = h_0(t)exp(\beta x)ke^{exp(\beta x)kt}. \tag{4.8}$$

Let $h_0(t) = \lambda$, we have

$$h(t|x, z(t)) = \lambda exp(\beta x)ke^{exp(\beta x)kt}. \tag{4.9}$$

Then, the cumulative hazard function is given by:

$$\begin{aligned} &H(t, x, z(t)) = \textstyle\int_0^t \lambda exp(\beta x)ke^{exp(\beta x)ku}\, du \\ &= \lambda(exp(exp(\beta x)kt) - 1). \end{aligned} \tag{4.10}$$

The survival function is related with the cumulative hazard function, $H(t|x)$, by

$$S(t|x) = exp[-H(t|x)]. \tag{4.11}$$

When $S(t|x)$ follows a continuous $\mathcal{U}(0, 1)$ distribution, a random variable $U = exp[-H(t|x)]$ also follows a continuous $\mathcal{U}(0, 1)$ distribution.

$$F = 1 - S = 1 - exp[-H(t|x)] = 1 - exp(-\lambda(exp(exp(\beta x)kt) - 1)). \tag{4.12}$$

$$\begin{aligned} &f(t, x|\lambda, k) = \tfrac{\partial F}{\partial t} \\ &= exp(-\lambda(exp(exp(\beta x)k \times t) - 1))\lambda(exp(exp(\beta x)kt) \times exp(\beta x)k. \end{aligned} \tag{4.13}$$

$$L(\lambda,k|x,t)=\prod_{i=1}^{n}f(t_i,x_i|\lambda,k)=exp(-\lambda\sum_{i=1}^{n}(exp(exp(\beta x_i)kt_i)-1))\\ \lambda^n exp(\sum_{i=1}^{n}exp(\beta x_i)t_i k)exp(\sum_{i=1}^{n}\beta x_i)k^n. \tag{4.14}$$

$$logL(\lambda,k|x,t)=-\lambda\sum_{i=1}^{n}(exp(exp(\beta x_i)kt_i)-1)\\ +nlog(\lambda)+\sum_{i=1}^{n}exp(\beta x_i)t_i k+\sum_{i=1}^{n}\beta x_i+nlog(k). \tag{4.15}$$

Let $\frac{\partial log(L(\lambda,k|x,t))}{\partial\lambda}=0$, we can get

$$\lambda_{mle}=n[\sum_{i=1}^{n}(e^{exp(\beta x_i)t_i k}-1)]^{-1}. \tag{4.16}$$

$$\frac{\partial log(L(\lambda,k|x,t))}{\partial k}=-\lambda\sum_{i=1}^{n}(exp(exp(\beta x_i)kt_i)exp(\beta x_i)t_i)\\ +\sum_{i=1}^{n}exp(\beta x_i)t_i+n/k, \tag{4.17}$$

which is a function of λ and k. We can use newton raphson method to solve λ and k for

$$\frac{\partial log(L(\lambda,k|x,t))}{\partial k}=0, \tag{4.18}$$

and

$$\frac{\partial log(L(\lambda,k|x,t))}{\partial\lambda}=0. \tag{4.19}$$

Therefore, an event time can be generated as

$$T=\left(\frac{1}{kexp(\beta x)}\right)log(1+\frac{(-log(U))}{\lambda_{mle}}). \tag{4.20}$$

When x increases, T decreases. x is associated with T. To make T in the range of "Time To Event" (TTE), multiply T with a scale factor and let

$$T=T\times max(TTE_i)/max(T_i). \tag{4.21}$$

Simulation of Full Cohorts and CCHs

Formulars above provides a non-proportional model. Firstly, we need to find a real DEG and use its information to estimate k_{mle} and λ_{mle}. P9906 cohort of ALL's

dataset was used here, as the "cage effect" does not need to be considered in a single cohort. Sex- and hemoglobin-related genes were removed according to the gene' "Probe.Set.ID". Then the package "SAMR" were used to find potential DEG candidates. Table 2.1 shows that when the median FDR is less than 0.05, the first related delta should be 0.7712257998. Using the value, we can compute the list of significant genes (the gene with large positive values or large negative values). The gene "KIAA0430", with Probe.Set.ID "202386_*s_at*", has a large negative score value, -2.867. It was randomly selected from negative expressed genes as the input to estimate k_{mle} and λ_{mle}, which is consistent with the designed negative association model above.

We assign $\beta = 5$ and input the gene's survival time and expression value to the model. Then we can use newton Rapson method to solve the derivative equations 4.18 and 4.19 to approximate k_{mle} and λ_{mle}.

To build a full cohort, we need to assign the number of patients in it. For example, we have n patients in a full cohort. Then we need to know whether the full cohort includes one gene or multiple genes. If it has only one gene and the gene is a DEG, we assume the expression value of patients follows a $N(\mu, \sigma)$ distribution, which can be calculated from the original expression value of gene 11823. The single gene expression values of n patients can be simulated from the $N(\mu, \sigma)$ distribution. Then input them to equation 4.20 and equation 4.21, we can acquire the event time of the n patients. For the null genes, we draw random numbers since we only care that they are unrelated to survival time. At last, we need to identify the control group and cases. For all patients, draw censoring times $C \sim Exp(\lambda_{cens})$ and compare them to their corresponding survival times. If censoring occurs after the survival time (censoring time is longer than the survival time) for an individual, they were observed to have experienced an event and are considered cases. Otherwise, they are in the control group.

For a multiple gene full cohort, we use exactly the same procedures above to acquire censoring, survival time, and gene expression value for a single gene. Then we need to know the percentage of DEGs in the full cohort. For DEG genes, the expression level of gene i for individual j is: $x_{ij} = x_j + e_{ij}$, where the e_{ij}s are $\mathtt{N}(0, 1)$ perturbations. We draw a set of perturbations to generate expression levels for each DEG. For the null genes, we draw random numbers since we only care that they are unrelated to survival time. Like proportional data simulation, a CCH sample consists of all cases (both in and out of the subcohort) but only the controls in the subcohort. The fraction of subcohort is 10% to 90%, respectively.

Type-I error and power for DEGs' identification in single gene datasets

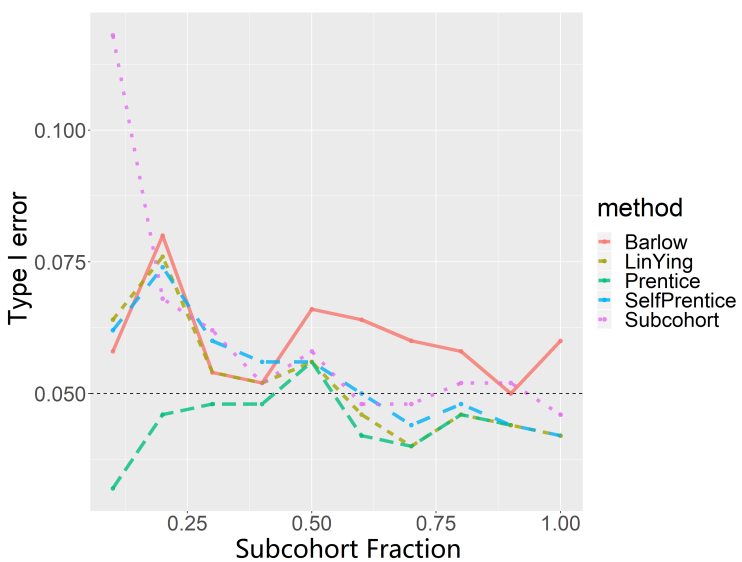


Figure 4.29: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.05, and $\beta = 5$.

For the second type of non-PH data, the five methods have similar performances on single gene datasets. When the subcohort is large enough and the percentage of

cases in patients is not low, or the number of cases in a dataset is not very rare, the "subcohort" method performs very similarly to other existing methods.

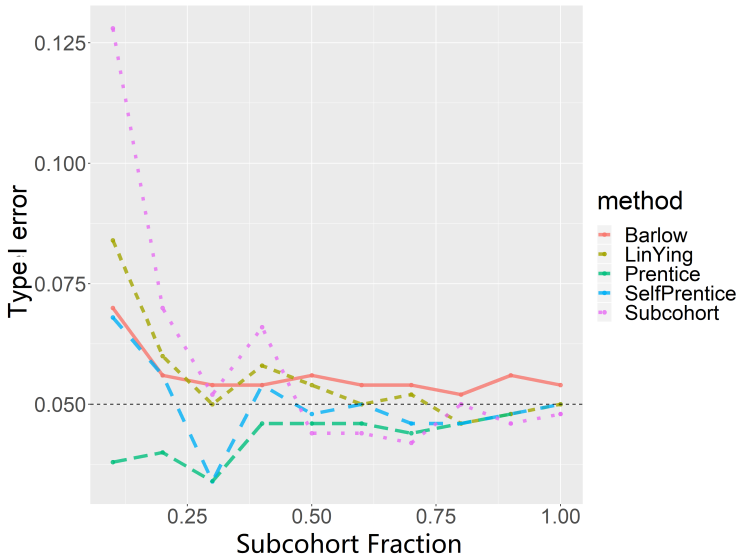


Figure 4.30: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.1, and $\beta = 5$.

Figure 4.31: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.15, and $\beta = 5$.

