LONGITUDINAL MULTIVARIATE NORMATIVE COMPARISONS AND ACCURACY IMPROVEMENT METRICS FOR COMPETING ENDPOINTS

by

Zheng Wang

B.S., University of Science and Technology of China, 2015

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Zheng Wang

It was defended on

March 26th 2020

and approved by

Yu Cheng, Ph.D., Department of Statistics

Kehui Chen, Ph.D., Department of Statistics

Satish Iyengar, Ph.D., Department of Statistics

Chung-Chou H. Chang, Ph.D., Department of Medicine and Biostatistics

Dissertation Director: Yu Cheng, Ph.D., Department of Statistics

LONGITUDINAL MULTIVARIATE NORMATIVE COMPARISONS AND ACCURACY IMPROVEMENT METRICS FOR COMPETING ENDPOINTS

Zheng Wang, PhD

University of Pittsburgh, 2020

Motivated by the Multicenter AIDS Cohort Study (MACS), we are interested in developing methods to address several challenges in analyzing the data. The objectives of the research include developing dementia classification procedures from longitudinal data to control family-wise error, modeling covariates effects on time to dementia, and evaluating the improvement of diagnostic accuracy when including a new biomarker in the model.

In order to properly define an event of interest, we adapt the cross-sectional multivariate normative comparison (MNC) method, which controls family-wise error by accounting for the intercorrelations among all covariates, to a longitudinal setting. Longitudinal MNC is constructed based on multivariate mixed effects models when hypothesis testing happens by the conclusion of study.

In practice, patients may require visit-by-visit classification. Prompt feedback can guide treatments for longer survival. We propose to modify longitudinal MNC statistics to build visit-by-visit test statistics. Then, based on predicted number of visits from survival model and Poisson regression, we can apply Bonferroni-type correction to control family-wise error for this prospective classification scenario.

Lastly, we examine a new biomarker's contribution to diagnostic accuracy for competing-risk outcomes. The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were originally proposed to characterize accuracy improvement in predicting a binary outcome, when new biomarkers are added to regression models. These two indices have been extended from dichotomous outcomes to multi-categorical and survival outcomes. We extend the NRI and IDI to competing-risk outcomes, by adopting the definitions of the two indices for multi-

category outcomes. The "missing" category due to independent censoring is handled through the inverse probability weighting.

Keywords: Competing risks; Cumulative incidence function; Family-wise error rate; Integrated discrimination improvement; Multivariate mixed-effect model; Net reclassification improvement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

This dissertation has been written to fulfill the graduation requirements for my doctoral degree. This research provides unsupervised classification methods from two very practical perspectives. They have strong applications in broad fields dealing with multivariate longitudinal data and can help guide clinicians in treating patients. The last work focuses on feature selection for competing-risk survival models in terms of diagnostic accuracy improvements. This is also valuable for doctors and researchers in building more accurate diagnostic model. All the methodologies have been developed closely around the collaboration work I had with Dr. James T. Becker on Multicenter AIDS Cohort Study from August of 2016.

Research was difficult, but the challenges I encountered brought me more knowledge and confidence. Looking back five years, I couldn't have imagined where I am right now, showing you my work over the years and meeting people who helped me enrich my intellectual and emotional capacities.

I would like to express my deepest appreciation to my adviser, Dr. Yu Cheng. Dr. Cheng has been a wonderful mentor and helped me a lot through research, career and life challenges. Her invaluable insight into survival analysis and statistical applications guided me through the dissertation. Dr. Cheng is also a great life mentor with many constructive advice and emotional support. I also gratefully acknowledge the help from my other committee members, Dr. Kehui Chen, Dr. Satish Iyengar and Dr. Chung-Chou Chang. I would like to extend my sincere thanks to Dr. James T. Becker and other collaborators in MACS neuropsychology working group for their insights and guidance of AIDS research.

Another person I would like to extend my deepest gratitude to is my spouse, Ryan Leslie Hoy. I have known Ryan throughout my graduate study and we got married during March 2019. He is a strong supporter to my graduate study and career plan. He is also always there with me when I

am going though challenges in life. This dissertation would not have gone smoothly without Ryan along the way. I also would like to thank my friends throughout the years, Bang Wang, Shihao Zhang, Huanqin Liu and Hanwen Liu.

Family also played an important part for me to pursue my doctoral degree. Although we had many disagreements, I would like to acknowledge their support for my education. I wish to thank my grandparents, Shuping Zhang and Chengxiang Gao, my mother, Gaoling Zhang, my father, Shengchao Wang, and my sisters, Siyuan Wang and Sirun Wang.

This dissertation incorporates the efforts from not just me but also everyone else in my life. I sincerely hope that you enjoy reading and learning about these methodologies and statistical analysis of clinical research.

1.0 GENERAL INTRODUCTION

In medical research, subjects are often followed up with physical and (or) mental evaluations over time. The researchers and clinicians may be interested in identifying the onset of some diseases, and evaluating how treatment affects the risks of onset along with other demographic and clinical variables. Furthermore, when a new biomarker is introduced into study, it is often of interest to know whether it will improve the diagnosis of the disease so that patients with higher risks will be treated more aggressively. The first step of establishing effective analysis is giving proper definition of event of interest. Then an appropriate modeling procedure can be used to relate these covariates with the events of interest, which can be further utilized to investigate the accuracy improvement over the course of variable additions.

For example, in the Multicenter AIDS Cohort Study (MACS), subjects will be periodically measured by their brain cognitive domain functioning as well as other characteristics. These brain cognitive domain functioning scores will provide clinicians with useful information about how a subject's cognitive functioning evolves longitudinally and when they have developed mild cognitive impairment and dementia. After the event of interest such as dementia is properly identified, one may want to examine how covariates and treatment affect the time to dementia. Naturally, death is a competing risk preventing dementia from being observed, and a competing risks survival model is helpful in evaluating the association between covariates and risks. On top of this, when a new biomarker is introduced to the study, researchers are interested in knowing whether this biomarker will improve diagnostic accuracy of dementia. Motivated by the MACS data, in this thesis we are developing methods to specifically address these practical problems.

First, to properly make a dementia classification, it is crucial to take family-wise error into account. Huizenga et al. (2007) proposed the cross-sectional multivariate normative comparisons (MNC) to control family-wise error by accounting for the inter-correlation among multiple domain

scores and making hypothesis testing on all domains simultaneously. However, in a longitudinal setting where covariates are recorded multiple times, applying the cross-sectional MNC at each visit will still inflate family-wise error rate due to multiple testing over all visits. We thus extend the cross-sectional MNC method to a longitudinal setting by taking into account multiple testing over different domains and repeated measures.

Approaches to addressing family-wise error inflation vary, depending on different research purposes in practice. If classification is desired after a study concludes or when researchers have collected sufficient number of visits for classification decisions, number of visits is fixed for each patient. We propose to use multivariate linear mixed effects model to characterize longitudinal multivariate data (Fieuws and Verbeke, 2004; Fang et al., 2006). With estimated parameters from the mixed model, a longitudinal MNC procedure is proposed by adapting cross-sectional MNC method to repeated multivariate measures to control the overall family-wise error rate. We will also show that a permutation test is appropriate when the data fail to satisfy the multivariate normal assumption.

On the other hand, clinicians might want to give prompt feedback to patients at each visit based on all the historical data up to that visit. Using MACS as an example, subjects are carefully monitored, and prompt interventions are provided if subjects' cognitive scores are far from normal. The problem is also closely related to Just-in-Time Adaptive Interventions (JITAIs), in which subjects are monitored over time and interventions are determined at each measurement based on solely the historical data (Klasnja et al., 2015; Nahum-Shani et al., 2018). However, few literature related to family-wise error control was found for this type of problems. Most existing procedures, such as Bonferroni (Bonferroni, 1936; Dunn, 1959, 1961) and Benjamini-Hochberg-Yekutiele procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001) do not work, because number of visits for each patient is unknown at the beginning of the study or during the study.

The only relevant works that we found come from the online testing literature. The α-investing rule, proposed by Foster and Stine (2008) and developed by Aharoni and Rosset (2013), exhibited the ability to control false discovery rate in an online multiple hypothesis testing scenario when number of tests is unknown and potentially infinite. However, the α-investing rule only takes into account historical testing results, and does not differentiate among subjects with different testing frequency. During an ongoing clinical study, historical data provide useful information about how

many visits a subject may have left. Therefore, we propose to use some frequency model to predict number of visits each subject will have. At the same time, we will introduce a dynamic procedure based on Bonferroni correction. As for visit-by-visit test statistic, we propose to modify previously mentioned longitudinal MNC method so that visit-by-visit test statistics are independent for the same subject and power can be preserved.

After the events of interest are properly defined, one may want to evaluate how covariates impact the risks associated. In the example of the MACS, when clinicians are concerned about the event of cognitive impairment, patients may die before such an event can be observed, and the event time may be competing-risk censored by death. Popular competing risks modeling strategies include applying Cox proportional hazard regression (Cox, 1972) to each cause, and using Fine-Gray's subdistribution hazard model (Fine and Gray, 1999). The former approach draws upon a familiar Cox model, where the parameters can be interpreted as the log hazard ratio. The latter is popular as it directly evaluates how covariates affect CIFs. However, the sum of estimated CIFs from the Fine and Gray model across all events may exceed one, causing difficulty in interpreting covariate effects. Gerds et al. (2012) proposed a logistic risk regression model in a competing risks setting. Due to its multinomial nature, the sum of event probabilities and survival probability at any given time point is strictly equal to one. The covariate effects can be explained as the changes in log odds ratio between event probability and survival probability, though the estimation procedure is not readily available in existing packages. Cubic B-spline functions defined by De Boor (2001) is flexible to be applied to the logistic risk regression model to approximate baseline multinomial logistic functions, and we can implement the maximum likelihood estimator of the logistic risk regression model in R and SAS.

Last but not least, after events are properly defined and appropriate survival model is established, introducing a new biomarker into the statistical model may change the risks associated with events of interests. Therefore, clinicians want to know whether including this new biomarker might help separate those who will develop the event of interest by a certain time point from those who will not. The proper discrimination can lead to different treatment strategies. The net reclassification improvement (NRI) and the integrated discrimination improvement (IDI) were originally proposed to characterize accuracy improvement in predicting a binary outcome, when new biomarkers are added to regression models (Pencina et al., 2011; Uno et al., 2013). These two indices have

been extended from dichotomous outcomes to multi-categorical and survival outcomes (Li et al., 2013b). Working on the MACS study where the onset of cognitive impairment is competing-risk censored by death, we extend the NRI and the IDI to competing risk outcomes, by using CIFs to quantify cumulative risks of competing events, and adopting the definitions of the two indices for multi-category outcomes. The "missing" category due to independent censoring is handled through the inverse probability weighting. Various competing risks models are considered, such as the Fine and Gray (Fine and Gray, 1999), multistate (Cox, 1972; Cheng, 2009), and multinomial logistic models (Gerds et al., 2012). Procedures of estimation for the NRI and the IDI with their asymptotic distribution from competing risks data are presented, and a bias-corrected and accelerated bootstrap (Efron, 1987) method are applied for inference of the IDI in light of the difficulty in establishing a general variance estimator across competing-risks survival models. However, Demler et al. (2017) pointed out that variance estimators based on U-statistic theory will fail if the model is under the null hypothesis. We adopt their remedial measures to handle the degeneracy in the NRI and the IDI under the null, and will validate the theoretical results and the robustness of their remedial method through simulation studies.

The rest of the proposal is organized as follows. Chapter 2 and 3 will introduce two approaches to control family-wise error in identifying disease onset based on longitudinal data. In Chapter 2, multivariate linear mixed effects model with time effect and dependency structure will be used for testing subject's cognitive status when cognitive impairment classification happens by the conclusion of study. Then in Chapter 3, we propose a family-wise error controlling procedure when cognitive impairment classification is desired at every visit, such that patients will get prompt results back and treatments can be prescribed on time. Last but not least, in Chapter 4 we extend the NRI and IDI to competing-risk survival outcomes, and elaborate inferential procedures for the extended metrics. Simulation studies are carried out in each chapter to demonstrate the effectiveness of our methods. Then, we will apply proposed methods to the MACS to illustrate how they should be applied for real data, followed by some discussions and future work.

2.0 LONGITUDINAL MULTIVARIATE NORMATIVE COMPARISON

2.1 INTRODUCTION

Classification plays an important role in many fields of medical science. For example, identifying participants with cognitive impairment will enable clinicians to provide patients with proper treatments. Several methods of counting the number of domains with abnormal cognitive functioning scores Antinori et al. (2007); Gisslén et al. (2011) have been used in the fields of HIV and Alzheimer's Disease, despite the evidence that these methods are associated with inflated family-wise error rate (FWER). In order to control FWER at a pre-determined level and correct for inter-correlations among multiple cognitive domains, Huizenga et al. Huizenga et al. (2007) developed the so-called Multivariate Normative Comparison (MNC) method which specifically takes the covariance of the domain scores into consideration. Let $\boldsymbol{X}_i$ denote a vector of q cognitive domains measured for participant i. The healthy control group contains n participants, and their sample mean and sample covariance matrix of the q domain scores are denoted as $\hat{\boldsymbol{\mu}}_{\boldsymbol{c}}$ and $\hat{\boldsymbol{\Psi}}_{\boldsymbol{c}}$. If each vector of q domain scores is independent and identically distributed over a multivariate normal distribution for every participant, one could build an F-statistic for testing cognitive impairment for a certain individual:

$$\frac{n(n-q)}{(n+1)(n-1)q}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_c)^T \hat{\boldsymbol{\Psi}}_{\boldsymbol{c}}^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_c) \sim F(q, n-q).$$

In principle, the MNC method can effectively control FWER in impairment classification as long as the data follow a multivariate normal distribution Su et al. (2015); Wang et al. (2019). In practice, participants may visit the same clinician or institution multiple times. For example, if participants come to an Alzheimer Disease Research Center (ADRC) with memory complaints or

concerns on cognition, they will be followed roughly annually and their cognitive functioning will be assessed repeatedly over time. If the MNC is employed at each visit and participants are tested at a pre-specified α level, the resulting FWER of being categorized as cognitively impaired would be greatly inflated for failing to account for multiple testing over repeated visits. Moreover, an F distribution may not be a good approximation of the test statistic when the assumption of multivariate normality is not satisfied in real data. Here we propose two longitudinal MNC procedures that specifically take into account multiple tests over repeated measures.