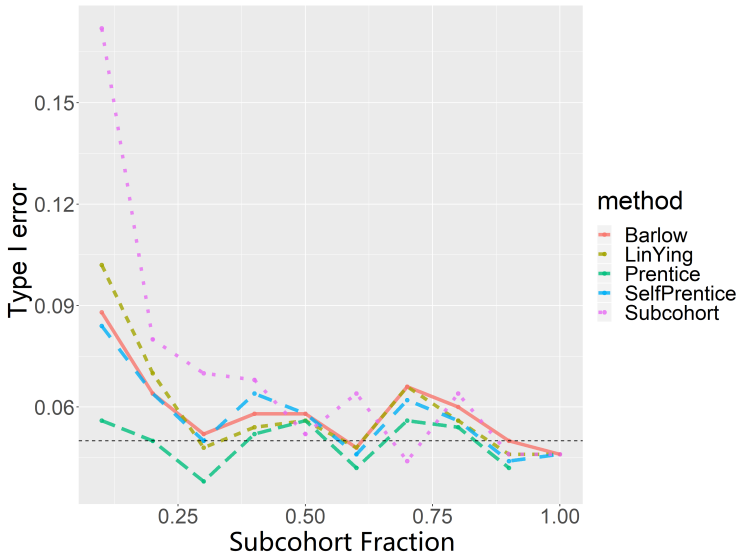


Figure 4.32: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.2, and $\beta = 5$.

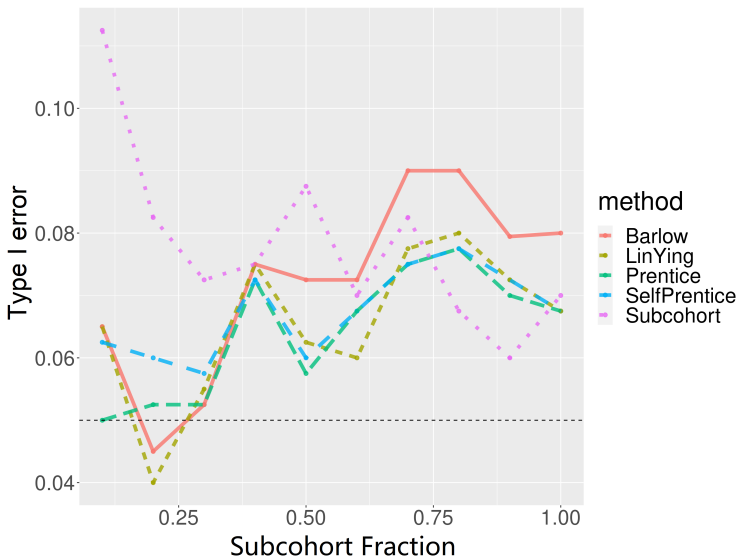


Figure 4.33: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.05, and $\beta = 10$.

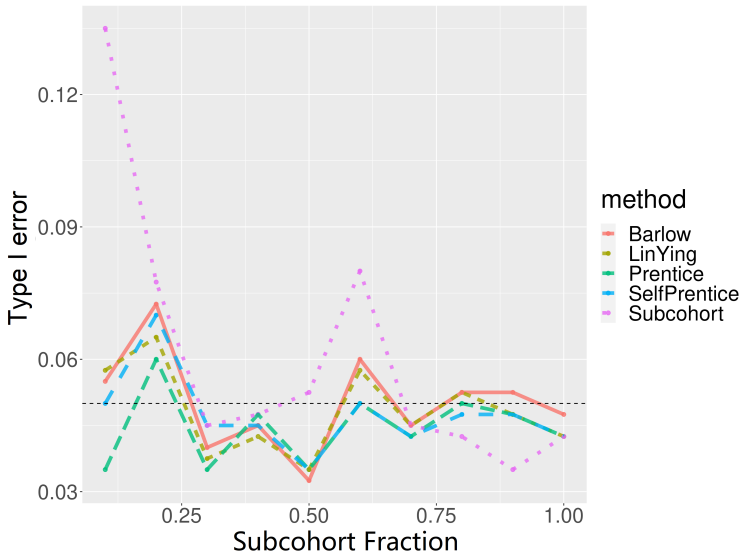


Figure 4.34: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.1, and $\beta = 10$.

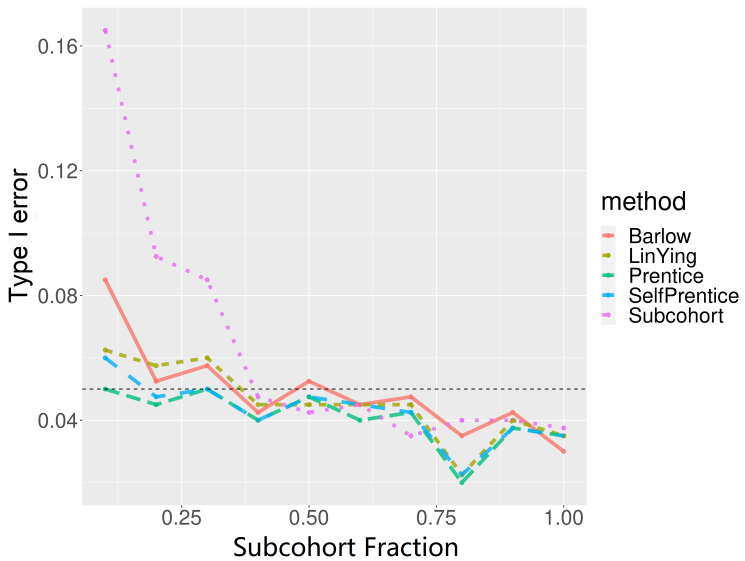


Figure 4.35: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.15, and $\beta = 10$.

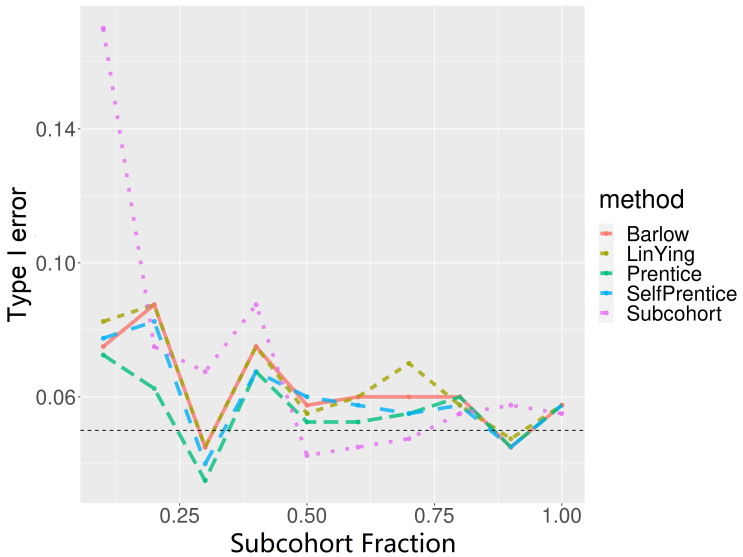


Figure 4.36: Type I error of subcohort and other existing four methods on simulated single gene data with full cohort size 1000, case rate 0.2, and $\beta = 10$.

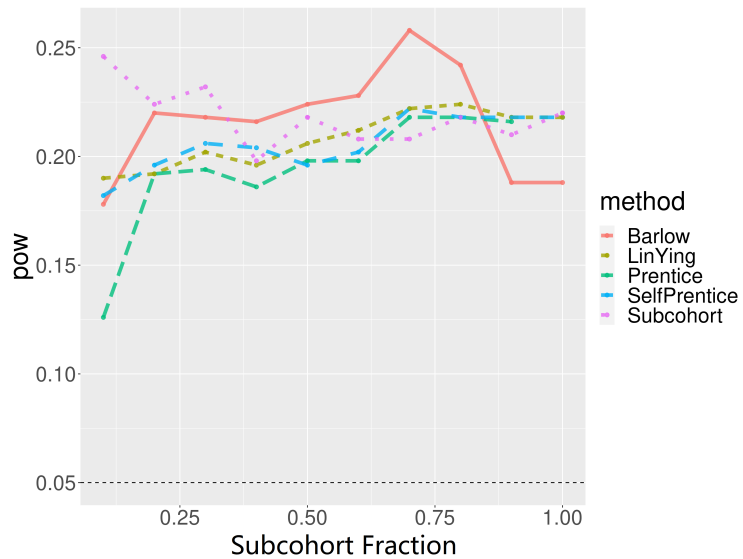


Figure 4.37: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.05, and $\beta = 5$.

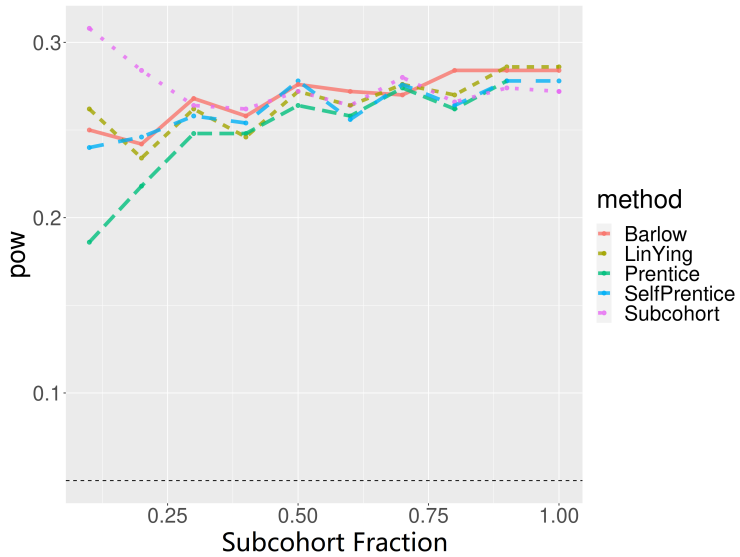


Figure 4.38: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.1, and $\beta = 5$.