One natural way to approach the longitudinal data is to utilize a multivariate linear mixed effects (MLME) model. Reinsel Reinsel (1982a) established theories for multivariate longitudinal models with repeated measures when data are balanced and parameters are unrestricted. Heitjan and Sharma Heitjan and Sharma (1997) further considered an autoregressive error structure for longitudinal data and estimated the parameters with the maximum likelihood approach. Fang et al. Fang et al. (2006) introduced a modified expectation-maximization (EM) algorithm to make it feasible to estimate unknown parameters in a MLME model with constrained intercepts. Fieuws and Verbeke Fieuws and Verbeke (2004) studied how the associations between different responses evolve over time and jointly modeled two responses by allowing a dependence structure among the random terms in the model. They further proposed to model longitudinal outcomes in a pairwise fashion for computation efficiency when too many outcome variables are considered Fieuws and Verbeke (2006). Verbeke et al. Verbeke et al. (2014) gave a rather comprehensive review of development in multivariate longitudinal analysis, and pointed out that joint modeling is preferred over univariate modeling to answer research questions about associations among various outcomes over time. Hout et al. proposed a longitudinal MLME model with change-point predictors for non-linear trends van den Hout et al. (2015).

Here, we start by assuming a multivariate normal distribution for longitudinal domain scores, and use the MLME to obtain the mean function and covariance structure of domain scores from healthy controls. Under multivariate normality, a testing procedure based on χ^2 is then proposed to classify cognitive status for each participant. However, multivariate normality is often difficult to assume for collected data. If the dependency structure is not sufficiently specified or the data fail to follow a multivariate normal distribution, the χ^2 procedure may still have an inflated family-wise error. Therefore, we propose a permutation test for our proposed test statistic which is robust

against distribution assumptions.

The rest of the chapter is organized as following. Modeling and testing procedures are detailed in Section 2.2. Then, in Section 2.3 some simulation studies are carried out 1) when multivariate normal distribution is satisfied, and 2) when the assumption is not satisfied. We will illustrate in Section 2.4 how to use MLME and the χ^2 and permutation tests for neuropsychological data collected from the Multicenter AIDS Cohort Study (MACS). Finally, we will conclude with some discussions on interpretation of the MNC and future work.

2.2 LONGITUDINAL MULTIVARIATE NORMATIVE COMPARISONS

2.2.1 Testing Procedure Based on χ^2

Assume there are n participants enrolled in a healthy group which is used as the reference, and each participant has q cognitive domains tested over m_i total visits during the study. Domain test scores are usually normalized so that a multivariate normal distribution holds for each visit. Let $Y_{ijk}, i = 1,...,n; j = 1,...,q; k = 1,...,m_i$ denote the tested score of participant i for domain j over k-th visit. Considering that scores of a single domain assessed across m_i visits are correlated with each other, and scores of two different domains from the same participant are correlated, we model Y_{ijk} using a MLME model:

$$Y_{ijk} = \beta_{j0} + \beta_{j1} t_{ik} + \beta_{j2} t_{ik}^2 + \beta_{j3} t_{ik}^3 + \nu_{ij} + \delta_{ik} + \epsilon_{ijk}. \tag{2.1}$$

Here we use q polynomial functions of degree 3 to describe the changes in the mean domain scores over time, and can add higher order terms if necessary. Alternatively, the B-spline technique can be used to approximate the true mean domain scores over time Bloxom (1985); De Boor (2001); Shumaker (2007); Rutherford et al. (2015); Harrell (2015). ϵ_{ijk} is assumed to be independent and identically distributed (i.i.d.) normal $N(0,\sigma^2)$, which is specific to each observation or measurement. Similarly, δ_{ik}, which represents the visit-specific effect, is also assumed to be i.i.d. normal $N(0,\theta^2)$. Because different domain functions tend to be correlated with each other for the same participant, $\boldsymbol{\nu}_i = (\nu_{i1},..,\nu_{iq})^\top$ is assumed to be $N(\mathbf{0},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{sr}], s,r = 1,...,q$. Generally, the

symmetric matrix $\boldsymbol{\Sigma}$ could be left unspecified, or assumed to have the structure of auto-regression or compound symmetry.

All unknown parameters can be estimated from an MLME modelFang et al. (2006); Fieuws and Verbeke (2004), which are denoted as $\hat{\beta}_{j0}, \hat{\beta}_{j1}, \hat{\beta}_{j2}, \hat{\beta}_{j3}, j = 1, \ldots, q, \hat{\rho}_{sr}, s, r = 1, \ldots, q$, $\hat{\theta}^2$, and $\hat{\sigma}^2$. For participant d to be tested, we take all q domain scores observed over m_d visits, and stack them into a single vector

$$\boldsymbol{U}_d = (Y_{d11}, ..., Y_{dq1}, Y_{d12}, ..., Y_{dq2}, ..., Y_{d1m_d}, ..., Y_{dqm_d})^\top. \tag{2.2}$$

From the linear mixed effects model in (2.1), the estimated mean vector of $\boldsymbol{U}_d$ is written as $\hat{\boldsymbol{\mu}}_d = (\hat{\beta}_{10} + \hat{\beta}_{11} t_{d1} + \hat{\beta}_{12} t_{d1}^2 + \hat{\beta}_{13} t_{d1}^3, \hat{\beta}_{20} + \hat{\beta}_{21} t_{d1} + \hat{\beta}_{22} t_{d1}^2 + \hat{\beta}_{23} t_{d1}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1} t_{d1} + \hat{\beta}_{q2} t_{d1}^2 + \hat{\beta}_{q3} t_{d1}^3, \ldots, \hat{\beta}_{10} + \hat{\beta}_{11} t_{dm_d} + \hat{\beta}_{12} t_{dm_d}^2 + \hat{\beta}_{13} t_{dm_d}^3, \ldots, \hat{\beta}_{q0} + \hat{\beta}_{q1} t_{dm_d} + \hat{\beta}_{q2} t_{dm_d}^2 + \hat{\beta}_{q3} t_{dm_d}^3)^\top$, which is of length qm_d. Further, based on the covariance matrix structured in this model, we can estimate the covariance matrix for $\boldsymbol{U}_d$ as $\hat{\boldsymbol{\Psi}}_d = [\tau_{sr}], s, r = 1, ..., qm_d$. Each element in $\boldsymbol{\Psi}_d$ corresponds to the covariance between a pair $Y_{dj_1k_1}$ and $Y_{dj_2k_2}$, which can be estimated as $\hat{\rho}_{j_1j_2} + \hat{\theta}^2 \mathbb{I}\{k_1 = k_2\} + \hat{\sigma}^2 \mathbb{I}\{j_1 = j_2, k_1 = k_2\}$, with domain indexes $1 \le j_1, j_2 \le q$, visit indexes $1 \le k_1, k_2 \le m_d$ and $\mathbb{I}\{\cdot\}$ being an indicator function.

Under the assumption of multivariate normal distribution for all observations measured over time, we now propose an extended longitudinal multivariate normative comparison (LMNC) statistic for testing whether the d-th participant has impaired cognition:

$$T_d = (\boldsymbol{U}_d - \hat{\boldsymbol{\mu}}_d)^\top \hat{\boldsymbol{\Psi}}_d^{-1} (\boldsymbol{U}_d - \hat{\boldsymbol{\mu}}_d) \sim \chi^2_{qm_d}, \tag{2.3}$$

which can be modified to an F test when number of participants is small in the healthy control group. For participant d, if we are generally concerned about whether this participant has cognitive functions with observations either too high or too low, we will use $(1-\alpha)$ quantile of $\chi^2_{qm_d}$ as the threshold for the significance level α. In practice, clinicians are usually more interested in screening for cognitive impairment with extremely low scores. One can conduct a statistical test considering the direction of domain scores by rejecting the null hypothesis if participant d's measured distance T_d exceeds the $(1-2\alpha)$ quantile of $\chi^2_{qm_d}$ and $\boldsymbol{U}_d^{'} \mathbf{1}_{qm_d} < \hat{\boldsymbol{\mu}}_d^{'} \mathbf{1}_{qm_d}$, where $\mathbf{1}_{qm_d}$ is the qm_d-vector of ones.

2.2.2 Permutation Testing

In practice, multivariate normality may not hold for the recorded measurements, and the test statistic in (2.3) might not follow an χ^2 distribution. In such case, the T_d statistic in (2.3) can still serve as a distance measure of individual scores to the norm, although we need to develop a new method to find the critical value for the test statistic without relying on a particular parametric distribution. We propose the innovative use of a permutation test to find such a critical value for each participant. Care should be taken in administrating a permutation test. In order that a test statistic from a permuted sample is comparable to the one from the original data, the permutation should respect the covariance structure of $\boldsymbol{\nu}_i$. Meanwhile, as the test statistic depends on the number of total visits m_d completed by the d-th participant, the permutation test should be done in a way specific to m_d.

For example, the covariance structure in Model (2.1), $\boldsymbol{\Sigma} = [\rho_{sr}]$, $s, r = 1, \ldots, q$, is set to be compound symmetric, where $\rho_{ss} = \rho_{r\,r}$ for $s, r = 1, \ldots, q$, and $\rho_{sr} = \rho_{ut}$ for $s, r, u, t = 1, \ldots, q$ and $s \neq r, u \neq t$. The compound symmetry is a reasonable covariance structure when all cognitive domain scores in the reference group have been standardized and their errors can be assumed to follow an identical distribution.

As the test statistic distribution varies by m_d, different thresholds are needed for each unique m_d observed from the testing group. Suppose there are M distinct numbers of visits in the testing group. We take M bootstrap samples, one for each unique number of visits. The following procedure details how permutation tests should be done for all of the participants in the testing group who have m total number of visits. We first take a bootstrap sample of the desired number N of participants with replacement (say 5,000) from the healthy control group. Then, we remove the time effect (i.e. $\beta_{j0} + \beta_{j1} t_{ik} + \beta_{j2} t_{ik}^2 + \beta_{j3} t_{ik}^3$ from model (2.1)) to obtain participant-specific errors over time for participant i from the bootstrap sample, $1 \leq i \leq N$. Next, to carry out the permutation test for each participant in the bootstrap sample, we consider errors of each domain function across all visits as a whole column. As a result, the multivariate longitudinal measures can be organized into a matrix of q-domain columns and m_i-visit rows. Then, we permute these q columns within the same participant so that this compound symmetric covariance structure will be sustained after each permutation.

For each participants i in the bootstrap sample, we sample m visits with replacement to repre-

sent the bootstrapped sample errors with the number of visits matching with that of those participants to be tested. The bootstrapped sample errors from the m visits can be stacked in a similar way as in equation (2.2) to a vector $\boldsymbol{V}_i = (E_{i11}, ..., E_{iq1}, E_{i12}, ..., E_{iq2}, ..., E_{i1m}, ..., E_{iqm})^\top$. Then, we calculate the bootstrap test statistic for the rearranged error sample from participant i as $T_i = (\boldsymbol{V}_i)^\top \hat{\boldsymbol{\Phi}}_{m_i}^{-1} \boldsymbol{V}_i$. However, the covariance structure $\boldsymbol{\Phi}_{m_i}$ used here is not the same as $\boldsymbol{\Psi}$ from equation (2.3), given that we draw errors with replacement for m times at the visit level within participant i. $\boldsymbol{\Phi}_{m_i}$ is a $mq \times mq$ matrix. For domain indexes $1 \le j_1, j_2 \le q$ and visit indexes $1 \le k_1, k_2 \le m$, its element can be estimated as $\hat{\rho}_{j_1 j_2} + \hat{\theta}^2 \left(\mathbb{I}\{k_1 = k_2\} + m_i^{-1} \mathbb{I}\{k_1 \ne k_2\}\right) + \hat{\sigma}^2 \left(\mathbb{I}\{j_1 = j_2, k_1 = k_2\} + m_i^{-1} \mathbb{I}\{j_1 = j_2, k_1 \ne k_2\}\right)$, where $\mathbb{I}$ is an indicator function. This covariance matrix $\boldsymbol{\Phi}_{m_i}$ cannot be inverted when $m > 1$ and the participant we bootstrapped has only one visit ($m_i = 1$). As a result, we will exclude participants with only one visit from the healthy control (reference) group when the permutation test is administered after longitudinal modeling. Considering that the number of participants with only one visit is small in a longitudinal setting, this exclusion seems to have a minimal impact on the MACS sample that we use.

With a sufficient number of permutation tests conducted, the $(1 - \alpha)$ quantile specific to m visits can be found among all $T_i \mathbb{I}\left\{(\boldsymbol{V}_i)^\top \mathbf{1}_{qm} < 0\right\}, i = 1, \ldots, N$, to serve as the critical value. Thus, we relax the assumptions that the test statistic follows a χ^2 distribution and that the upper tails and the lower tails of the domain scores are symmetric. Participant d with total $m_d = m$ visits will be classified as cognitively impaired if their test statistic exceeds this critical value and $\boldsymbol{U}_d^\top \mathbf{1}_{qm} < \hat{\boldsymbol{\mu}}^\top \mathbf{1}_{qm}$.

2.3 SIMULATION ANALYSIS

We ran a series of simulation studies to evaluate the performance of the proposed procedures. Given that the MACS data analyses in Section 2.4 involve 6 cognitive domains, in the simulation studies we also considered $q = 6$ hypothetical domains. We first generated longitudinal multivariate data following the multivariate normal distribution with several forms of polynomial mean functions over time. The testing procedure based on χ^2 was evaluated by FWER over different levels of α. Then, we considered data that do not follow multivariate normality to evaluate the performance

of the proposed permutation test. Two forms of data were examined. One type was generated from multivariate t distributions with symmetric but heavier tails than normal distributions. The other was generated by transforming Gamma distributions to get negative skewness.

We carried out 1,000 simulations for each scenario. For each simulation, we generated longitudinal scores for 1,000 participants supposedly from the healthy control group, and generated longitudinal scores for another 1,000 participants independently as the test group. For each participant, we simulated survival time from an exponential distribution with mean 30 years and censored at 15 years. Since participants in the MACS were tested semiannually (around 0.5 year between any consecutive two visits) or biannually on their cognitive performance Miller et al. (1990); Becker et al. (2014), the time between any consecutive two visits was assumed to follow independent and uniform (0,1) distribution with the first visit at time 0. We continued to simulate visits until the accumulated visit times exceeded the censored survival time for the ith participant. The number of visits at the last visit before the boundary was recorded as m_i.

In practice, one might be interested in determining whether cognitive functions are significantly better in one group compared to another. Thus, we examined and compared various testing groups with different visit frequencies and mean functions under alternatives and under the null. Detailed simulation specification and results are described below.

2.3.1 Multivariate Normal Distribution

After the set of visits m_i was generated for participant i, six domain scores were simulated from the multivariate normal distribution at each visit. The covariance matrix for $\boldsymbol{U}_i$ was specified as following. We set $\sigma^2 = 30$, $\theta^2 = 10$, $\rho_{sr} = 20$, for $s = r$, and $\rho_{sr} = 60$, for $s \neq r$ with $s, r = 1, ..., 6$. Each element for covariance matrix can then be computed. Diagonal elements are $\sigma^2 + \theta^2 + \rho_{11} = 100$. Covariance of different cognitive domains at the same visit is $\theta^2 + \rho_{12} = 30$. Covariance of the same cognitive domains at different visits is $\rho_{11} = 60$. The rest elements are $\rho_{12} = 20$.