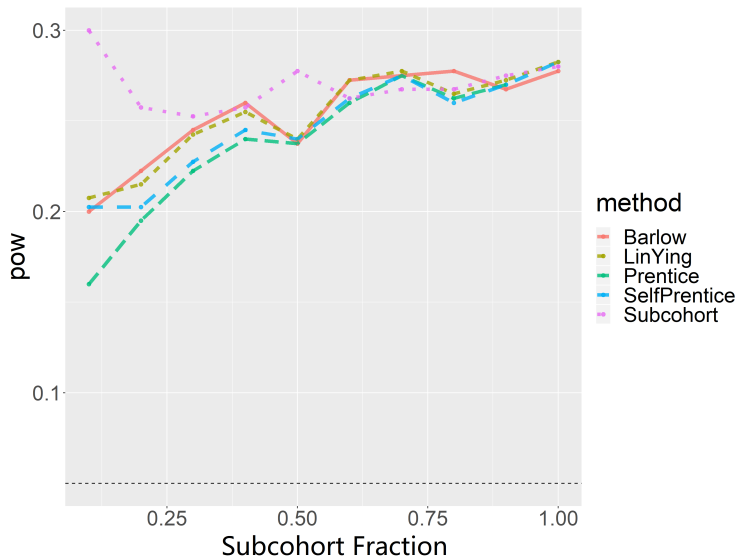


Figure 4.39: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.15, and $\beta = 5$.

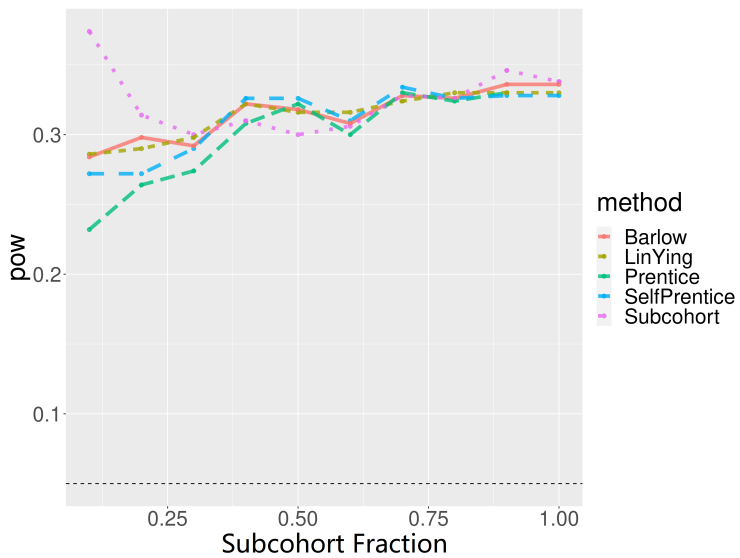


Figure 4.40: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.2, and $\beta = 5$.

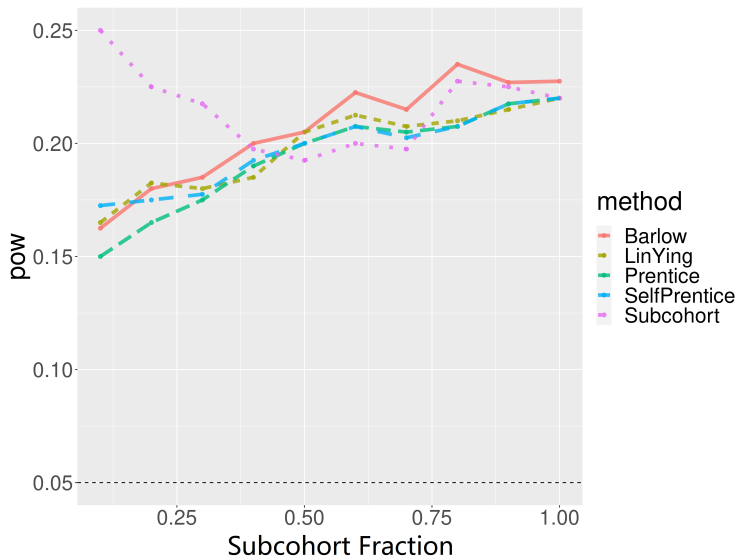


Figure 4.41: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.05, and $\beta = 10$.

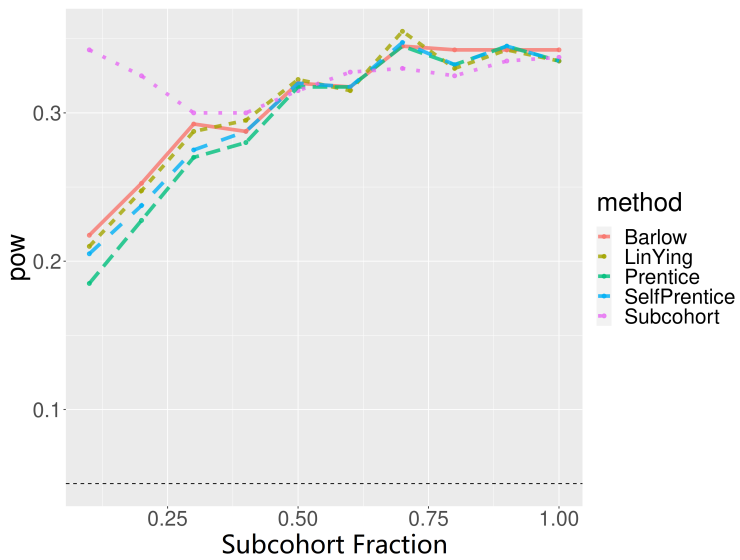


Figure 4.42: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.1, and $\beta = 10$.

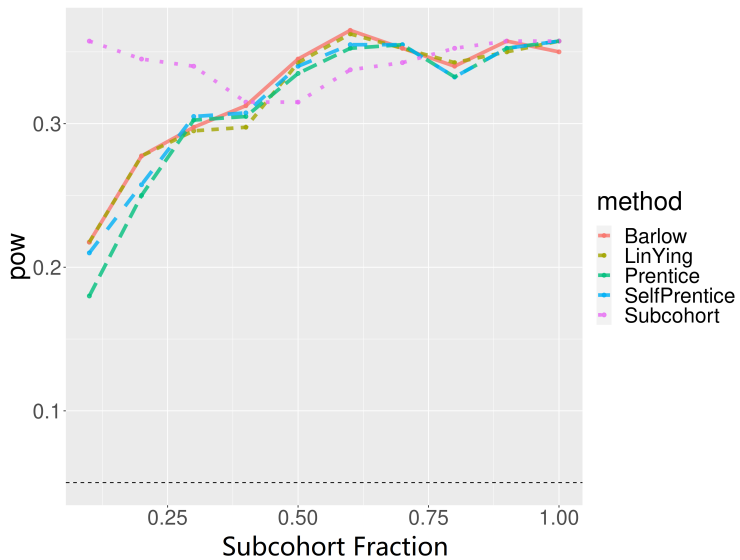


Figure 4.43: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.15, and $\beta = 10$.

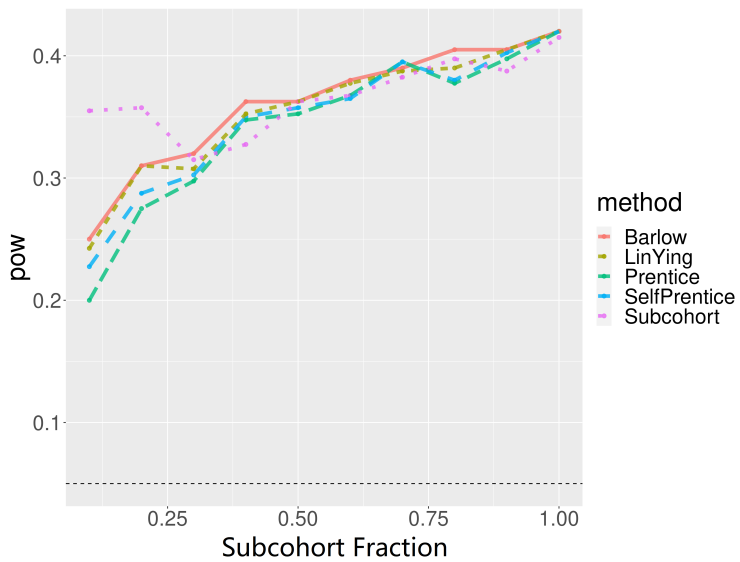


Figure 4.44: The power of subcohort and the other four existing methods are on simulated single gene data with full cohort size 1000, case rate 0.2, and $\beta = 10$.

As shown in Figure 4.29 to 4.32 for $\beta = 5$, the type I errors of five methods are comparable, although in case rate 0.05, the type I error of the "Barlow" method is higher than the other four methods. But when the case rate increases to 0.1, 0.15, or 0.2, the type I error curves of the five methods become very similar in the middle and high subcohort fraction range. When we increase β from 5 to 10 (shown in Figure 4.33 to 4.36), the situation is similar.

The situation is the same when we compare the "power" of the five methods. We found that in case rate 0.05 and $\beta = 5$, the power of the "Barlow" method is higher than the other four methods (Shown in Figure 4.37). But when the case rate increases to 0.1, 0.15, or 0.2, the power curves of the five methods almost overlap in the middle and high subcohort fraction range (Shown in Figure 4.38 to 4.40). To consider the effect of β, we increase β from 5 to 10 (shown in Figure 4.41 to 4.44). Results show that the situation is similar. The power of the "Barlow" method is higher than the other four methods in the case rate 0.05, but the power curves of the five methods almost overlap in the middle and high subcohort fraction range when

the case rate increases to 0.1, 0.15, or 0.2.

FDR and power for DEGs' identification in high-throughput gene datasets

We used p-value BH-adjustment and FDR to measure the performance of the CCH-based permutation test on high-throughput datasets. For the second type of non-PH data, the six methods have similar performances on high-throughput datasets.

When the subcohort is large enough and the percentage of cases in patients is not low, or the number of cases in a dataset is not very rare, the "Subcohort" method performs very similarly to other existing methods. As shown in Figure 4.45 to 4.48, their FDR is comparable, although in case rate 0.05 and 0.1, their FDR are low. But when the case rate increases to 0.15, or 0.2, the FDR curves of the six methods become very similar in the middle and high subcohort fraction range.

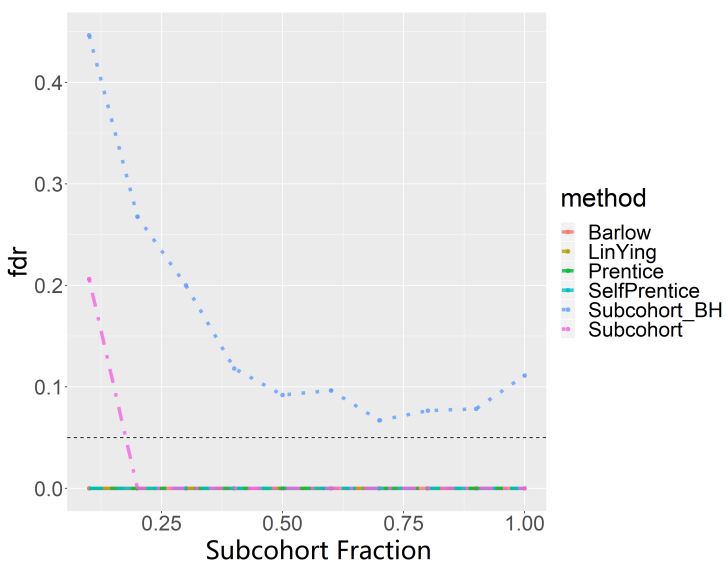


Figure 4.45: FDR of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.05. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

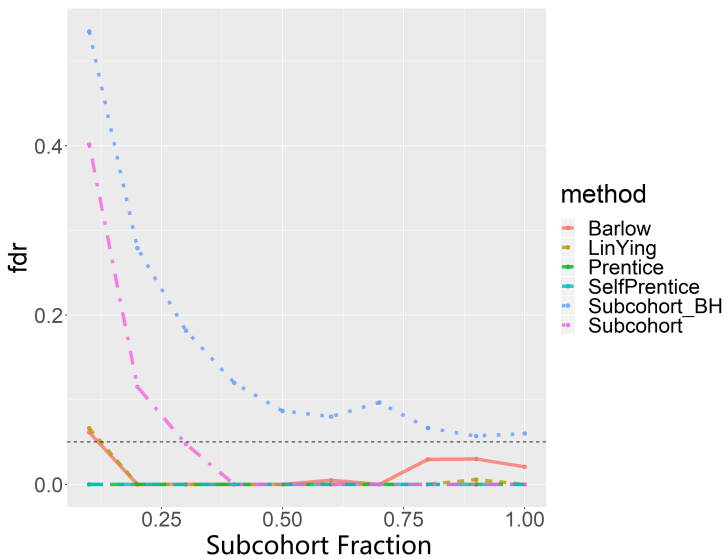


Figure 4.46: FDR of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.1. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

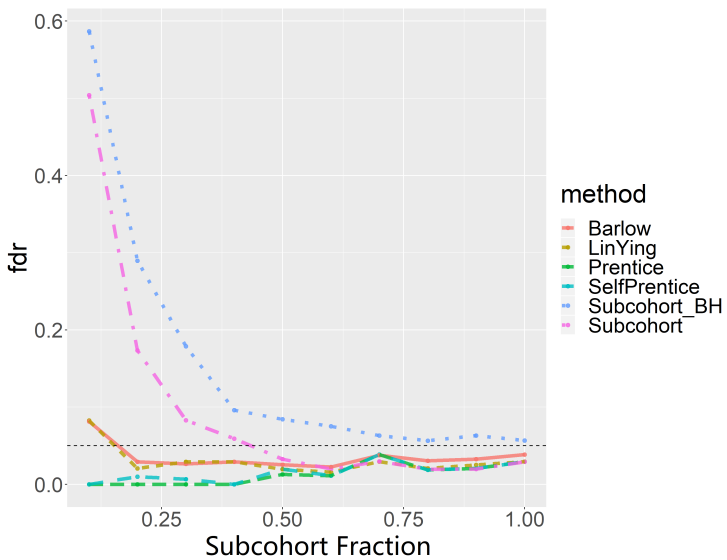


Figure 4.47: FDR of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.15. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

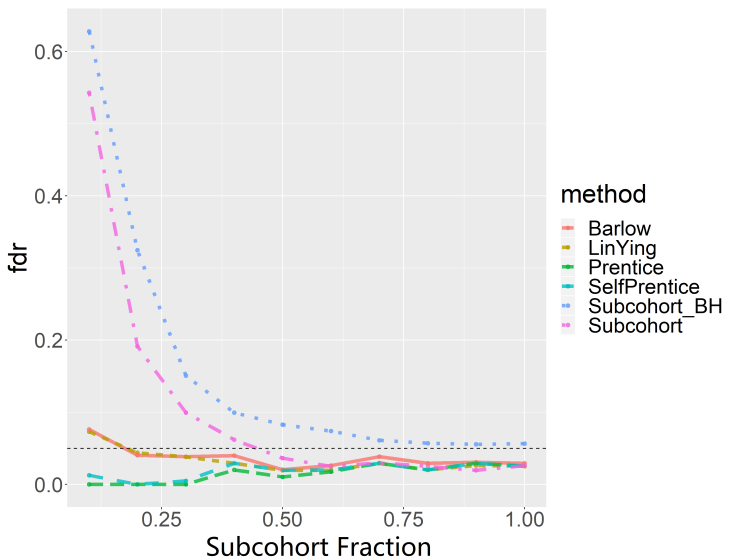


Figure 4.48: FDR of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.2. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

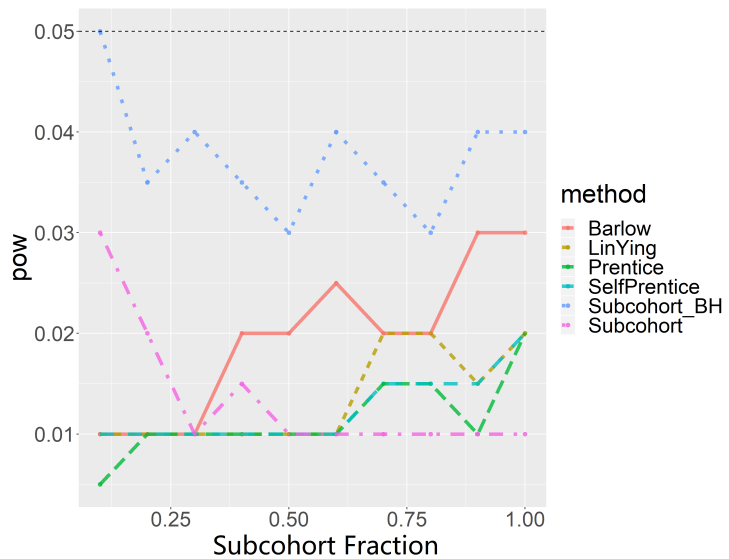


Figure 4.49: Power of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.05. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

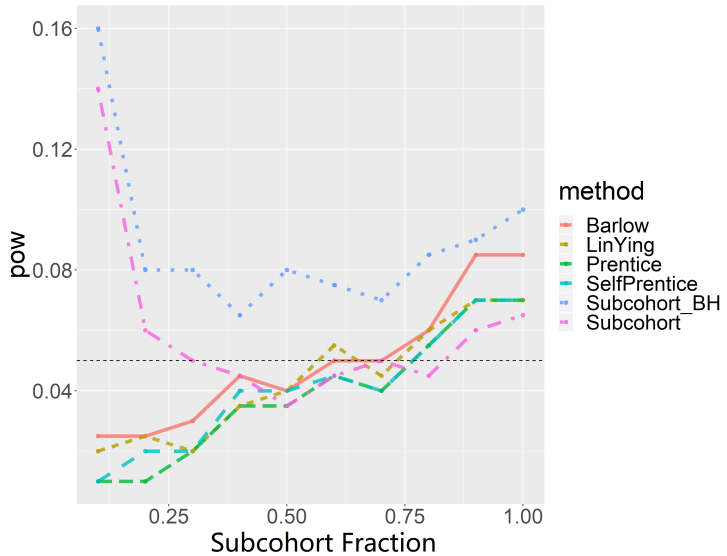


Figure 4.50: Power of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.1. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

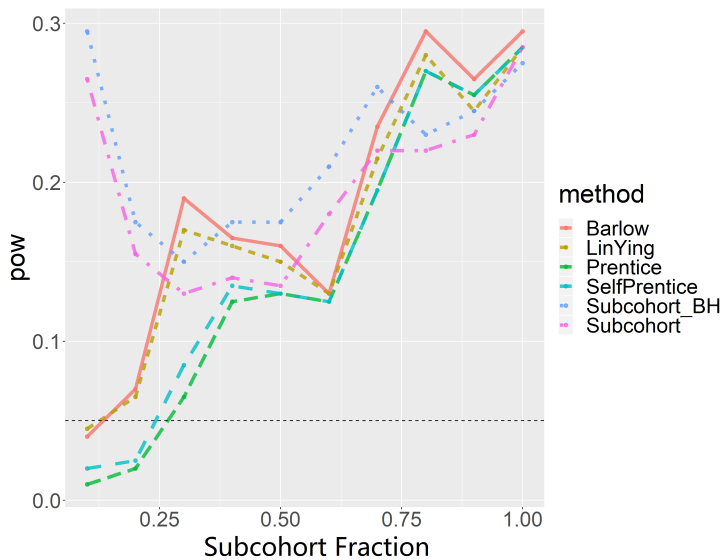


Figure 4.51: Power of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.15. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