We considered four types of polynomial mean functions over time. For the constant trend, all six cognitive domains are assumed to have mean 50 at any given t. For the linear trend, the first three cognitive domains are set to have means $50 - 0.3t$, and the other three have means $50 - 0.5t$. For the quadratic trend, the first three cognitive domains have means $50 - 0.02t^2 + 0.1t$, and the

other three have means $50-0.15t^2+0.2t$. Lastly, for the cubic trend, the first three have means $50-0.001t^3+0.05t^2+0.3t$ and the rest have means $50-0.0015t^3+0.07t^2+0.6t$.

The `mvrnorm` from the R library `MASS` was then used to generate longitudinal cognitive errors following the multivariate normal distribution with means 0 and the covariance matrix as described above. The mean polynomial functions with the four forms (see above) were added to the errors to represent the simulated longitudinal cognitive scores. For the healthy control group, the `lmer` from the library `lme4` was used to implement model (2.1). Without assuming any prior knowledge of the true longitudinal mean trend, cubic polynomial functions were used to describe the mean functions for all four sets of data. For each type of mean functions, the test statistics were then computed for 1,000 testing participants using the sample mean and covariance matrix obtained from the corresponding healthy control group. The χ^2 tests were conducted for each simulated dataset at different levels of α (from 0.001 to 0.1), and the average FWER was computed based on 1,000 simulations for each type of mean functions. Figure 1 illustrates the obtained FWERs of the LMNC χ^2 test across all α levels. The estimated FWERs are denoted by the black solid lines, and the nominal α levels are denoted by gray dash lines. The two lines are almost identical under the four mean trends. The LMNC χ^2 test seems to have exact FWER when domain scores follow multivariate normal distributions and the underlying means and covariance structure are correctly specified.

2.3.2 Multivariate *t* and Gamma Distributions

Real data often violate multivariate normal distributions. Skewness and heavy tails are often observed. In this simulation setting, we considered the same four mean functions described in Section 2.3.1, but non-normal errors were used in simulations. One set of errors has symmetric heavy tails from multivariate t distributions, and the other set has negative skewness transformed from correlated Gamma distributions.

We generated longitudinal random errors from multivariate t distributions with 5, 25 and 50 degrees of freedom. The covariance matrix for the error terms was assumed to follow the same structure as described in Section 2.3.1, and the means of the errors were set at 0. The `rmt` from the library `csampling` was used for multivariate t random error generation. Then we added four

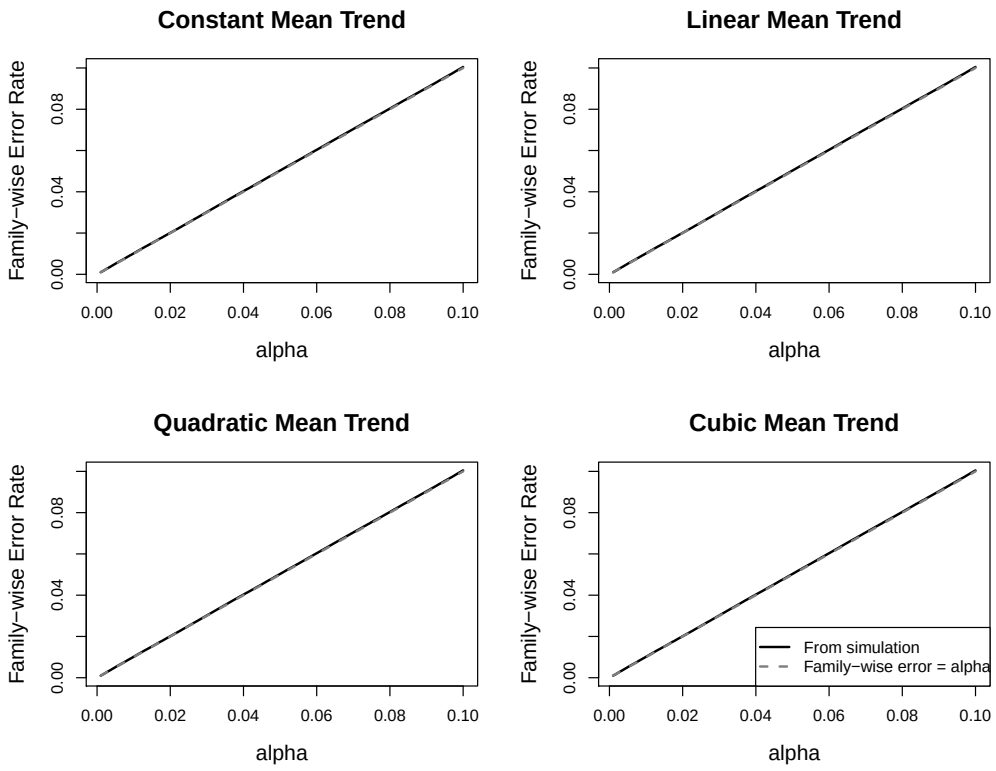

Figure 1: The LMNC χ^2 test when data follow multivariate normal distribution

polynomial mean trends to the simulated random errors to represent observed longitudinal scores with heavy symmetric tails.

Next, a gamma distribution was utilized to simulate data with negative skewness. In order to comply with certain covariance structure, i.e. compound symmetric, we first generated longitudinal multivariate normal errors ς_{ijk}, $j = 1,...,6, k = 1,..., m_i$ for participant i with the zero means. The covariance matrix from Section 2.3.1 divided by 100 was used here. Then we considered three different gamma distribution designs. For the first one, we calculated $70 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where Γ is the cumulative distribution function (CDF) of the gamma distribution with shape of 4 and scale of 5 and Φ is the CDF of the standard normal distribution. For the second design, we calculated $100 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where we assumed shape of 25 and scale of 2 for the gamma distribution. For the third design, we used $150 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where the gamma distribution has shape of 100 and scale of 1. The same longitudinal mean functions from Section 2.3.1 were again added to the simulated errors to obtain observed longitudinal cognitive domain scores with negative skewness. All three designs have baseline scores with mean 50 and variance 100.

For each scenario we generated longitudinal cognitive domain scores for 1,000 participants from the healthy control group and scores for the other 1,000 as the test group. Other simulation setups are the same as those from Section 2.3.1. To implement the permutation test, we first fit an MLME with cubic polynomial terms to data from the healthy control group as specified in Model (2.1) and obtained the estimates for the mean trends and the covariance matrix. Then, for each unique number of visits M observed in the test group, we bootstrapped 5,000 participants with replacement (N=5,000). For each participant, we subtracted the estimated mean trend from their longitudinal scores. The resulting errors were rearranged randomly as illustrated in Section 2.2.2, and then added back to the estimated mean trend to get permuted participant's scores. The α-th quantile was found among these 5,000 test statistics to serve as the threshold for cognitive impairment classification in the test group. After 1,000 simulations, summarized FWERs at various levels of α are shown in Figure 2 for data generated from multivariate t distributions and in Figure 3 for data generated from gamma distributions. For comparison, we also carried out the testing procedure based on χ^2 to examine how FWERs are controlled relative to different α levels. Their FWERs at various levels of α are also shown in respective figures.

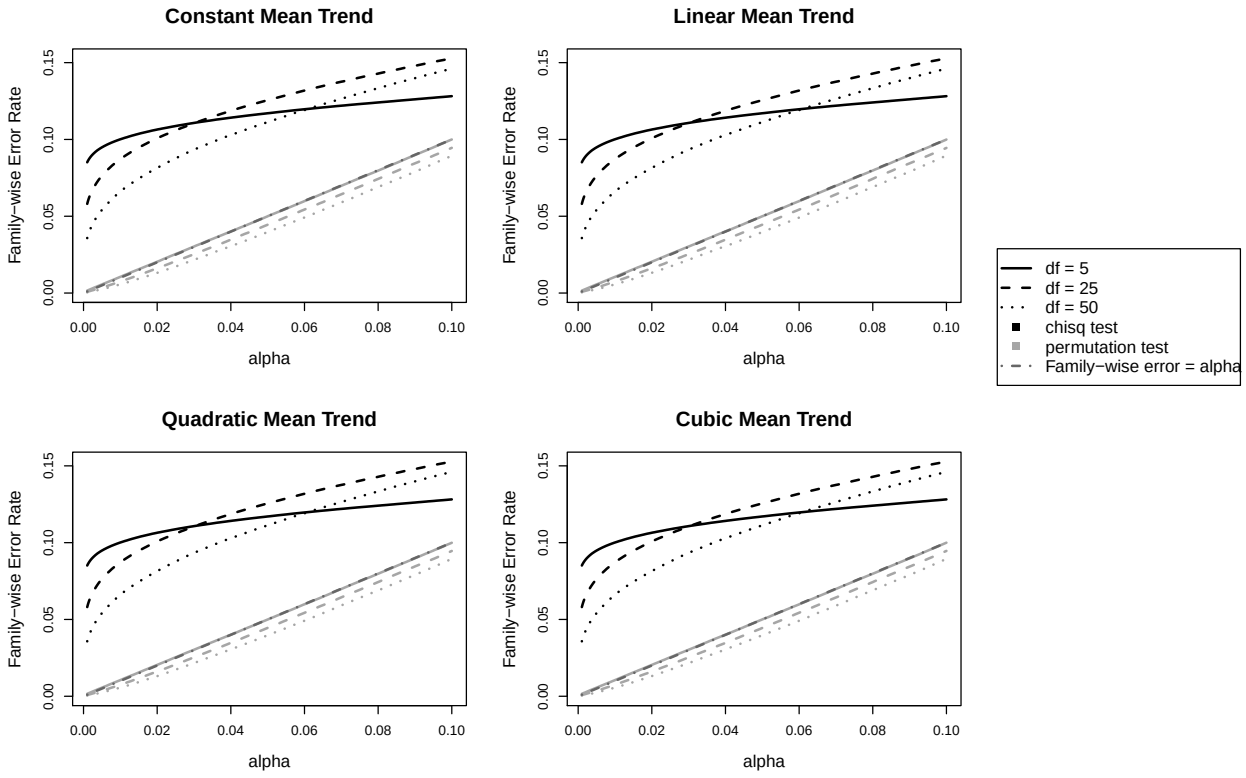

Figure 2: The LMNC χ^2 and permutation tests when data follow multivariate t distributions (permutation test when df=5 overlapped with the nominal α line)

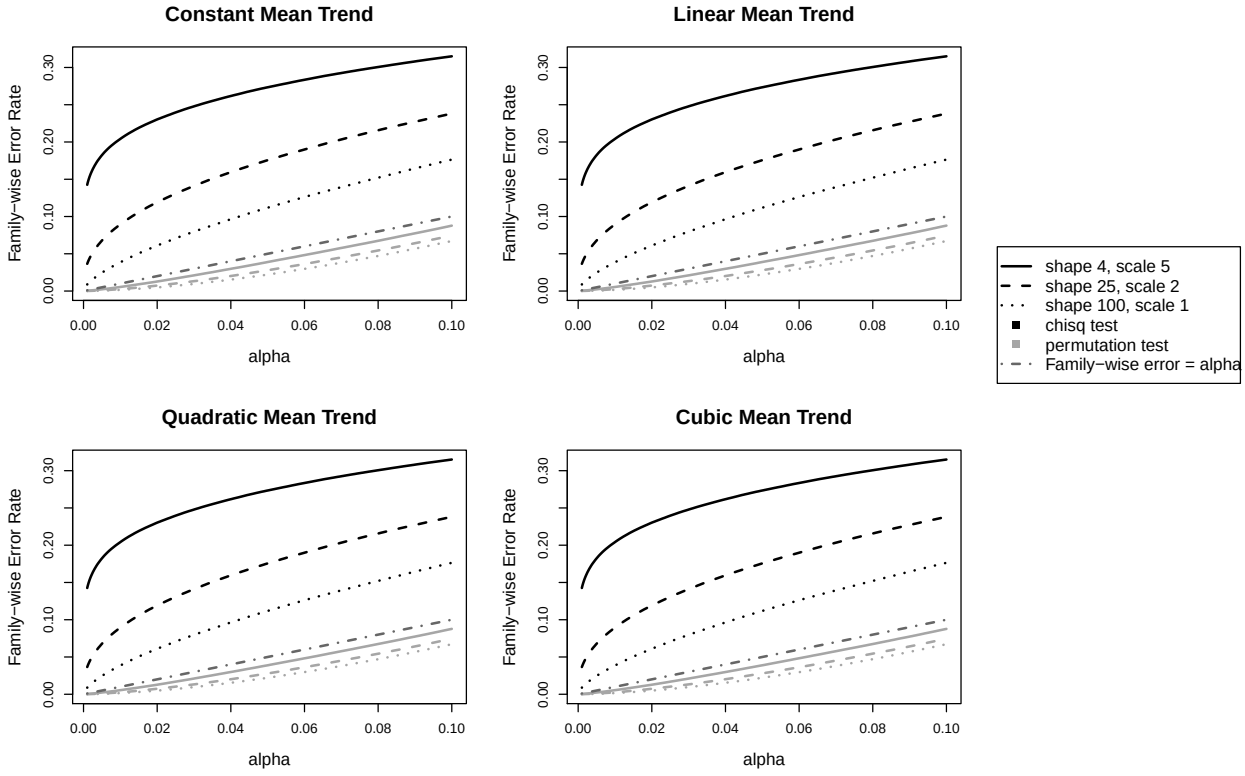

Figure 3: The LMNC χ^2 and permutation tests when data are transformed from Gamma distributions

When multivariate normality does not hold, the FWER based on the χ^2 procedure can be greatly inflated, as shown in Figures 2 and 3 where the three black curves denoting the FWERs from the χ^2 test are way above the empirical α levels denoted by the gray broken dash lines. Moreover, the inflation seems more drastic at smaller levels of α. On the other hand, the permutation test can successfully guard FWER at or below any pre-determined level as illustrated in Figures 2 and 3. Since the permutation test was carried out on the error terms, this suggests that Model 1 still works well in capturing the mean functions even when the data do not follow multivariate normal. Another interesting phenomenon about the permutation test that we observed from the plots is that FWER will be smaller compared to α when the multivariate t distribution has less heavy tailedness and the gamma distribution has less skew. In other words, when the data move closer to normality, the permutation test becomes more conservative. Though the conservativeness of permutation tests has been observed in early work Berger (2000), our permutation test is more complicated and the dependency on the skewness of the data requires further investigation. Therefore, it remains important to check the normality of data before determining whether the χ^2 or permutation test should be used when applying the LMNC for classification.