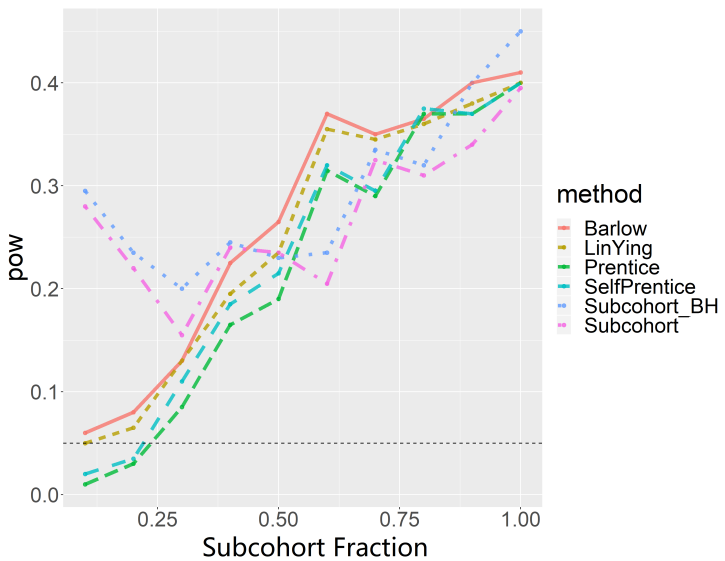


Figure 4.52: Power of subcohort, p-value BH-adjustment of subcohort, and other four existing methods on simulated multiple gene data with full cohort size 1000, and case rate 0.2. β is 10. There are 1000 genes in each dataset, and 10% are DEGs.

The situation is the same when we compare the "power" of the six methods. We found that in case rate 0.05 and 0.1, their power is low (Shown in Figure 4.49 to 4.50). But when the case rate increases to 0.15, or 0.2, the power curves of the six methods almost overlap in the middle and high subcohort fraction range (Shown in Figure 4.51 to 4.52).

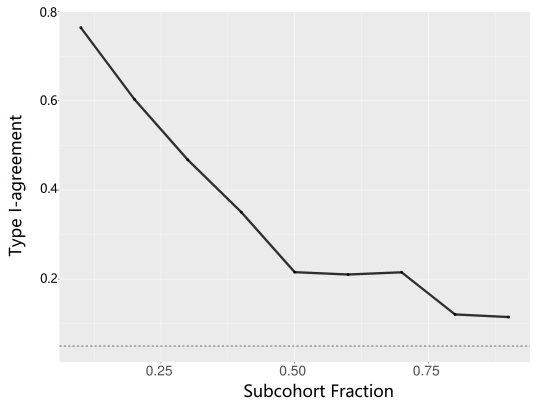
To sum up, the "subcohort" method has a very similar performance of FDR and power to other existing methods in high-throughput gene datasets of the second type non-PH model.

4.4 Application

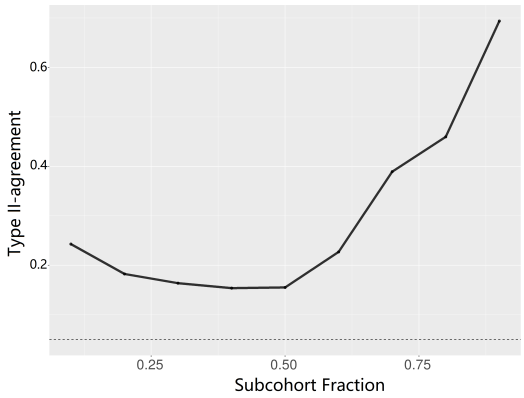
We applied the "subcohort" method to two real datasets. Pediatric B-Cell acute lymphoblastic leukemia (ALL) dataset has 801 patients from two cohorts, including

54,504 probe sets [25]. Its 20,000 probe sets with the largest interquartile range (IQR) were selected as the input dataset. For the BRCA data (the same data we used in Chapter 3), the 16000 genes (the data has only 16005 genes) with the largest interquartile range (IQR) were selected as a dataset. We choose 0.1, 0.2, ... and 0.9 as sub-cohort fraction. We reconstructed full cohort simulations 100 times for each real dataset for each subfraction and replicated a 1000 times permutation-based score test on each full reconstructed cohort.

As stated in the Method part, we defined "Type I-agreement" and "Type II-agreement" to measure the consistency between an original full cohort and a full reconstructed cohort. For example, for the top 16000 largest interquartile genes of the BRCA database, in a one-time simulation with a sampling fraction of 0.9, a full cohort analysis identified 373 DEGs, and a full reconstructed cohort identified 376 DEGs. They have 367 DEGs in common (shown in Figure 3.23). The related Type I-agreement = 9/(367+9)= 2.4% and Type II-agreement=367/(367+6) = 98.4%.

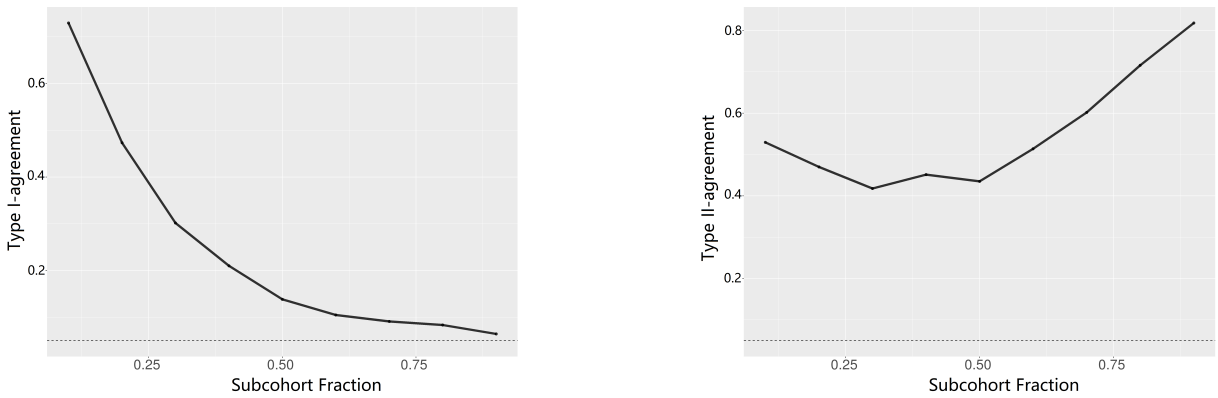


(a) Type I-agreement



(b) Type II-agreement

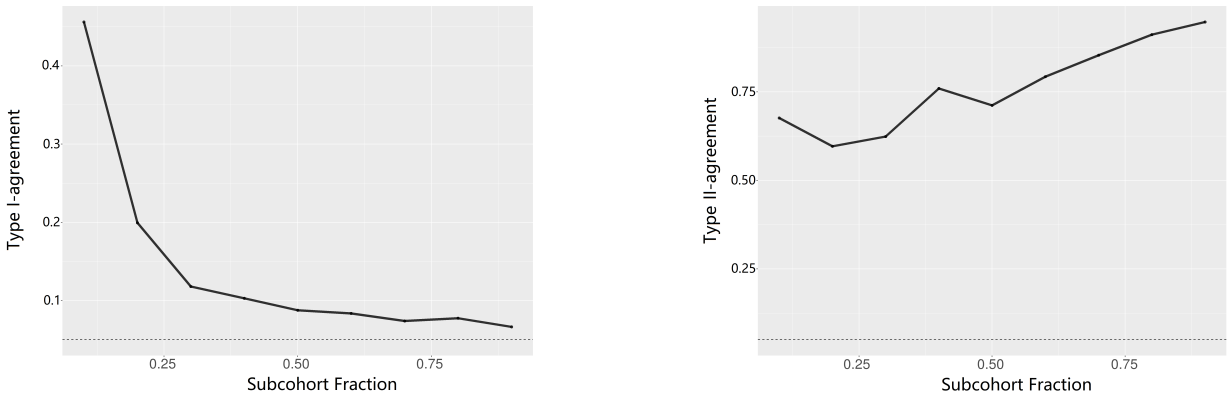
Figure 4.53: Performance of "subcohort" methods on BRCA.



(a) Type I-agreement (b) Type II-agreement

Figure 4.54: Performance of "subcohort" methods on ALL.

For a CCH method to be considered an adequate substitute for complete cohort analysis, it should have decreasing Type I-agreement and increasing Type II-agreement with increasing subcohort. However, Type I-agreement need not be below 5%, and Type II-agreement need not be above 95%, following the nominal 0.05 and 0.95 significance levels, respectively.



(a) Type I-agreement (b) Type II-agreement

Figure 4.55: Performance of subcohort on simulated multiple gene data with hazard ratio 1.5-1.6, full cohort size 1000, genes 2000, case rate 0.1 and DEG 10%.

The results of BRCA is shown in Figure 4.53. Type I-agreement decreases, and

Type II-agreement increases with the increase of subcohort fraction. The results of ALL and simulated data show a similar pattern (Figure 4.54 and 4.55). The only difference among the three figures is how quickly the Type I-agreement will decrease and how quickly Type II-agreement will increase within the scope of the subcohort fraction.

Chapter 5

Conclusion

High-throughput gene expression profiling technologies, such as microarray and RNA-Sequencing, have been widely used in medical research. One of the applications of these technologies is to identify, among thousands of genes, those whose expressions are associated with survival outcomes. The case-cohort (CCH) design is an efficient way to analyze survival data, particularly for large cohorts with low failure rates. However, the application of CCH design in a high-throughput gene expression analysis has not been seen in the published literature. In this dissertation, we sought to promote the use of the CCH design in gene expression analysis and to develop statistical methods for identifying the genes associated with survival outcomes under the CCH design.

A score test is usually preferred in a typical genomic study because it does not need to fit the Cox PH model iteratively when thousands of hypothesis tests must be performed simultaneously. Hence, it can save computing time and avoid potential convergence issues. Combining the advantage of the CCH study design and score test, we developed a score test under the CCH design to identify DEGs associated with survival outcomes. We provided asymptotic distribution theory and inferential

procedures for the test. a CCH-based score test was proposed, in which the covariance matrix $\Sigma(\beta_0)$ and $\Delta(\beta_0)$ in the asymptotic chi-square distribution of CCH were estimated at $\beta_0 = 0$ under the null hypothesis, rather than at $\tilde{\beta}$. Furthermore, the "Score Process" was acquired by deriving the first derivative of the log Pseudo-likelihood function. Then, a test statistic with asymptotic Chi-squared distribution was established to calculate the p-value under the null hypothesis, $\beta = 0$, which is equivalent to the null hypothesis that the tested gene is a NONDEG (survival time has no relationship with gene expression data).

For the CCH-based score test, we verified the validity of the inferential procedure in finite samples through simulation studies with datasets generated from PH models. For simulated single gene datasets, its type I error decreases quickly and around the nominal significance level with the increase of sub-cohort fraction, and power increases quickly and approach the power of the full cohort with the increase of sub-cohort fraction. For simulated high-throughput datasets of PH models, its FDR decreases quickly, and its power increase quickly with the increase of sub-cohort fraction. Besides we used an RNA-sequencing data set (BRCA) from a breast cancer study as the full cohort and draw a large number of CCH samples with different sampling fractions. We applied the CCH-based score test on each of the CCH samples and compared the result to that from applying an ordinary score test on the full cohort. The type-I agreement and Type-II agreement show that there was an excellent agreement between the set of genes identified both methods. Furthermost, we compared the proposed method with some existing methods that can also applied to the gene expression analysis under the CCH Design. The power of our CCH-based score method is comparable with other methods because all methods have similar performance in all combinations of event rate and full cohort size.

Permutation tests are non-parametric methods. When the permutation test and CCH design are combined on DEG identification, strong semi-parametric and prob-

ability distribution assumptions for p-value do not need to be concerned. In this project, a CCH-based permutation test procedure was proposed and applied for high-dimensional gene differential expression analysis, which can reduce the cost of DEG discovery and avoid statistical assumption violation. Another advantage of this method is that it estimates the false discovery rate (FDR) directly from the permutation procedure, which considers the correlation among the genomic features (genes).

We developed a procedure to reconstruct a full cohort from a CCH sample and then perform the permutation-based score test on the reconstructed full cohort to identify the DEGs associated with survival outcomes. To illustrate the usage and advantages of this method, we evaluated our testing procedures through simulation studies with datasets generated based on PH and non-PH models, respectively. For simulated single gene datasets of both PH and non-PH models, its type I error decreases quickly and around the nominal significance level and power increase quickly and approach the power of the full cohort with the increase of sub-cohort fraction. For simulated high-throughput datasets of both PH and non-PH models, its FDR decreases quickly, and its power increase quickly with the increase of sub-cohort fraction. Besides, we applied the proposed method to the same RNA-sequencing data set (BRCA) and a microarray data set (ALL) from a leukemia study. The results of the Type-I agreement and the Type-II agreement show good consistency between the proposed permutation test and the full cohort method. Furthermost, we compared the proposed method with some existing CCH methods. Results show they have a good consistency. When the number of patients' events is scarce, our proposed method performs better than the others.

The results indicates that our methods can be effectively used to identify DEGs in high-throughput gene expression dataset while reducing the costs for genomics experiments.

Chapter 6

Future Work

In the two CCH-based tests proposed in this dissertation, there is possible room to extend our investigation. The following areas are recommendations for further work.

For the CCH-based score test, we applied it to simulated datasets and a real question with one cohort using χ^2_1 distribution. Individuals may have different distributions for some categorical variables at different variable levels. We can extend our method by using a stratified design for those variables to consider their main effects. The degree of freedom of the extended CCH-based score test needs to be considered.

We proposed a CCH-based permutation test. We validated it through simulation studies and applied it to real problems. Further theoretical exploration may give a more precise mechanism of the method.

A nested case-control study design is a case-control study within a cohort study by the selection of several healthy controls for each case. A pseudo-likelihood approach was proposed [47]. Another future work is to develop similarly a "nested case-control study-based score test" and a "nested case-control study-based permutation test".

Appendices

Appendix A

Proving Estimators' Consistency for CCH Asymptotic Chi-Square Distribution

The defination of $\tilde{\triangle}(\beta)$, $\triangle$, $\tilde{G}(\beta_0, x, w)$ and $G(\beta_0, x, w)$ are in Prentice and Self [49] .

$$\triangle = \int_0^1 \int_0^1 G(\beta_0, x, w) S^{(0)}(x) S^{(0)}(w) \lambda_0(x) \lambda_0(w)\, dx\, dw. \tag{A.1}$$

$$\tilde{\triangle}(\beta) = \tfrac{1}{n^2} \int_0^1 \int_0^1 \tilde{G}(\beta, x, w)\, d\bar{N}(x)\, d\bar{N}(w). \tag{A.2}$$

Suppose β_0 is the true value of β. Under the null hypothesis of the score test, the real value of β is known. Pick $\beta = \beta_0$ in A.2, we get

$$\tilde{\triangle}(\beta_0) = \tfrac{1}{n^2} \int_0^1 \int_0^1 \tilde{G}(\beta_0, x, w)\, d\bar{N}(x)\, d\bar{N}(w). \tag{A.3}$$

We want to prove

$$\tilde{\triangle}(\beta_0) \to_P \triangle(\beta_0). \tag{A.4}$$

$$
\begin{aligned}
&G(\beta_0, x, w) = \tfrac{1-\alpha}{\alpha}[s^{(0)}(\beta_0, x)s^{(0)}(\beta_0, w)\}^{-1}h^{(1)}(\beta_0, x, w) \\
&+s^{(0)}(\beta_0, x)s^{(0)}(\beta_0, w)\}^{-2}s^{(1)}(\beta_0, x)s^{(1)}(\beta_0, w)^T h^{(0)}(\beta_0, x, w) \\
&-s^{(0)}(\beta_0, x)^{-1}s^{(0)}(\beta_0, w)^{-2}s^{(1)}(\beta_0, w)h^{(2)}(\beta_0, w, x) \\
&-s^{(0)}(\beta_0, w)^{-1}s^{(0)}(\beta_0, x)^{-2}s^{(1)}(\beta_0, x)h^{(2)}(\beta_0, x, w)],
\end{aligned} \tag{A.5}
$$

where $\alpha = (1 - eventRate) \times sr + eventRate$. The $eventRate$ is the event rate in the full cohort, and sr is the subfracton of the subcohort out of the full cohort.

$$
\begin{aligned}
&\tilde{G}(\beta_0, x, w) = \tfrac{1-\tilde{\alpha}}{\tilde{\alpha}}[\{\tilde{S}^{(0)}(\beta_0, x)\tilde{S}^{(0)}(\beta_0, w)\}^{-1}\tilde{H}^{(1)}(\beta_0, x, w) \\
&+\{\tilde{S}^{(0)}(\beta_0, x)\tilde{S}^{(0)}(\beta_0, w)\}^{-2}\tilde{S}^{(1)}(\beta_0, x)\tilde{S}^{(1)}(\beta_0, w)^T\tilde{H}^{(0)}(\beta_0, x, w) \\
&-\tilde{S}^{(0)}(\beta_0, x)^{-1}\tilde{S}^{(0)}(\beta_0, w)^{-2}\tilde{S}^{(1)}(\beta_0, w)\tilde{H}^{(2)}(\beta_0, w, x) \\
&-\tilde{S}^{(0)}(\beta_0, w)^{-1}\tilde{S}^{(0)}(\beta_0, x)^{-2}\tilde{S}^{(1)}(\beta_0, x)\tilde{H}^{(2)}(\beta_0, x, w)],
\end{aligned} \tag{A.6}
$$

where $\tilde{\alpha} = (1 - eventRate) \times sr + eventRate$. The $eventRate$ is the event rate in the full cohort, and sr is the subfracton of the subcohort out of the full cohort. First, we need to show

$$
\tilde{G}(\beta_0, x, w) \to_P G(\beta_0, x, w). \tag{A.7}
$$

Prentice and Self ([40] and [49]) listed condtions A-G to ensure the desired asymptotic distribution of $\tilde{\beta}$ and $\tilde{\Lambda}$. By condition G(iv): $\tilde{S}^{(0)}(\beta_0, x)$, $\tilde{S}^{(1)}(\beta_0, x)$, and $\tilde{Q}^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) converge to $s^{(0)}(\beta_0, x)$, $s^{(1)}(\beta_0, x)$, and $q^{(i)}(\beta_0, x, w)$ ($i =$ 0, 1, and 2), respectively. $\tilde{H}^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) are functions of $\tilde{S}^{(0)}(\beta_0, x)$, $\tilde{S}^{(1)}(\beta_0, x)$, and $\tilde{Q}^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) (Shown in Chapter 3), and $h^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) are functions of $s^{(0)}(\beta_0, x)$, $s^{(1)}(\beta_0, x)$, and $q^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) (Similar with the definition of $\tilde{H}^{(i)}(\beta_0, x, w)$. Only change $\tilde{S}^{(0)}(\beta_0, x)$, $\tilde{S}^{(1)}(\beta_0, x)$, and $\tilde{Q}^{(i)}(\beta_0, x, w)$ ($i = 0$, 1 and 2) to $s^{(0)}(\beta_0, x)$, $s^{(1)}(\beta_0, x)$, and $q^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2), respectively). So, $\tilde{H}^{(i)}(\beta_0, x, w)$ converge to $h^{(i)}(\beta_0, x, w)$ ($i = 0$, 1, and 2) uniformly on $\beta_0 \times [0, 1]^2$. So, $\tilde{G}(\beta_0, x, w)$ converges to $G(\beta_0, x, w)$.