2.3.3 Comparing Groups under Different Visit Frequencies

In this section, we examined the power and the FWER of the proposed tests under different settings of visit frequency for the test group. The MACS study, which inspired us to develop the LMNC method, followed seronegative and seropositive participants at roughly the same frequency. Thus the two comparison groups have similar distributions for the numbers of visits as shown later in Section 2.4. However, this may not hold when a new study with certain treatment/condition is tested against an old study, because various factors can contribute to significant differences in visit frequencies. Even within the same study, participants from different cohorts may have different follow-up visits. Therefore, we carried out the following numerical studies to examine how different visit frequencies affect FWER as well as power if comparison between groups is desired. Four different designs were considered for the test group by changing mean survival time and censoring time:

1. Survival time follows exponential distribution with mean 30 years and is censored at 15 years

(median visit number 28);

2. Survival time follows exponential distribution with mean 50 years and is censored at 15 years (median visit number 29);
3. Survival time follows exponential distribution with mean 30 years and is censored at 10 years (median visit number 19);
4. Survival time follows exponential distribution with mean 30 years and is censored at 25 years (median visit number 42).

Here for the healthy control group, we adopted the same multivariate t setting with 5 degrees of freedom and the same quadratic mean trend from Section 2.3.1. Under the first design the test and control groups have an identical visiting frequency. To evaluate the FWER under the null, we generated 1,000 participants following the same mean trend and covariance structure as the test group at each simulation. The only difference is the survival time observed and subsequent visit frequency. To examine power under alternatives, we assumed the first, third and fifth cognitive domains of the test group to have mean trends of $50 - 0.02t^2$ and the rest cognitive domains to have mean trends of $30 - 0.04t^2$. As the two sets of domain scores had quite different means, we lowered the dependence in our covariance structure by setting $\sigma^2 = 60$, $\theta^2 = 10$, $\rho_{sr} = 5$, for $s = r$, and $\rho_{sr} = 25$, for $s \neq r$ with $s, r = 1, ..., 6$. Therefore, covariance of different cognitive domains at the same visit is $\theta^2 + \rho_{12} = 15$. Covariance of the same cognitive domains at different visits is $\rho_{11} = 25$. The remaining elements are $\rho_{12} = 5$. Again, we considered the four different designs of visit frequencies for the test group and 1,000 participants were simulated for each design at every simulation run.

At each simulation for every participant, both the χ^2-test and the permutation test based on 5,000 permutations were used for cognitive impairment classification. One thousand simulations were implemented and summarized in Figure (4). From the graph under the null hypothesis, we can see that frequency-specific permutation test can effectively control FWER at per-determined α level for all different survival time designs. Because data do not follow multivariate normal distribution, using the χ^2-test will inflate FWER and the inflation in cognitive impairment differs with various survival time designs and visit frequencies. Not surprisingly, the χ^2-test has more power than the permutation test under alternative hypotheses. Although close, groups with much different visit number distributions may have different power using permutation test. Therefore,

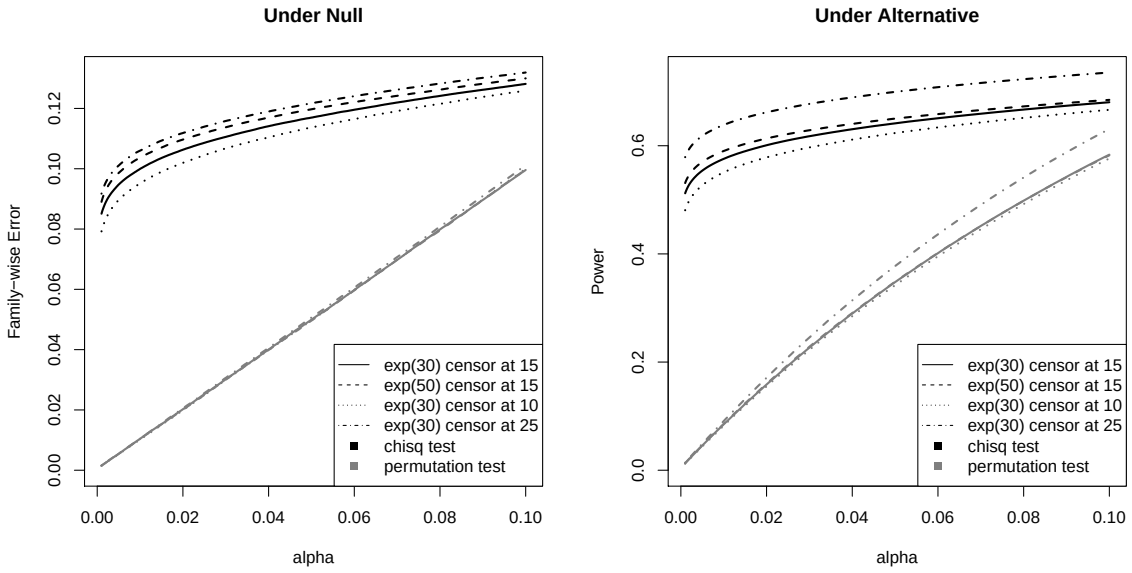

Figure 4: The LMNC χ^2 test and permutation test when multivariate t data have different visit frequencies

it is important to make sure that the visit number distributions are comparable when comparison between groups is desired.

2.4 APPLICATION TO THE MULTICENTER AIDS COHORT STUDY

We applied the proposed LMNC to the neuropsychological (NP) data that were collected from an ongoing Multicenter AIDS Cohort Study. The MACS study has been administered by the University of Pittsburgh, Johns Hopkins University, Northwestern University and the University of California at Los Angeles Kingsley et al. (1987); Kaslow et al. (1987). Since its first enrollment in 1984, the MACS has recruited more than 7,000 men who have sex with men (MSM), either infected with HIV or at risk for infection at study entry. participants have been regularly interviewed and examined semiannually about a broad range of variables including their age, depressive symptoms, sexual activity, substance use, cognitive functioning and physical measurements. HIV infection

negatively impacts patients' brain, and the effect of HIV on brain functioning was found to be less drastic after the highly active antiretroviral therapy (HAART) became available in early 1990's. In a MACS NP substudy, participants have been repeated tested on a NP test battery assessing six cognitive domains which included learning, motor speed & coordination, speed of information processing, memory, working memory & attention, and executive functioning Farinpour et al. (2003); Popov et al. (2019). As of October 2017, some participants had more than 20 years of longitudinal NP data. This provides a unique opportunity to examine how cognitive impairment compares between those infected with HIV and those not infected in the HAART era.

At each NP visit, the battery of tests was administrated, and these test scores were summarized by T-scores which were calculated from regression models adjusting for education, race/ethnicity, age, and number of tests administrated, and standardized to have mean 50 and standard deviation of 10. Then, summary T-scores were obtained from taking the arithmetic mean of all T-scores in each domain, except for motor speed & coordination domain, where the lowest T score is used. As a result, assuming multivariate normal distribution on the MACS NP data is of concern for the motor domain score which tends to be negatively skewed. Thus, permutation may work better for cognitive impairment classification to control FWER at a pre-determined level.

In this analysis, we focus on visits where participants had all six cognitive domain scores available, and include 3,701 participants who have at least one such visit. Among participants included in this analysis, 1,667 were seronegative (279 having one visit), while 2,034 were infected with HIV (328 having one visit) at the study entry. Those not infected with HIV serve as the "healthy" control group, representing HIV-uninfected MSM. Because motor speed & coordination domain used the lowest T score instead of the average, we can see from Figure 5 that baseline motor domain score for seronegative participants failed to follow a normal distribution. The LMNC using the χ^2-test may be of concern and the permutation test should be considered. For both seropositive and seronegative groups, we calculated the number of participants for each total visit frequency and plotted them by group in Figure 5. The Kolmogorov-Smirnov test shows that the visit number distributions do not differ ($p = 0.90$), and the visit frequencies are comparable between the two groups. Thus, the LMNC permutation test is expected to work well in this application.

Specifically, we first fit the model described in (2.1) with cubic mean trends in the healthy control group. After estimates were obtained, both the χ^2 test and permutation test were applied to data

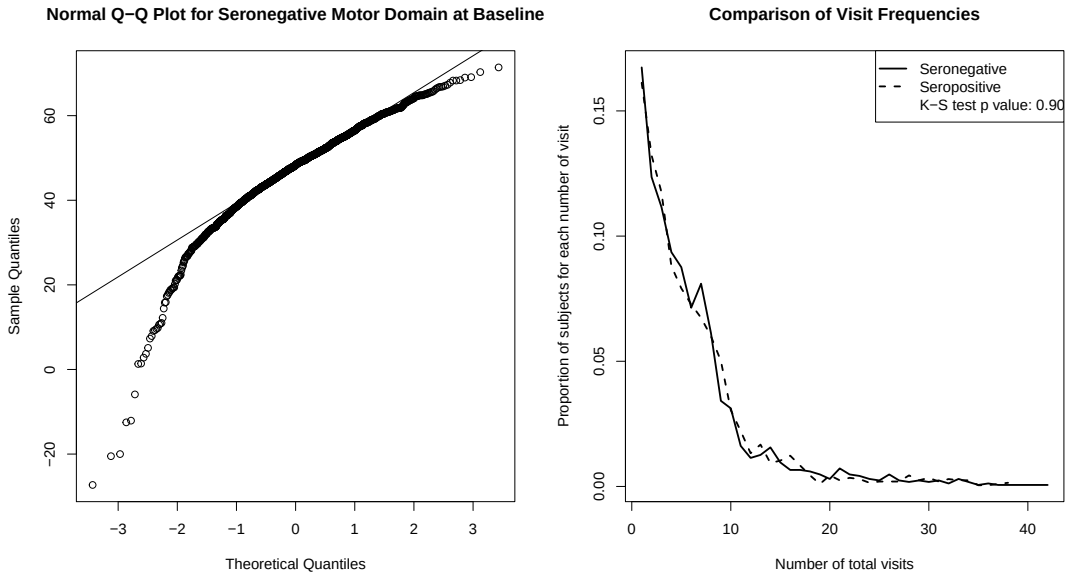

Figure 5: Q-Q plot of baseline motor score in the seronegative group and visit frequencies of two serostatus groups

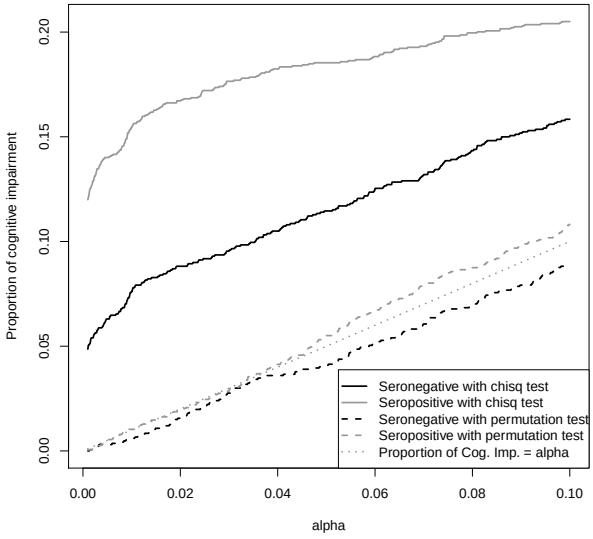

Figure 6: Comparing proportion of cognitive impairment in seronegative and seropositive groups in the MACS

Table 1: Mean scores of six cognitive domains for seronegative and seropositive groups at different visit

	Cognitive Domain	Motor	Executive	Speed	Learning	Memory	Working Memory
Visit 1	**Seronegative**	47.12	49.81	49.92	49.67	49.90	49.64
	Seropositive	46.73	49.77	49.25	49.71	49.98	49.40
Visit 4	**Seronegative**	45.86	50.17	50.45	49.33	49.04	48.99
	Seropositive	45.79	49.31	49.33	48.96	49.04	48.36
Visit 10	**Seronegative**	48.14	53.14	51.25	50.94	50.74	51.86
	Seropositive	48.26	51.93	51.00	52.41	52.38	51.51

from the healthy control group across different levels of α. For both tests, 5-fold cross validation was used to test cognitive impairment among those not infected with HIV. The results are shown in Figure 6. The first thing we can see is that the permutation test (N=100,000) can effectively control FWER at pre-determined α levels. By contrast, the χ^2 test would have inflated the family-wise error when the data fail to follow multivariate normal distribution but the model is sufficiently specified. We also applied both the permutation test and the χ^2 test to data from seropositive men. The results are also shown in Figure 6. The permutation test identified about the same proportion of seropositive men with cognitive impairment as in the seronegative group across αs. Meanwhile, the χ^2 test identified a much higher proportion of cognitively impaired men in the seropositive group than in the seronegative group. This has serious clinical and research implications. Not only would the χ^2 test identify more people with cognitive impairment in seronegative and seropositive groups, but also wrong conclusions might be drawn about the relationship between serostatus and cognitive impairment during the HAART era. By contrast, the permutation test shows that the association between cognitive impairment and HIV infection becomes insignificant, leading to the conclusion that people infected with HIV can enjoy equally healthy cognitive functioning after being properly treated with cART. To further support this conclusion, Table 1 shows the mean scores of all six cognitive domains for both seronegative and seropositive groups at the first visit (100% participants), the forth visit (50% participants) and the tenth visit (15% participants). We can see that, relative to standard deviation of 10, the score differences are very small between the two groups. A wrong conclusion would be drawn if a method failing to control family-wise error, like the χ^2 test in this case, is used otherwise.

2.5 DISCUSSIONS

From the numerical studies, we can see that our proposed LMNC method can effectively control FWER. Multivariate normality is a key assumption in using the χ^2 test for cognitive impairment classification. When such an assumption is not satisfied by data or the model in use does not fully address random effects, the permutation test can still guard FWER at a pre-determined level.

The MNC method specifically takes inter-correlations among domain scores into account, and

may lead to different results as some existing methods that are used in AIDS research. As an example, we only consider two cognitive domains at a single visit. Suppose that the variance of two domain scores is 1 and the correlation is 0.5, and both mean cognitive scores are zero. The participant having cognitive scores of $(-1,-2)$ will have a larger p value than the one with scores $(0,-2)$. This is contrary to the intuition that the first participant seems to have more extreme scores. However, the correlation between two domains is high. Thus, the scores $(0,-2)$ from the second participant is more unusual than $(-1,-2)$ under the strong positive correlation, and consequently, the second participant has a longer "distance" from the means, after inversely "weighted" by the covariance matrix. If the correlation between two domains is set to be zero, then the first participant will have a smaller p value. Therefore, the MNC results may not be consistent with some existing ad-hoc diagnoses methods such as counting number of domains with scores 1 or 1.5 standard deviations below the means Antinori et al. (2007); Gisslén et al. (2011).