And by the definition of $n^{-1}\bar{N}(t)$, it converges uniformly to $\int_0^1 S^{(0)}(x)\lambda_0(x)dx$. As $n^{-1}\bar{N}(t)$ is bounded in probability and $G(\beta_0, x, w)$ resides in the product space

of left-continuous functions, $\tilde{\triangle}(\beta_0)$ converges to $\triangle(\beta_0)$. Now, we want to prove

$$\tilde{\Sigma}(\beta_0) \to_P \Sigma(\beta_0). \tag{A.8}$$

The defination of $\tilde{\Sigma}(\beta)$, $\Sigma(\beta_0)$, $\tilde{V}(\beta_0, t)$ and $v(\beta_0, t)$ are in Prentice and Self [49].

$$\Sigma(\beta_0) = \int_0^1 v(\beta_0, t) s^{(0)}(\beta_0, t) \lambda_0(\beta_0, t)\, dt. \tag{A.9}$$

$$\tilde{\Sigma}(\beta_0) = \tilde{\Sigma}(\beta)|(\beta = \beta_0) = \tfrac{1}{n} \int_0^1 \tilde{V}(\beta_0, t)\, d\bar{N}(t). \tag{A.10}$$

$$v(\beta_0, t) = s^{(2)}(\beta_0, t)/s^{(0)}(\beta_0, t) - e(\beta_0, t)^{\otimes 2}. \tag{A.11}$$

$$e(\beta_0, t) = s^{(1)}(\beta_0, t)/s^{(0)}(\beta_0, t). \tag{A.12}$$

$$\tilde{V}(\beta_0, t) = \tilde{S}^{(2)}(\beta_0, t)/\tilde{S}^{(0)}(\beta_0, t) - \tilde{E}(\beta_0, t)^{\otimes 2}. \tag{A.13}$$

Similarly, as $\tilde{S}^{(0)}(\beta_0, x)$, $\tilde{S}^{(1)}(\beta_0, x)$, and $\tilde{S}^{(2)}(\beta_0, x)$ converge to $s^{(0)}(\beta_0, x)$, $s^{(1)}(\beta_0, x)$ and $s^{(2)}(\beta_0, x)$, respectively. We get

$$\tilde{V}(\beta_0, t) \to_P v(\beta_0, t). \tag{A.14}$$

And by the definition of $n^{-1}\bar{N}(t)$, it converges uniformly to $\int_0^1 S^{(0)}(x)\lambda_0(x)dx$. As $n^{-1}\bar{N}(t)$ is bounded in probability and $v(\beta_0, t)$ resides in the product space of left-continuous functions, we have:

$$\tilde{\Sigma}(\beta_0) \to_P \Sigma(\beta_0). \tag{A.15}$$

Appendix B

The test statistics of CCH-based permutation test

For ith gene, x_i is the gene expressions of the original full cohort, and x_i^R is the gene expressions of the reconstructed full cohort. e_i is the difference between x_i^R and x_i ($e_i = x_i^R - x_i$). We assume x_i and x_i^R have the same mean and variance. e_{ij} is 0 for patient j inside of CCH. We assume that e_{ij} has the mean 0 and finite variance σ_e^2 for patient j outside of CCH.

The definition of the score S_i of the reconstructed full cohort for gene i is:

$$S_i = \frac{rc_i}{sc_i + sc_0}. \tag{B.1}$$

where rc_i is the numerator of the score, sc_i is a standard deviation and sc_0 is an exchangeability factor. We use $sc_0 = 0$ for simplification. For censored survival data, rc_i is defined as

$$rc_i = \sum_{k=1}^{K}(x_{ik}^* - d_k \bar{x}_{ik}). \tag{B.2}$$

And sc_i is defined as

$$sc_i = [\sum_{k=1}^{K}((\frac{d_k}{m_k}) \sum_{j \in R_k}(x_{ij}^R - \bar{x}_{ik}))^2]^{1/2}. \tag{B.3}$$

Where x_{ij}^R is the expression value of gene i for patient j. x_{ij}^R is either from CCH or be imputed. As the same as the original full cohort analysis, D be the indices of the K unique death times z_1, z_2, ... z_K, and R_1, R_2, ... R_K be the indices of the observations at risk at these unique death times, that is $R_k = \{i : t_i \geq z_k\}$. Let $m_k = \#inR_k$. Let d_k be the number of deaths at time z_k. $x_{ik}^* = \sum_{t_j=z_k} x_{ij}^R$ and $\bar{x}_{ik} = \sum_{j \in R_k} \frac{x_{ij}^R}{m_k}$.

$$\begin{aligned} rc_i &= \textstyle\sum_{k=1}^{K}(x_{ik}^* - d_k \bar{x}_{ik}) \\ &= \textstyle\sum_{k=1}^{K}(\sum_{t_j=z_k} x_{ij} - d_k \sum_{j \in R_k} \frac{x_{ij}+e_{ij}}{m_k}) \\ &= \textstyle\sum_{k=1}^{K}(\sum_{t_j=z_k} x_{ij} - d_k \sum_{j \in R_k} \frac{x_{ij}}{m_k} - d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k}) \\ &= r_i + \textstyle\sum_{k=1}^{K}(-d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k}), \end{aligned} \tag{B.4}$$

where r_i is completely from the original full cohort.

As $E(\sum_{k=1}^{K}(-d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k})) = 0$,

$$E(rc_i) = r_i. \tag{B.5}$$

$$\begin{aligned} Var(rc_i) &= Var(\textstyle\sum_{k=1}^{K}(-d_k \sum_{j \in R_k} \frac{e_{ij}}{m_k})) \\ &= \textstyle\sum_{k=1}^{K}(d_k^2 Var(\sum_{j \in R_k} \frac{e_{ij}}{m_k})) \\ &= \textstyle\sum_{k=1}^{K}(\frac{d_k^2}{m_k^2} Var(\sum_{j \in R_k} e_{ij})) \\ &= \textstyle\sum_{k=1}^{K}(\frac{d_k^2}{m_k^2} \sum_{j \in R_k} Var(e_{ij})) \\ &= \textstyle\sum_{k=1}^{K}(\frac{d_k^2}{m_k^2} \sigma_e^2 \#_k), \end{aligned} \tag{B.6}$$

where $\#_k$ is the number of imputed gene expressions in the risk set R_k, and $\#_k$ is less than or equal to m_k.

$$Var(rc_i) = \textstyle\sum_{k=1}^{K}(\frac{d_k^2}{m_k^2} \sigma_e^2 \#_k) \leq \sigma_e^2 \sum_{k=1}^{K}(\frac{d_k^2}{m_k}) < \infty. \tag{B.7}$$

Equation B.7 shows that the "score process" of the CCH-based permutation test

has a limited variance.

$$
\begin{aligned}
&{sc_i}^2 = \sum_{k=1}^{K}((\tfrac{d_k}{m_k})\sum_{j\in R_k}(x_{ij}^R - \bar{x}_{ik}))^2 \\
&= \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(x_{ij}^R - \bar{x}_{ik}))^2 \\
&= \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(x_{ij} + e_{ij} - \sum_{j\in R_k}\tfrac{x_{ij}+e_{ij}}{m_k}))^2 \\
&= \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k} + e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2 \\
&= \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k}) + \sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2 \\
&= \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k}))^2 \\
&+2 * (\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k})) \times (\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k})) + (\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2 \\
&= {s_i}^2 + \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2 2 * (\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k})) \times (\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k})) + \\
&(\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2,
\end{aligned}
\tag{B.8}
$$

where ${s_i}^2$ is only dependent on the original full cohort.

$$
\begin{aligned}
&E({sc_i}^2) = E({s_i}^2 + \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2 2 * (\sum_{j\in R_k}(x_{ij} + \sum_{j\in R_k}\tfrac{x_{ij}}{m_k})) \times (\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k})) + \\
&\sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2) \\
&= {s_i}^2 + 0 + E(\sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\sum_{j\in R_k}(e_{ij} - \sum_{j\in R_k}\tfrac{e_{ij}}{m_k}))^2) \\
&= {s_i}^2 + \sigma_e^2 \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\#_k - \tfrac{2\#_k}{m_k} + \tfrac{{\#_k}^2}{{m_k}^2}).
\end{aligned}
\tag{B.9}
$$

From equation B.6 and B.9, we know $E({sc_i}^2)$ is a function of $Var(rc_i)$.

$$E({sc_i}^2) = Q * Var(rc_i) + {s_i}^2 \tag{B.10}$$

where $Q = \sum_{k=1}^{K}(\tfrac{d_k}{m_k})^2(\#_k - \tfrac{2\#_k}{m_k} + \tfrac{{\#_k}^2}{{m_k}^2}) / \sum_{k=1}^{K}(\tfrac{d_k^2}{m_k^2}\#_k)$.

As $(\#_k - \tfrac{2\#_k}{m_k} + \tfrac{{\#_k}^2}{{m_k}^2}) = (1 - \tfrac{2\#_k}{m_k} + \tfrac{{\#_k}^2}{{m_k}^2}) + \#_k - 1 \geq (1 - \tfrac{\#_k}{m_k})^2 \geq 0$, we have $E({sc_i}^2) \geq {s_i}^2$.

References

[1] *Rare List*. Global Genes, 15 April 2016. Retrieved 15 April 2016.

[2] *Rare Disease Act of 2002*. United States Congress, 2002. Retrieved 21 January 2022.

[3] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):106, 2010.

[4] M. J. Anderson. Permutation tests for univariate or multivariate analysis of variance and regression. *Can. J. Fish. Aquat. Sci.*, 58:626–639, 2001.