This paradox also exists in a longitudinal setting. For illustration purpose, let us assume that only one cognitive domain is tested, with a mean of 0 and variance of 1. The correlation between any two visits is 0.5. One participant with the domain score tested at two visits as $(0,-2)$ will have a larger p value than another participant with the domain score tested at three visits as $(0,0,-2)$. This is also against the intuition as the first participant seems to have worse cognition earlier. However, the second participant has longer records of being "normal", so the "distance" from the means is also larger after weighted by the inverse covariance matrix. Consequently, the second participant has a smaller p value. If domain scores are independent among all visits, the p value for the second participant would be larger, because of more visits and larger degree of freedom when performing the χ^2 test. This may serve as an explanation to why we observed higher power under a higher visit frequency design, even through they follow the same mean trends. To generalize our proposed method to groups with very different visit number distributions, further efforts should be made to improve our proposed permutation test. Nevertheless, our proposed LMNC method provides insights into how "abnormal" domain scores may be, which could be missed by naive methods ignoring inter-correlations among domain scores and repeated visits.

3.0 DYNAMIC ARRAYED COMPARISON

3.1 INTRODUCTION

Motivated by the Multicenter AIDS Cohort Study (MACS), researchers in psychiatry and brain science fields are often interested in cognitive impairment classification with family-wise error rate (FWER) under control. Such classification can further guide treatments on these patients infected with HIV to effectively prolong their life expectancy. Given normalized neuropsychological scores collected from a battery of tests across several cognitive brain domains, existing popular methods use t tests on each domain separately (Antinori et al., 2007; Gisslén et al., 2011). This approach fails to control family-wise error by not taking into account inter-correlations among all brain domains. Huizenga et al. (2007) introduced a method called the Multivariate Normative Comparison (MNC) by incorporating the covariance structure in the test statistic.

For patient i, we assume q cognitive brain domains are tested and normalized neuropsychological scores are summarized into a vector of $\mathbf{X}_i$. Patients are separated into a healthy group of size n for reference and a testing group which can be of a disease population in need of particular attention for treatment. If X_i follows identical multivariate normal distribution $N(\mu, \Psi)$ independently, mean and covariance matrix in healthy control can be estimated by $\hat{\mu}$ and $\hat{\Psi}$. Then we can construct an F-statistic to test whether subject i from the testing group has impaired cognition

$$\frac{n(n-q)}{(n+1)(n-1)q}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}) \sim F(q, n-q).$$

Under the assumption that cognitive brain domain scores follow multivariate normal distribution, the MNC method will control family-wise error effectively at a pre-determined level. Su et al. (2015) and Wang et al. (2019) have studied and demonstrated the effectiveness of the MNC with

cross-sectional neuropsychological data collected among AIDS patients. However, patients may visit the same doctor more than once so that their cognitive status is closely monitored. In order to extend the MNC method from cross-sectional data to a longitudinal setting, we proposed in Chapter 2 a longitudinal MNC (LMNC) method by modeling multiple neuropsychological scores with a multivariate linear mixed effects (MLME) model. Correlations among different brain domains and across all visits within the same subject are explicitly considered within the model. A permutation procedure is used when the model cannot sufficiently explain the data. For research purposes, we often carry out analysis after all necessary information has been collected and look at data retrospectively. Therefore, the longitudinal MNC method is adequate for providing cognitive impairment classification with proper control of family-wise error.

In practice, this is probably not enough, especially when treatment should be prescribed in the early stage of cognitive decline. In other words, patients cannot wait until the end of study to know their cognitive status. Rather classification of impaired cognition should be done fluidly at each visit. This will have practical use in Just-in-Time Adaptive Interventions (JITAIs) which track health status on mobile device (Klasnja et al., 2015; Nahum-Shani et al., 2018). Proper identification of departures from normal health can lead to effective treatments. For an ongoing study like the MACS, the 30-year history of data can give us insight about participants' behaviors, such as how their cognitive functions change over time, how long they might survive, and how frequently they visit research centers.

Reinsel (1982b) built theoretical foundations for MLME model for the analysis of multiple responses with repeated measures and unrestricted parameters. Heitjan and Sharma (1997) then studied multiple correlated series and proposed a linear model with autoregressive error structure to be estimated by maximum likelihood method. Fieuws and Verbeke (2004, 2006) jointly modeled multiple longitudinal responses with certain dependence structure specified in the model. Fang et al. (2006) further considered a modified expectation-maximization approach to estimate MLME parameters with constraints on intercepts. In the review of multivariate longitudinal analysis, Verbeke et al. (2014) mentioned that longitudinal correlations among multiple outcomes can be better addressed by joint modeling.

In an ongoing study, we assume we have collected sufficient data from the same or similar cohort. Using survival analysis and frequency modeling techniques, such as cox proportional

hazard regression and Poisson regression, we can predict how many visits each subject will have. MLME model will be used to characterize longitudinal cognitive domain scores while accounting for inter-correlations among various cognitive domains and repeated measures for the same subject. A test statistic can be built from MLME estimations for each subject. We can then apply correction procedures to control family-wise error.

We will organize the rest of the chapter as following. In Section 3.2, frequency prediction and MLME models will be formulated. Then, we will construct test statistics for dynamic arrayed comparisons (DAC) and the Bonferroni-type adaptive procedure. Permutation test will be considered when data do not follow multivariate normal distributions. In Section 3.3, numerical studies will be carried out to examine the performance of the proposed tests. Applications of DAC to a neuropsychological substudy in the MACS are shown in Section 3.4. In the end of this chapter, we will conclude with some discussions on the performance of DAC and future work.

3.2 FAMILY-WISE ERROR CONTROLLING PROCEDURE

3.2.1 Dynamic Arrayed Comparison Based on χ^2

Before starting to identify cognitive impairment prospectively, we assume certain information has been collected from the same or a similar cohort in history. Following the notation from Chapter 2, n subjects enrolled in a healthy reference group and was evaluated on q cognitive domains over m_i visits. Cognitive domain j is measured as $Y_{ijk}, i = 1,...,n; j = 1,...,q; k = 1,...,m_i$ for subject i at k-th visit. A multivariate normal distribution is assumed on cognitive scores since they are generally normalized in practice. Within the same subject, the MLME model is used for cognitive functions from different domains over time with a covariance structure capturing dependences among cognitive domains and repeated measures for different visits. Thus we have:

$$Y_{ijk} = \beta_{j0} + \beta_{j1} t_{ik} + \beta_{j2} t_{ik}^2 + \beta_{j3} t_{ik}^3 + \nu_{ij} + \delta_{ik} + \epsilon_{ijk}. \tag{3.1}$$

Here we use q polynomial functions of degree 3 to characterize the mean cognitive scores over time. If desired, polynomials with a higher degree can be added. The B-spline technique, besides

polynomial, can also approximate the true average cognitive scores over time (Bloxom, 1985; De Boor, 2001; Shumaker, 2007; Rutherford et al., 2015; Harrell, 2015). We assume ϵ_{ijk}, representing random error from each observation, to be independent and identically distributed following normal $N(0,\sigma^2)$. We also assume δ_{ik}, which are errors specific for each visit, to follow independent and identical normal $N(0,\theta^2)$. Due to the inter-correlations among various cognitive domains for the same subject, $\boldsymbol{\nu}_i = (\nu_{i1},..,\nu_{ip})^\top$ is assumed to follow multivariate normal $N(\mathbf{0},\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = [\rho_{sr}], s,r = 1,...,q$. The structure of covariance matrix depends on research designs and data collected, and can be unspecified, auto-regressive or compound symmetric.

Fang et al. (2006) and Fieuws and Verbeke (2004) provided estimation procedures used to estimate unknown parameters from the MLME model. Estimated parameters are denoted as $\hat{\beta}_{j0},\hat{\beta}_{j1},\hat{\beta}_{j2},\hat{\beta}_{j3}, j = 1,\ldots,q,\hat{\rho}_{sr}, s,r = 1,\ldots,q$, $\hat{\theta}^2$, and $\hat{\sigma}^2$. Assuming subject d is tested, we take all q cognitive scores observed over m_d visits, and stack them into m_d vectors

$$\boldsymbol{U}_\omega^d = (Y_{d11},...,Y_{dq1},Y_{d12},...,Y_{dq2},...,Y_{d1\omega},...,Y_{dp\omega})^\top, \omega = 1,...,m_d. \tag{3.2}$$

From the MLME model in (3.1), the estimated mean vector of $\boldsymbol{U}_\omega^d$ is written as $\hat{\boldsymbol{\mu}}_\omega^d = (\hat{\beta}_{10}+\hat{\beta}_{11}t_{d1}+\hat{\beta}_{12}t_{d1}^2+\hat{\beta}_{13}t_{d1}^3,\hat{\beta}_{20}+\hat{\beta}_{21}t_{d1}+\hat{\beta}_{22}t_{d1}^2+\hat{\beta}_{23}t_{d1}^3,\ldots,\hat{\beta}_{q0}+\hat{\beta}_{q1}t_{d1}+\hat{\beta}_{q2}t_{d1}^2+\hat{\beta}_{q3}t_{d1}^3,\ldots,\hat{\beta}_{10}+\hat{\beta}_{11}t_{d\omega}+\hat{\beta}_{12}t_{d\omega}^2+\hat{\beta}_{13}t_{d\omega}^3,\ldots,\hat{\beta}_{q0}+\hat{\beta}_{q1}t_{d\omega}+\hat{\beta}_{q2}t_{d\omega}^2+\hat{\beta}_{q3}t_{d\omega}^3)^\top$, which is of length $q\omega$. Moreover, from the covariance matrix specified in this model, we can estimate the covariance matrix for $\boldsymbol{U}_\omega^d$ as $\hat{\boldsymbol{\Psi}}_\omega^d = [\tau_{sr}], s,r = 1,...,q\omega$. Each element in $\boldsymbol{\Psi}_\omega^d$ corresponds to the covariance between a pair $Y_{dj_1k_1}$ and $Y_{dj_2k_2}$, which can be estimated as $\hat{\rho}_{j_1j_2}+\hat{\theta}^2\mathbb{I}\{k_1 = k_2\}+\hat{\sigma}^2\mathbb{I}\{j_1 = j_2, k_1 = k_2\}$, with domain indexes $1 \le j_1, j_2 \le q$, visit indexes $1 \le k_1, k_2 \le \omega$ and $\mathbb{I}\{\cdot\}$ being an indicator function.

With the assumption of multivariate normal distribution on q longitudinal cognitive functioning scores, we stated in Chapter 2 that the LMNC test statistic for subject d at visit ω is

$$G_\omega^d = (\boldsymbol{U}_\omega^d-\hat{\boldsymbol{\mu}}_\omega^d)^\top(\hat{\boldsymbol{\Psi}}_\omega^d)^{-1}(\boldsymbol{U}_\omega^d-\hat{\boldsymbol{\mu}}_\omega^d) \sim \chi^2_{q\omega}, \omega = 1,...,m_d, \tag{3.3}$$

which can be further adjusted to an F test when we only have a small number of subjects in the healthy control group. In order to construct a DAC test statistic to be used for cognitive impairment identification at each visit ω, we use the difference between consecutive LMNC test statistics and set $S_\omega^d = G_\omega^d - G_{\omega-1}^d$ for $2 \le \omega \le m_d$ and $S_1^d = G_1^d$. We can show that they are independent from each

other. Without loss of generality, denote $\boldsymbol{X}_\omega = \boldsymbol{U}_\omega - \boldsymbol{\mu}_\omega = (\boldsymbol{X}_{\omega-1}^\top, \boldsymbol{W}_\omega^\top)^\top$ for $\omega \geq 2$ and $\boldsymbol{X}_1 = \boldsymbol{W}_1$. The covariance matrix of $\boldsymbol{X}_\omega$ is

$$\boldsymbol{\Psi}_\omega = \begin{pmatrix} \boldsymbol{\Psi}_{\omega-1} & \boldsymbol{\Delta}_\omega \\ \boldsymbol{\Delta}_\omega^\top & \boldsymbol{\Phi}_\omega \end{pmatrix}, \omega \geq 2,$$

with $\boldsymbol{\Psi}_1 = \boldsymbol{\Phi}_1$. Then for $\omega \geq 2$,

$$\begin{aligned} S_\omega &= G_\omega - G_{\omega-1} \\ &= \boldsymbol{X}_\omega^\top \boldsymbol{\Psi}_\omega^{-1} \boldsymbol{X}_\omega - \boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1} \\ &= \boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{\Delta}_\omega \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1} - 2\boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{\Delta}_\omega \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{W}_\omega + \boldsymbol{W}_\omega^\top \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{W}_\omega, \end{aligned}$$

where $\boldsymbol{\Theta}_\omega = \boldsymbol{\Phi}_\omega - \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{\Delta}_\omega$. We also know from properties of multivariate normal distribution that $\boldsymbol{X}_{\omega-1}$, with covariance $\boldsymbol{\Psi}_{\omega-1}$, and $\boldsymbol{W}_\omega - \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1}$, with covariance $\boldsymbol{\Theta}_\omega$, are independent. Therefore,

$$\begin{aligned} &\boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1} = G_{\omega-1} \\ &\left(\boldsymbol{W}_\omega - \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1}\right)^\top \boldsymbol{\Theta}_\omega^{-1} \left(\boldsymbol{W}_\omega - \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1}\right) \\ &= \boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{\Delta}_\omega \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{\Delta}_\omega^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{X}_{\omega-1} - 2\boldsymbol{X}_{\omega-1}^\top \boldsymbol{\Psi}_{\omega-1}^{-1} \boldsymbol{\Delta}_\omega \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{W}_\omega + \boldsymbol{W}_\omega^\top \boldsymbol{\Theta}_\omega^{-1} \boldsymbol{W}_\omega \\ &= S_\omega \sim \chi_q^2. \end{aligned}$$

As a result, we can claim DAC test statistics $\{S_\omega^d, \omega = 1, ..., m_d\}$ are independent.

If m_d is known, we can construct visit-by-visit testing procedures easily and apply the Bonferroni or Banjamini-Hochberg procedure to control family-wise error rate (Bonferroni, 1936; Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). In order to identify cognitive impairment for subject d at visit ω, we will use $(1 - 2\alpha_\omega^d)$ quantile of χ_q^2 as the threshold for the significance level α while $\mathbf{1}_{q\omega}^\top \hat{\boldsymbol{X}}_\omega^d < \mathbf{1}_{q(\omega-1)}^\top \hat{\boldsymbol{X}}_{\omega-1}^d$ when $\omega > 1$ or $\mathbf{1}_q^\top \hat{\boldsymbol{X}}_1^d < 0$, since clinicians are more interested in screening subjects with lower cognitive scores after adjusting for previous ones. However, m_d is generally unknown for a prospective study.

3.2.2 Frequency Prediction

Given that we have observed some data with respect to our interested population in an ongoing study, the difficulty of unknown m_d can be addressed by using some frequency model to estimate the expected number of visits each patient will have based on their baseline characteristics and historical data $\mathscr{H}$. One may first use a survival modeling approach to predict the mean residual lifetime ($\hat{T}$) of each patient and then use Poisson regression to predict the frequency of visiting during a unit of time ($\hat{\Lambda}$). For example, we assume that the survival time follows a proportional hazards model

$$\lambda(t|\boldsymbol{Z}_1) = \lambda_0(t)\exp(\boldsymbol{\beta}^\top \boldsymbol{Z}_1), \tag{3.4}$$

and that the number of visits per time unit follows a log-linear model

$$\log(\Lambda|\boldsymbol{Z}_2) = \gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{Z}_2, \tag{3.5}$$

where covariate vectors $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ can be completely different or share some predictors. Commonly used techniques like (partial) maximum likelihood estimation can be used to obtain the estimates $\hat{\boldsymbol{\beta}}$, $\hat{\gamma}_0$ and $\hat{\boldsymbol{\gamma}}_1$, and the Nelson-Aalen type estimator can be computed for $\hat{\lambda}_0(t)$. As a result, the product of the expected survival time

$$\hat{T}_i = \int \exp\{-\hat{\Lambda}_0(t)e^{\hat{\boldsymbol{\beta}}^\top \boldsymbol{Z}_{1i}}\}dt$$

and the expected frequency at a unit of time

$$\hat{\Lambda}_i = \exp(\hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_1^\top \boldsymbol{Z}_{2i})$$

can be used to predict the number of visits, i.e., $\hat{N}_i = \hat{T}_i\hat{\Lambda}_i$ for subject i. We assume $\hat{N}_i$ is an unbiased estimator for subject's expected number of future visits. The nearest integer greater than $\hat{N}_i$ is used if the prediction is a decimal number and it is denoted by $\lceil \hat{N}_i \rceil$. Based on the estimated expected number of visits, we propose a Bonferroni-type adaptive procedure on controlling family-wise errors. We are interested in identifying first significant cognitive impairment in general and will stop after the first null rejection.

3.2.3 Bonferroni-type Adaptive Procedure

The Bonferroni procedure seems to be a natural choice in controlling family-wise error with expected number of visits known (Bonferroni, 1936; Dunn, 1959, 1961). If a uniform $\frac{1}{\lceil \hat{N} \rceil_i}\alpha$ is applied to every visit, the subjects who have fewer visits than expected will have a much smaller family-wise error than those who have more visits than expected. Therefore, we propose a new procedure to control family-wise error rate, where subjects having more visits than expected will be subject to more stringent testing. We define M_i to be the actual number of visits for subject i, which is unknown at the beginning of the study. For subject i at any visit $1 \le j \le M_i$, if $j \le \lceil \hat{N}_i \rceil$, the p value at this visit is compared with $\frac{C(\lceil \hat{N}_i \rceil - j + 1)}{\lceil \hat{N}_i \rceil^2}\alpha$. When $j > \lceil \hat{N}_i \rceil$, the p value is compared with $\frac{C}{\lceil \hat{N}_i \rceil^2}\alpha$. C is chosen to be $\frac{2c\lceil \hat{N}_i \rceil^2}{\lceil \hat{N}_i \rceil(\lceil \hat{N}_i \rceil + 1)}$. When $c = 1$ and under the null hypothesis H_0 that subject is not impaired cognitively at any visit, we have

$$
\begin{aligned}
& P\left(\text{identified cognitive impairment during the study} \mid Z_{1i}, Z_{2i}, H_0, \mathscr{H}_S\right) \\
=& 1 - \prod_{j=1}^{\infty} \left[1 - P\left(\text{identified cognitive impairment at visit } j \mid Z_{1i}, Z_{2i}, H_0, \mathscr{H}_S\right)\right] \\
=& 1 - \mathbb{E}\left[\prod_{j=1}^{M_i} \left(1 - \alpha\left(\frac{C(\lceil \hat{N}_i \rceil - j + 1)}{\lceil \hat{N}_i \rceil^2}\mathbb{I}\left(j \le \lceil \hat{N}_i \rceil\right) + \frac{C}{\lceil \hat{N}_i \rceil^2}\mathbb{I}\left(j > \lceil \hat{N}_i \rceil\right)\right)\right) \mid Z_{1i}, Z_{2i}, H_0, \mathscr{H}_S\right] \\
\le& 1 - \prod_{j=1}^{\lceil \hat{N}_i \rceil} \left(1 - \alpha\frac{C(\lceil \hat{N}_i \rceil - j + 1)}{\lceil \hat{N}_i \rceil^2}\right) \\
\le& \alpha,
\end{aligned}
$$

where $\mathbb{I}(\cdot)$ is an indicator function. The first equation is valid because of the independence of DAC test statistics. The second last inequality establishes because of Jensen's inequality and $\hat{N}_i$ is assumed to be an unbiased estimator of $\mathbb{E}[M_i] = N_i$. We can choose c to be a greater number adaptively such that power is improved with family-wise error rate strictly controlled. However, if number of visits is small, the loss of power under independent tests is expected to be small even with $c = 1$. Alternatively, we can also replace expected mean survival $\hat{T}_i$ with median survival time, which is supposedly smaller as a survival time distribution is often right skewed. Median survival time is easier to estimate and will tend to have less stringent family-wise error while controlled at a pre-determined level.

3.2.4 Permutation Test

Generally, it can be hard to justify multivariate normal distribution for recorded data and the testing procedure may fail to follow an χ^2 distribution. Intuitively, the test statistic can still measure the departure from the norm, while there is no known distribution to use in finding thresholds. Here, we propose an innovative use of permutation tests to obtain a series of critical values for the test statistics over time, when we cannot assume multivariate normality.

First, we bootstrap B (i.e. 10,000) participants from the study with replacement. For the b-th participant bootstrapped, we will remove the time effect obtained from model (3.3) (i.e. $\hat{\beta}_{j0} + \hat{\beta}_{j1} t_{bk} + \hat{\beta}_{j2} t_{bk}^2 + \hat{\beta}_{j3} t_{bk}^3$) to obtain subject-specific errors over M_b visits. Under the assumption that the covariance matrix $\boldsymbol{\Sigma} = [\rho_{sr}], s, r = 1, ..., q$ characterizing cognitive domains follows a compound symmetry structure, we can permutate the errors in the following way without disrupting the covariance structure. We rearrange the errors as a matrix with M_b rows representing all visits and q columns representing all cognitive domains. We permutate the columns in whole and then premutate the rows in whole to preserve the assumed structure.

Then, based on the permutated errors, we obtain an error vector as

$$\boldsymbol{V}_b = (E_{b11}, ..., E_{bq1}, E_{b12}, ..., E_{bq2}, ..., E_{b1M_b}, ..., E_{bqM_b})^\top.$$

The DAC test statistics can be calculated as $\{S_\omega^b \mathbb{I}\{(E_{b1\omega}, ..., E_{bq\omega})^\top \mathbf{1}_{qm} < 0\}, \omega = 1, ..., M_b\}$ without assuming any specific error distributions. Pooling them together after B bootstraps, we can calculate any α_0 based on the predicted number of visits from models (3.5) and (3.4). The $(1 - \alpha_0)$ quantile may serve as a critical value. Participant d at the ω-th visit will be identified as cognitively impaired if this visit-specific test statistic exceeds this critical value while $\mathbf{1}_{q\omega}^\top \hat{\boldsymbol{X}}_\omega^d < \mathbf{1}_{q(\omega-1)}^\top \hat{\boldsymbol{X}}_{\omega-1}^d$ when $\omega > 1$ or $\mathbf{1}_q^\top \hat{\boldsymbol{X}}_1^d < 0$.

As we have pointed out in Section 3.2.3, the factor c, which is used to control how much we can spend α, can be set at 1 under multivariate normal distribution, because the independence of tests guarantees the loss of power is small. However, this independence becomes questionable when the multivariate normal distribution cannot be justified. As a result, c should be adjusted to preserve power while controlling family-wise error. Cross-validation can be used here for determining an appropriate c value. We first randomly divide the healthy reference group by several folds. For

each fold, we apply the MLME model to the rest of participants and calculate DAC test statistics for this fold. Then permutation test statistics are built from the healthy reference group leaving out the fold to be tested. Putting the DAC test statistics together with their corresponding permutation test statistics, an iterative process is used to determine an appropriate c value such that the family-wise error rate is controlled just around the pre-determined level.

3.3 NUMERICAL STUDIES

Here, we ran some simulation studies to assess how the proposed DAC method works under various distribution assumptions. As described in Section 3.4, the MACS study evaluates 6 cognitive domains regularly among participants, so we also considered $q = 6$ cognitive domains in our simulations. First, longitudinal multivariate data were generated with various forms of mean score functions over time. When a multivariate normal distribution was assumed, the DAC testing procedure based on χ^2 was evaluated by family-wise error over various levels of α. Otherwise, the permutation testing procedure was used to evaluate DAC. Multivariate t and Gamma distributions were considered for non-normal errors with heavier tails or skewness.

For these simulations, we assumed 1,000 participants have had their cognitive functioning tested in the past and they can serve as the healthy controls. At the same time, we generated longitudinal scores for 1,000 more subjects as the testing group assuming they will enroll in the study in the future. For each participant in the healthy control and testing groups, we first simulated their enrollment time uniformly over 20 years. Their survival time follows a Weibull distribution with five covariates,

$$\log(T) = \beta_0 + \sum_{i=1}^{5} \beta_i z_i + \sigma W,$$

where W was generated from the standard extreme value distribution. This error distribution gives the proportional hazard interpretations for all covariates. We set $\beta_0 = 3$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.2$, $\beta_5 = -0.2$ and $\sigma = 0.1$. Covariates are independently generated. z_1-z_4 follow the standard normal distribution and z_5 follows a uniform (0,1) distribution. Participants were assumed to be censored at year 20 for both historical and newer cohorts.

Based on the time in the study, each visit time was generated such that the time between two

studies follows independent exponential distribution with the first visit happening at time 0. The hazard rate follows Poisson regression from equation (3.5). We used the same five covariates were used, and set $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.1$, and $\gamma_5 = -0.1$. At all the visit times, 6 cognitive domain scores were generated from different multivariate distributions as detailed below. One thousand simulations were carried out for each scenario.

As we have mentioned in Section 3.2.4, it's important to determine an appropriate c value to relax the thresholds and to preserve power. At the same time, study designs or participants might have an impact on the study duration and visit frequency in the future. Therefore, we examined various scenarios, where the duration of the prospective study changes, or when participants visit the study more and less frequently. FWER and power were examined under the null and the alternative, respectively. The details of each simulation and results are described below.

3.3.1 Multivariate Normal Distribution

After obtaining the number of visits m_i for participant i, we generated six domain scores from the multivariate normal distribution at each visit. For the covariance matrix $\boldsymbol{U}^i_{m_i}$, we set $\sigma^2 = 30$, $\theta^2 = 10$, $\rho_{sr} = 60$, for $s = r$, and $\rho_{sr} = 15$, for $s \neq r$ with $s, r = 1, ..., 6$. Covariance of different cognitive domains at the same visit is $\theta^2 + \rho_{12} = 25$. Covariance of the same cognitive domains at different visits is $\rho_{11} = 60$. The rest elements are $\rho_{12} = 15$.

Four forms of polynomial mean trends were considered. For the constant trend, all six cognitive domains are assumed to have mean 50 at any given t. For the linear trend, the first three cognitive domains are set to have means $50 - 0.6t$, and the other three have means $50 - 0.8t$. For the quadratic trend, the first three cognitive domains have means $50 - 0.08t^2 + 0.2t$, and the other three have means $50 - 0.06t^2 + 0.1t$. Lastly, for the cubic trend, the first three have means $50 - 0.008t^3 + 0.08t^2 + 0.55t$ and the rest have means $50 - 0.007t^3 + 0.06t^2 + 0.55t$. The `mvrnorm` from the R library `MASS` was then used to simulate longitudinal cognitive errors following the multivariate normal distribution with means 0 and the covariance matrix $\boldsymbol{U}^i_{m_i}$. The mean polynomial functions mentioned above were then added to the errors to represent the generated longitudinal cognitive scores.

`lmer` from the library `lme4`, `coxph` from the library `survival`, and `glm` were used to

implement models (3.1), (3.4) and (3.5). The MLME, which is used for modeling mean scores, has cubic polynomial functions specified without assuming the true polynomial functions are known. At various levels of α (from 0.001 to 0.1), χ^2 tests were conducted for each simulated subject in the newer cohort. Results based on 1,000 simulations are summarized in Figure 7. The estimated FWERs are denoted by the black solid lines, and the nominal α levels are denoted by gray dash lines. The DAC χ^2 test can have FWER controlled below the pre-determined level when domain scores follow a multivariate normal distribution and we can correctly model the visit frequency and multiple longitudinal domain scores. As expected, the departures of the FWERs from the nominal level re small. The FWERs of four scenarios are around 0.046 when $\alpha = 0.05$.

3.3.2 Multivariate t and Gamma Distributions

In practice, multivariate normal distributions can be difficult to justify and real data may present skewness and heavy tails. Here, we considered two sets of non-normal errors. One set follows multivariate t distributions for heavy tails, while the other set present negative skewness from correlated Gamma distributions. The same four mean trends from Section 3.3.1 are used here.

To simulate longitudinal scores with heavy tails, multivariate t distributions with 5, 25 and 50 degrees of freedom were used. The `rmt` from the library `csampling` was used for multivariate t random error generation. Means of the random errors were set to 0, and the covariance matrix used here is the same as $\boldsymbol{U}^i_{m_i}$ from Section 3.3.1. The four polynomial score trends were then added to the generated errors as the observed longitudinal scores.