[5] A. Anjum, S. Jaggi, E. Varghese, S. Lall, A. Bhowmik, and A. Rai. Identification of differentially expressed genes in rna-seq data of arabidopsis thaliana: A compound distribution approach. *J Comput Biol*, 23(4):239–247, 2019.

[6] W.E. Barlow, L. Ichikawa, D. Rosner, and S. Izumi. Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, 52(12):1165–1172, 1999.

[7] R. Bender, T. Augustin, and M. Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

[8] W. C. Blackwelder. Current issues in clinical equivalence trials. *Journal of Dental Research*, 83:113–115, 2004.

[9] N. E. Breslow. Analysis of survival data under the proportional hazards model. *Int Stat Rev*, 43:45–57, 1975.

[10] G. Chu, M. Seo, J. Li, B. Narasimhan, R. Tibshirani, and V. Tusher. *SAM "Significance Analysis of Microarrays" Users guide and technical document*. 2001.

[11] D. S. Collingridge. A primer on quantitized data analysis and permutation testing. *Journal of Mixed Methods Research*, 7(1):79–95, 2013.

[12] D. Commenges. Transformations which preserve exchangeability and application to permutation tests. *Journal of Nonparametric Statistics*, 15(2):171–185, 2003.

[13] S. J. Costa, D. Domingues, and F. M. Lopes. Rna-seq differential expression analysis: An extended review and a software tool. *PLoS ONE*, December 2017.

[14] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–220, 1972.

[15] D. R. Cox. Partial likelihood. *Bioetrika*, 62(2), 1975.

[16] D. Lin D and Z. Ying. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, 88(424):1341–1349, 1993.

[17] C. W. Dunnet. Pairwise multiple comparison in the homogenous variance, unequal sample size case. *Journal of the American Statistical Association*, 75:789–795, 1980.

[18] R. R. Esteban and X. Y. Jiang. Differential gene expression in disease: a comparison between high-throughput studies and the literature. *BMC Medical Genomics*, 10(59), 2017.

[19] J. Fan and J. Jiang. Non-and semi-parametric modeling in survival analysis. *New Developments in Biostatistics and Bioinformatics*, 1:3–33, 2009.

[20] R. A. Fisher. *The Design of Experiments*. New York-Hafner, New York, 1935.

[21] S. C. Gad. *Encyclopedia of Toxicology (Third Edition)*. Elsevier Inc, 2014.

[22] J. Han, M. J. Chen, Y. H. Wang, B. X. Gong, T. W. Zhuang, L. Y. Liang, and H. Qiao. Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Scientific Reports*, 8(9912), 2018.

[23] D. T. Jones, A. Banito, T. G. Grünewald, M. Haber, N. Jäger, and M. Kool M. Molecular characteristics and therapeutic vulnerabilities across paediatric solid tumours. *Nature Reviews Cancer*, 19(8):420–438, 2019.

[24] L. Jun and T. Robert. Finding consistent patterns: A nonparametric approach for identifying differential expression in rna-seq data. *Stat Methods Med Res*, 22(5):519–536, 2013.

[25] H. Kang, I. M. Chen, C. S. Wilson, E. J. Bedrick, R. C. Harvey, S. R. Atlas, M. Devidas, C. G. Mullighan, X. Wang, M. Murphy, K. Ar, W. Wharton, M. J. Borowitz, W. P. Bowman, D. Bhojwani, W. L. Carroll, B. M. Camitta, G. H. Reaman, M. A Smith, J. R Downing, S. P. Hunger, and C. L. Willman.

Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric b-precursor acute lymphoblastic leukemia. *Blood*, 115(7):1394–1405, 2010.

[26] R. S. Kim. A new comparison of nested case-control and case-cohort designs and methods. *Eur J Epidemiol*, 30(3):197–207, 2015.

[27] D. G. Kleinbaum and M. Klein. *Evaluating the Proportional Hazards Assumption*. Springer, New York, 2011.

[28] S. Kulathinal, J. Karvanen, O. Saarela, K. Kuulasmaa, and the MORGAM Project. Case-cohort design in practice – experiences from the morgam project. *Epidemiol Perspect Innov*, 4(15), 2007.

[29] T. K. Lam, S. D. Schully, S. D. Rogers, R. Benkeser, B. Reid, and M. J. Khoury. Provocative questions in cancer epidemiology in a time of scientific innovation and budgetary constraints. *Cancer Epidemiol Biomarkers Prev*, 22:496–500, 2013.

[30] B. Langholz and D. Thomas. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. *American Journal of Epidemiology*, 131:169–176, 1990.

[31] B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 1(12), 2011.

[32] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550–558, 2014.

[33] Y. F. Lu, D. B. Goldstein, M. Angrist, and G. Cavalleri. Personalized medicine and human genetic diversity. *Cold Spring Harbor Perspectives in Medicine*, 4(9), 2014.

[34] S. Ogino, P. Lochhead, E. Giovannucci, J. A. Meyerhardt, C. S. Fuchs, and A. T. Chan. Discovery of colorectal cancer pik3ca mutation as potential predictive biomarker: power and promise of molecular pathological epidemiology. *Oncogene*, 33(23):420–438, 2014.

[35] N. C. Onland-Moret, Y. T. van der Schouw, W. Buschers, S. G. Elias, C. H. Van Gils, J. Koerselman, M. Roest, D. E. Grobbee, and P. H. Peeters. Analysis of case-cohort data: a comparison of different methods. *Journal of clinical epidemiology*, 60(4):350–355, 2007.

[36] C. E. Parker, D. Domanski, A. J. Percy, A. G. Chambers, A. G. Camenzind, D. S. Smith, and C. H. Borchers. *Chemical Diagnostics*. Springer, New York, 2012.

[37] J. C. Pesko. *Contributions to Statistical Testing, Prediction, and Modeling*. Doctoral dissertation, University of New Mexico, Albuquerque, NM, USA, 2017.

[38] E. J. G. Pitman. Significance tests which may be applied to samples from any population. *Royal Statistical Society*, Supplement 4(1):119–130 and 225–232, 1937.

[39] R. L. Prentice. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73:1–11, 1986.

[40] R. L. Prentice and S. G. Self. Asymptotic distribution theory for cox-type regression models with general relative risk form. *Ann. Statist*, 11:804–813, 1983.

[41] P. Ranganathan and C. S. Pramesh. Censoring in survival analysis: Potential for bias. *Perspect Clin Res*, 3(1):40, 2012.

[42] C. R. Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 44(1):50–57, 1948.

[43] N. Reid. A conversation with sir david cox. *Statistical Science*, 9(3):439–455, 1994.

[44] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[45] D. Rossell and F. J. Rubio. Additive bayesian variable selection under censoring and misspecification. *arXiv*, 2019.

[46] S. A. Rusticus and C. Y. Lovato. Impact of sample size and variability on the power and type i error rates of equivalence tests: A simulation study. *Practical Assessment, Research, and Evaluation*, 19(11), 2021.

[47] S. Samuelsen. A pseudo-likelihood approach to analysis of nested case-control studies. *Biometrika*, 84:379–394, 1997.

[48] M. K. Samur. Rtcgatoolbox: a new tool for exporting tcga firehose data. *PLos One*, 9(9), 2014.

[49] S. G. Self and R. L. Prentice. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, 16(1):64–81, March 1988.

[50] M. S. Setia. Methodology series module 1: Cohort studies. *Indian J Dermatol*, 61(1):21–25, 2016.

[51] M. S. Setia. Methodology series module 2: Case-control studies. *Indian J Dermatol*, 61(2):146–151, 2016.

[52] S. J. Sharp, M. Poulaliou, S. G. Thompson, I. R. White, and A. M. Wood. A review of published analyses of case-cohort studies and recommendations for future reporting. *PLoS One*, 9(6), 2014.

[53] G. K. Smyth. *Limma: linear models for microarray data*. Springer, New York, 2005.

[54] J. W. Song and K. C. Chung. Observational studies: Cohort and case-control studies. *Plast Reconstr Surg*, 126(6):2234–2242, 2016.

[55] M. J. Stensrud and M. A. Hernan. Why test for proportional hazards? *JAMA Guide to Statistics and Methods*, 323(14), 2021.

[56] J. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(1):9440–9445, 2003.

[57] Z. Tevak, M. Kondratovich, and E. Mansfield. Us fda and personalized medicine: In vitro diagnostic regulatory perspective. *Personalized Medicine*, 7(5):517–530, 2010.

[58] V. G. Tusher, R. Tibshirani, and G. Chu. Survival analysis with high-dimensional covariates. *Proceedings of the National Academy of Sciences of the United States of America*, 98:5116–5121, 2001.

[59] D. M. Witten and R. Tibshirani. Survival analysis with high-dimensional covariates. *Stat Methods Med Res*, 19(1):29–51, 2010.

[60] H. I. Woo and S. W. Lim. Differentially expressed genes related to major depressive disorder and antidepressant response: genome-wide gene expression analysis. *Experimental & Molecular Medicine*, 92(50):1–10, 1983.

[61] S. Xue, J. Qiao, F. Pu, M. Cameron, and J. J. Yang. Design of a novel class of protein-based magnetic resonance imaging contrast agents for the molecular imaging of cancer biomarkers. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*, 5(2):163–179, 2013.

[62] T. O. Yau. Precision treatment in colorectal cancer: Now and the future. *JGH Open*, 3(5):361–369, 2019.

[63] A. Zaman. Urn models for markov exchangeability. *Annals of Probability*, 12:223–229, 1984.

[64] G. Zhang, Z. J. Xue, C. K. Yan, J. L. Wang, and H. M. Luo. A novel biomarker identification approach for gastric cancer using gene expression and dna methylation dataset. *Frontiers in Genetics*, 12(644378), 2021.

[65] X. M. Zhao and G. Qin. Identifying biomarkers with differential analysis. *Bioinformatics for Diagnosis, Prognosis and Treatment of Complex Diseases.*, 2013.