To simulate longitudinal scores with negative skewness, gamma distributions were used. Since we have considered a special covariance structure, i.e. compound symmetric, we transformed longitudinal multivariate normal errors to get correlated gamma errors. We first simulated multivariate standard normal errors $\varsigma_{ijk}, j = 1, ..., 6, k = 1, ..., m_i$ with means 0 and covariance $\boldsymbol{U}^i_{m_i}/100$ from Section 3.3.1. Three gamma distribution designs were considered. The first one took transformation of $70 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$, where Γ is the cumulative distribution function (CDF) of the gamma distribution with shape of 4 and scale of 5 and Φ is the CDF of the standard normal distribution. For the second design, we used $100 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ as our negative skewed errors, where Γ has shape of 25 and scale of 2. We calculated $150 - \Gamma^{-1}(\Phi(\varsigma_{ijk}))$ for the third design, where Γ has

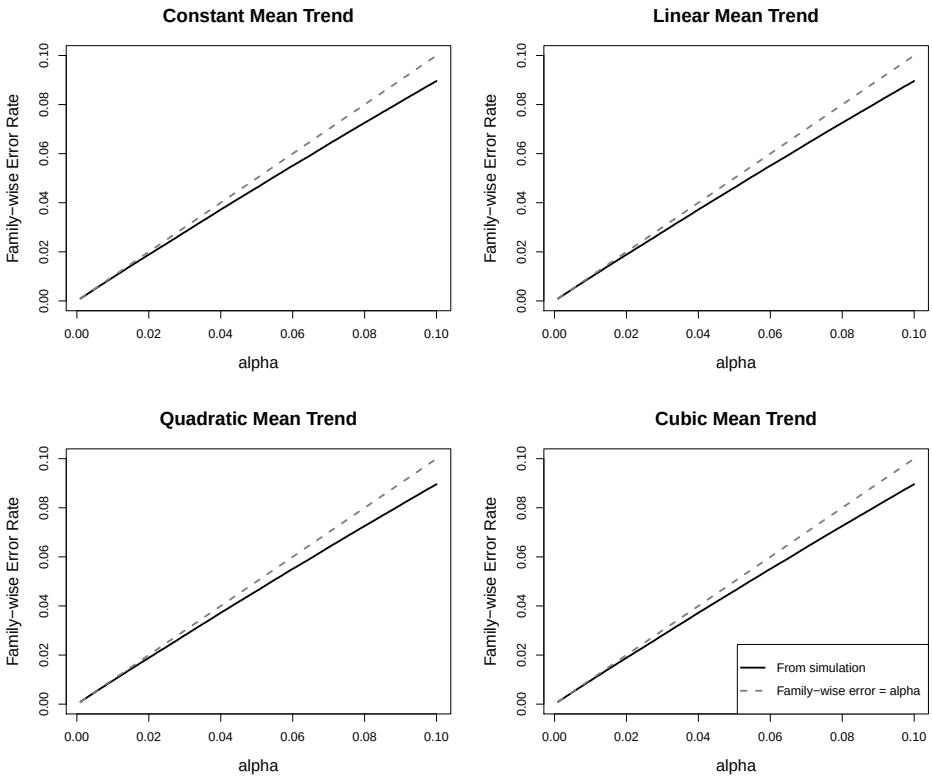

Figure 7: The DAC χ^2 test when data follow a multivariate normal distribution

shape of 100 and scale of 1. Each score error has mean of 0 and variance of 100. The four polynomial score trends were again added to the generated scores to represent measured longitudinal domain scores with negative skewness.

For each setting, survival time and visit frequency were generated in the same way as mentioned at the beginning of Section 3.3. We assumed 1,000 participants have been measured in the historical healthy control group, while 1,000 more participants were going to enroll in the new study and get tested. As described in Section 3.2.4, the proposed permutation test was used here. We bootstrapped 10,000 participants from the historical healthy control with replacement. For each participant selected, we got the longitudinal errors by subtracting estimated means from the original scores. Then we rearranged the errors, added back the longitudinal mean scores and got a series of DAC test statistics. After repeating 10,000 times, visit-by-visit classification of cognitive impairment is done in the newer cohort with respective quantiles from these permutation test statistics as the thresholds. The results of FWER at various α levels after 1,000 simulations for multivariate t and correlated gamma distributions are summarized in Figure 8 and 9, respectively. As a comparison, results from the DACs with χ^2 tests are also shown in the figures to illustrate how well permutation tests did in controlling FWER.

When multivariate normality cannot be justified as shown in Figure 8 and 9, the FWER based on DAC with χ^2 test can be greatly inflated. Illustrated by the black curves, FWER inflation is smaller when the multivariate t distribution has less heavier tails or when the gamma distribution is less skewed. When the permutation test is used instead, FWER can be strictly controlled below any pre-specified α level. This conservativeness has been observed by other work (Berger 2000), and we can relax the thresholds by increasing the factor c through cross-validation as described in Section 3.2.4, so that FWER is adjusted around the pre-determined level. In Section 3.3.3, we will discuss how we can adjust the value of c for power analysis.

3.3.3 Different Number of Visits and Power Analysis

In practice, not just multivariate normality is often violated by collected measurements, but also a future study design may change as compared with historical ones, such as in study duration and visit frequency. For example, in the MACS study, there were several enrollment waves as

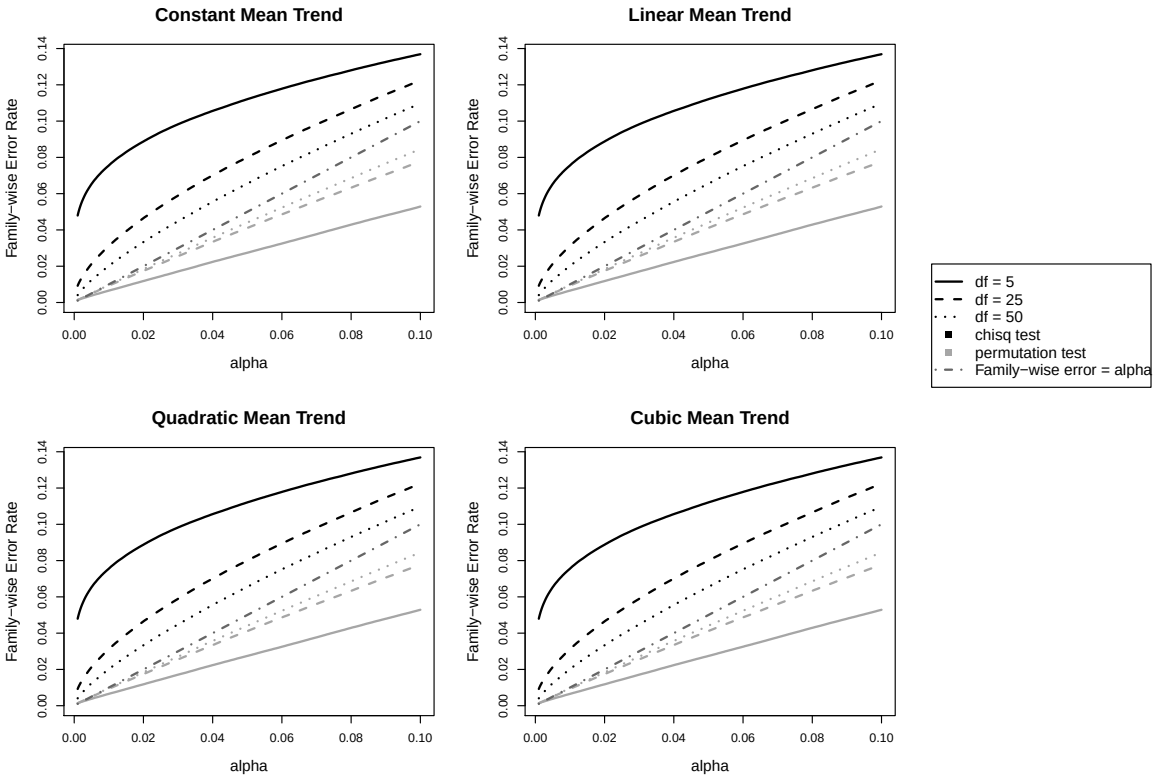

Figure 8: The DAC χ^2 and permutation tests when data follow multivariate t distributions

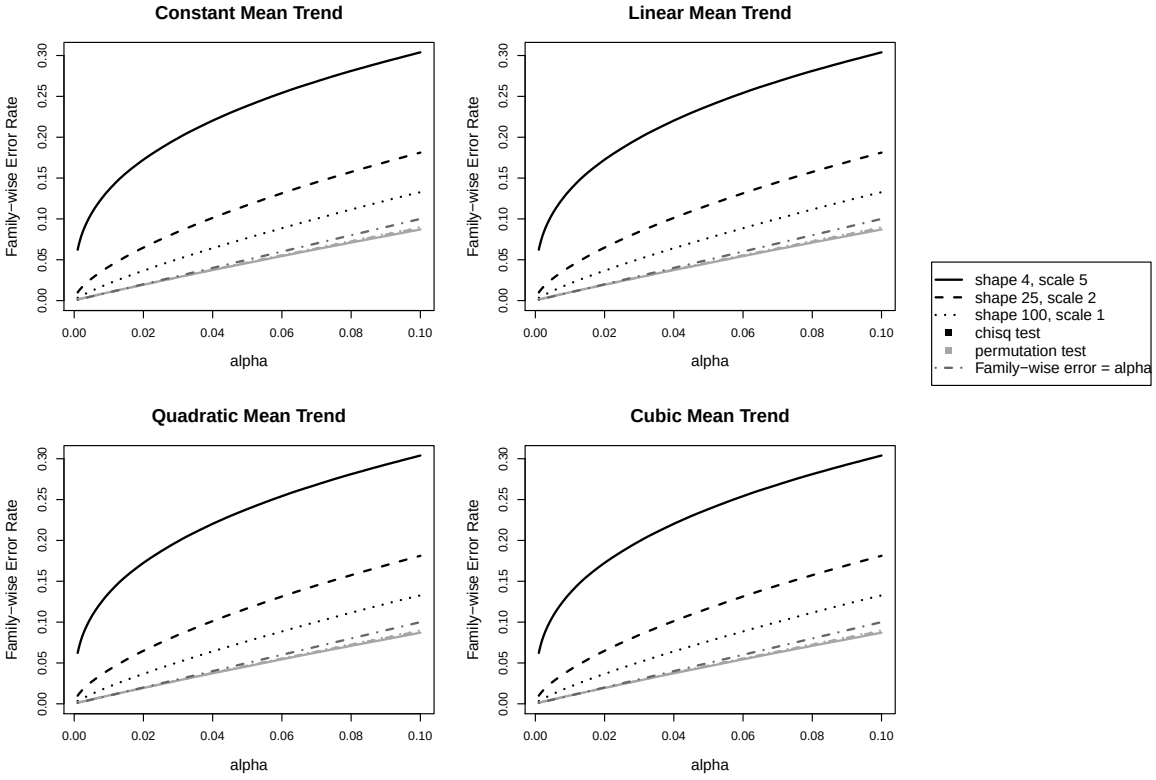

Figure 9: The DAC χ^2 and permutation tests when data are transformed from Gamma distributions (permutation test when shape=25 and when shape=100 overlapped)

time goes and we have more participants enrolling in this study in the later cohort. If we use data collected prior to today, which is of more than 30-year record, we may want to follow up 20 or 40 more years for future participants. At the same time, the MACS proposed some study design changes, so participants visit the centers less frequently than in 1990s. In this section, we examined the FWER and the power of the proposed DAC method under settings of different study durations and visit frequencies in the newer cohort. Five different designs were evaluated for the newer testing group by changing the study duration (survival distribution remains the same as in the previous two sections) and the visit frequencies through Poisson regression parameters. For historical healthy control group, the survival time and visit frequency remain the same as in the previous two sections. The distributions of five covariates remain the same as above. Here are the five settings in the newer cohort:

1. **Original:** In the Weibull model used for survival time, we set $\beta_0 = 3$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.2$, $\beta_5 = -0.2$ and $\sigma = 0.1$. Participants enroll uniformly until we end the study in 20 years. In the Poisson model used for visit frequency Λ, we set $\alpha_0 = 0$, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.1$, and $\alpha_5 = -0.1$.
2. **Shorter Study:** Same as setting (1), except we would end the study in 10 years.
3. **Longer Study:** Same as setting (1), except we would end the study in 30 years.
4. **Less Visits:** Same as setting (1), except we divided the visit frequency Λ by 2.
5. **More Visits:** Same as setting (1), except we multiplied the visit frequency Λ by 2.

Multivariate t with 5 degrees of freedom was used here to generate longitudinal scores in the historical healthy control group and in the testing group from the prospective study. For both groups under the null, the setup is the same as in Section 3.3.2. We first simulated the number of visits based on each scenario of survival and visit frequency. Then we generated multivariate t errors, which were added with quadratic mean trends from Section 3.3.1. For the testing group under the alternative, the setup is the same except for mean trends. We specified first three cognitive domains to have means $20 - 0.08t^2$, and the other three to have means $50 - 0.1t^2$. One thousand participants were assumed to have enrolled in the historical study, and 1,000 participants were expected to enroll in the new study. As mentioned in Section 3.2.4 and shown in Figure 8, the FWER is conservative with the permutation test and cross-validation can be used to determine an

appropriate factor c used for thresholds. After 5-fold cross-validation on 1,000 simulations, we determined that $c = 2.0$. In prediction of number of visits within the newer study, we assumed policy changes had already been known, such as when the study ends and how often the frequency of visits changes. The results of FWER and power with the permutation test (10,000 times) after 1,000 simulations under five cases are shown in Figure 10.

As illustrated in Figure 10, different study durations or visit frequencies have an impact on FWER and power. Less visits in the newer study, originated from a shorter study period or lower visit requirements, seem to have inflated FWER and larger power. However, the FWER inflation is in general small and cases (2)-(5) are quite different from case (1). Thus, it remains important to make sure the visit frequency does not deviate too much from the historical study. At the same time, factor c used to relax the thresholds can be increased or reduced slightly based on researchers' understanding of how study policy changes the visit frequency in the future.

3.4 APPLICATION TO THE MULTICENTER AIDS COHORT STUDY

Here, the proposed DAC method was applied to the neuropsychological (NP) sub-study data collected from the Multicenter AIDS Cohort Study (MACS), which is an ongoing study with first enrollment happening in 1984. The MACS has been administered by John Hopkins University, Northwestern University, The University of California at Los Angeles and the University of Pittsburgh (Kingsley et al., 1987; Kaslow et al., 1987). More than 7,000 who have sex with men (MSM) have been recruited in the study. Participants were either infected with HIV or at risk for infection enrollment. They have been regularly interviewed and examined about a wide range of variables, such as drug use, depressive symptoms, age, sexual disorder, cognitive functioning and physical measurements. Participants' cognitive functioning is negatively impacted by HIV infection. However, highly active antiretroviral therapy (HAART) was found to have positive effects on cognitive functioning among people infected with HIV since its first availability in early 1900's. Participants in the NP substudy have been regularly evaluated with a NP test battery for six cognitive domains, including motor speed & coordination, speed of information processing, executive functioning, learning, memory, and working memory & attention (Farinpour et al., 2003; Popov et al., 2019).

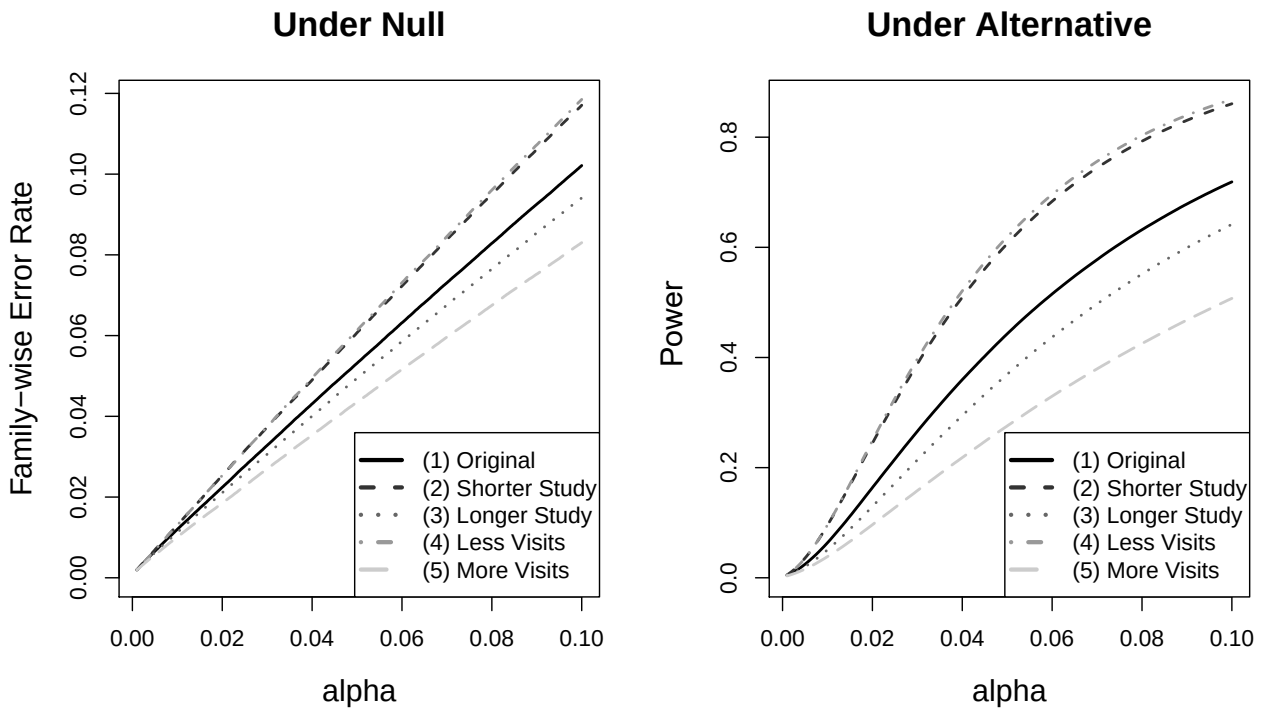

Figure 10: FWER and power of the DAC permutation tests when data follow multivariate *t* distributions and newer cohort has different study period or visit frequency (Case (2) and Case (4) are close to each other)

These test scores provide a rare opportunity to assess, in the HAART era, how those infected with HIV and those without differ in terms of cognitive impairment.

The MACS has four enrollment waves starting at 1984, 1987, 2001 and 2010. We took the first two pre-HAART cohorts as our historical data, and looked at the latest two cohort data prospectively to examine how cognitive impairment is developing over time after HARRT. During each NP visit, a battery of tests was carried out, and collected scores were summarized by T-scores, which were computed from regression models adjusting for ethnicity, education, age and number of tests administered. The T-scores have mean 50 and standard deviation of 10. For motor speed & coordination, the lowest T score is used for summary, while the other five use the arithmetic mean of all T-scores in each specific domain as summary T-scores. The multivariate normal distribution assumption on NP data from the MACS is of concern, because the summary T-scores in motor speed & coordination are very skewed. Consequently, the permutation test may work better for identifying cognitive impairment with FWER controlled at a pre-specified level.

For this analysis, only the participants with complete scores on six cognitive domains were included. We used participants prior to 2001 as the historical cohort for visit-by-visit cognitive impairment classification starting 2 001. Prior to 2001, 1,231 were infected with HIV, while 870 were not. Five-fold cross-validation on the 870 participants without HIV infection was used to find an appropriate factor c to relax the thresholds on testing. For each fold, we used Cox proportional hazard regression to model the survival and Poisson regression to model the visit frequency on the rest participants. For survival modeling, participants were censored at 4 years past the last NP visit or 2001, whichever is earlier, if death was not observed or death happened beyond 4 years past the last NP visit or 2001. Covariates of Cox regression included CD4+ cell count along with its quadratic transformation, age at first NP visit, Center for Epidemiologic Studies Depression (CES-D) score, hepatitis C status, four testing centers, and HIV serostatus. For Poission regression, we included quadratic transformation, age at first NP visit, four testing centers, and HIV serostatus as covariates. MLME was applied to the rest participants without HIV infection to estimate the mean trends and covariance structure. Based on the frequency prediction and MLME results, we treated each fold as if they were newer cohort and conducted visit-by-visit classification of cognitive impairment using DAC and the permutation test. After summarizing the rates from 5 folds on healthy controls from the historical study, we found that we can relax the factor c to 1.4

while keeping FWER around the pre-specified level. The results shown in Figure 11 are based on this cross-validated c of 1.4.

Then, we applied DAC to these participants enrolled from 2001, where 803 were infected with HIV and 796 were not. However, due to study policy changes, participants on average halved their frequencies to the NP substudy since 2001. Given this knowledge, we also halved the predicted frequencies for these participants. We anticipated that the study would conclude at 2017, and truncated the predicted survival times by 2017. The results are also shown in Figure 11 using the permutation test and the factor c of 1.4 from cross-validation. First, we can observe that the newer cohorts have significantly more people with cognitive impairment identified compared with the historical healthy control. Second, since 2001, the impairment rates between seronegative and seropositive group are not significantly different if we set $\alpha = 0.05$. This is consistent with the findings from Wang et al. (2019) and Chapter 2. Table 2 shows the mean scores for all six cognitive domains at visits when participants were evaluated around the same time. Because participants were measured at a half frequency in the newer cohort due to policy changes since 2001, the number of visits reflects this in Table 2. Across three comparison points, we can see that domains scores form the newer cohort are generally lower than the historical seronegative group, while motor speed & coordination exhibits significant difference and this difference increases as time progresses. However, seronegative and seropositive participants in this newer cohort do not seem to differ that much .

3.5 DISCUSSION

LMNC proposed in Chapter 2 considered a research setting, where data has been collected and we only need one diagnosis for each patient by looking retrospectively. As a comparison, the proposed DAC method is more practical, in a way that we can provide visit-by-visit diagnosis to the patients. This can encourage participation of preventive care or enable doctors to prescribe treatments on time. Not just for identifying cognitive impairment, the DAC method, along with the LMNC method, has a broad application in classification, as long as measurements collected are of longitudinal nature.

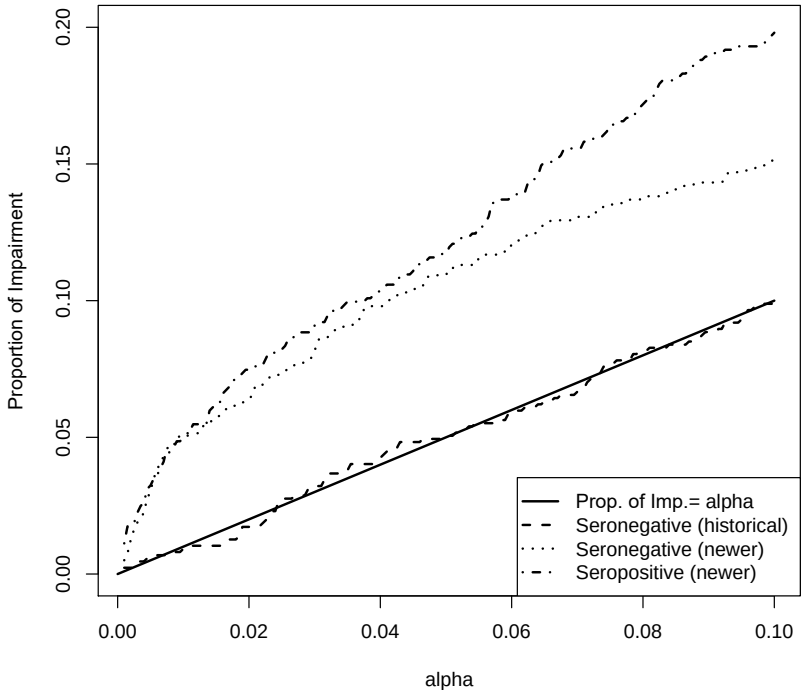

Figure 11: Comparing proportion of cognitive impairment in the MACS

Table 2: Mean scores of six cognitive domains for seronegative and seropositive groups at comparable times

Group (visit, % available)	Motor	Executive	Speed	Learning	Memory	Working Memory
Seronegative (historical, visit 1, 100%)	48.06	50.18	50.45	51.08	51.31	49.28
Seronegative (newer, visit 1, 100%)	46.10	49.41	49.34	48.12	48.35	50.04
Seropositive (newer, visit 1, 100%)	46.11	49.17	48.42	48.78	48.78	50.00
Seronegative (historical, visit 5, 46%)	50.34	52.12	51.11	50.98	51.02	51.08
Seronegative (newer, visit 3, 68%)	44.04	49.42	49.54	48.39	48.54	49.10
Seropositive (newer, visit 3, 72%)	44.21	47.99	48.31	48.57	48.21	48.74
Seronegative (historical, visit 9, 16%)	50.51	54.72	52.21	52.29	51.45	52.55
Seronegative (newer, visit 5, 44%)	42.81	50.06	50.14	48.68	48.85	47.94
Seropositive (newer, visit 5, 53%)	42.78	48.57	48.40	48.41	48.64	46.92

The proposed DAC method can effectively control FWER. Multivariate normality is an important assumption for the χ^2 test, although FWER is slightly lower than the pre-determined level. When such an assumption is violated, permutation test can also control FWER as shown in the simulation studies. It remains crucial to select an appropriate factor to relax the thresholds while keeping FWER under control. Cross-validation is a good way to exploit data collected from the historical healthy control group and set up this factor, but it doesn't generalize well to the newer study when study duration or visit frequency changes due to new study protocol. As a result, researchers need to decide whether they want to adjust such factor and how much if they do before conducting the DAC method.

4.0 QUANTIFYING DIAGNOSTIC ACCURACY IMPROVEMENT OF NEW BIO-MARKERS FOR COMPETING RISK OUTCOMES

4.1 INTRODUCTION

For clinicians, introducing a new biomarker into a statistical model may change the risks associated with various outcomes of interest, and subsequently may influence treatment decisions. Risk prediction algorithms using statistical modeling are among the most popular tools to evaluate significance of biomarkers. Although effect size and statistical significance are important, they do not provide direct information on the contribution of new biomarkers to diagnostic accuracy. For the latter, we are interested in evaluating the improvement in correctly "classifying" patients into several outcome categories, such as dementia, death and "nonevent," with the additional information from new biomarkers. In contrast, risk prediction algorithms typically attempt to predict the risks associated with each outcome in the course of time.

To investigate accuracy improvement over the course of variable additions for binary outcomes, the commonly used Receiver Operating Characteristic (ROC) curve and its corresponding Area Under the Curve (AUC) were shown to be insensitive to detecting the added values of new markers (Greenland and O'Malley, 2005; Pepe et al., 2004; Ware, 2006), and novel indicators were developed to complement the AUC measure (Pencina et al., 2008), such as the net reclassification improvement (NRI) and the integrated discrimination improvement (IDI). The NRI is the improvement in classification rates of disease categories by the "new" model which incorporates additional markers over those by the "old" model without the additional markers. On the other hand, the IDI quantifies the improvement in the integrated sensitivity minus that of specificity over all possible cutoff values, from the model without new biomarkers to the model with new biomarkers. Both indices have become popular in medical fields and been extended from categorical outcomes to

survival outcomes (Pencina et al., 2011; Uno et al., 2013).

However, there are few works in quantifying accuracy improvement for competing risks outcomes. Shi et al. (2014b) were among the first to consider accuracy improvement for competing risks, where the population is divided into two groups at a fixed time point – the "disease" group including subjects who have developed the event of interest, and the "healthy" group including those who have not had any event and those who have experienced competing events. Such a definition of the "healthy" group, which is in line with the augmented "at-risk" set in a popular regression model by Fine and Gray (1999) for competing risks data, is reasonable if competing events are not of interest, and those who have developed competing events are more or less similar to those who have not failed yet. However, there are many situations where we would like to separate subjects with competing events from those without any events. As an example, the Multicenter AIDS Cohort Study (MACS) involves two endpoints, death and dementia, where the age of dementia onset may be competing-risk censored by death. When the dementia onset is of concern, it does not seem appropriate to group those subjects who died with those who were alive and stayed healthy. Ideally they could be treated as separate categories in evaluation of accuracy improvement.

Li et al. (2013b) proposed reclassification statistics for assessing improvements in diagnostic accuracy for multi-level outcomes. Here we specifically consider how the definitions of the NRI and the IDI for multi-category outcomes can be extended to the competing risks setting, where one event prevents others from occurring. The detailed definitions are given in Sections 4.2.2 and 4.2.3 for two competing risks outcomes. One issue with estimating the adapted NRI and IDI is that independent censoring often occurs in additional to competing risks censoring, and a subject's disease status may not be determinable if this subject was censored before the time of interest. As detailed in Sections 4.2.2 and 4.2.3, the "missingness" due to censoring can be overcome by using the method of inverse probability of censoring weighting.

Demler et al. (2017) have evaluated the feasibility of establishing U-statistics theory under different assumptions for changes in the NRI and the IDI. If the models are under the alternative, both the NRI and the IDI are non-degenerate and variance estimators based on the U-statistics theory should work, though some adjustments are needed for the IDI. The bootstrap technique is valid under this situation. On the other hand, if the null model is the true model, and it is nested within the alternative model, both the NRI and the IDI are degenerate and the theoretical formulas for

estimating their variance do not apply. This raises special concerns in practice, because, in evaluating the accuracy improvement associated with new biomarkers, we are comparing the "new" model with the additional variables and the "old" model without them. Since these two models are nested, the degeneracy of the NRI and the IDI under the null should be and can be remedied as suggested by Demler et al. (2017), which will be evaluated later via extensive simulations.

Though the focus of this project is to evaluate diagnostic accuracy, it remains crucial to select a proper regression model to distinguish all survival outcomes and identify covariate effects on each outcome at different time points. In this work, we adopted three models, the proportional hazards regression, Fine-Gray's model (Fine and Gray, 1999) and the multinomial logistic risk regression model (Gerds et al., 2012). Three simulation designs were considered in Section 4.3. For each of these three models, two data designs were examined, with and without added covariate improving diagnostic accuracy. In Section 4.4, we applied both NRI and IDI estimators to the MACS data for assessing whether including a new biomarker, CD4 cell count, would improve predictive ability over the old model. Some discussions are given in Section 4.5.

4.2 METHODS

4.2.1 Notation

In a competing-risk setting, there are two or more types of events. To simplify the notation, only two types are considered here, which are denoted as $\epsilon = 1, 2$, though the proposed methods can be naturally extended to more than two competing events. Let T be the time to first event from either type. With two competing events, we can define three categories according to their disease status at a fixed time point t_0. For the i-th subject, if $T_i \leq t_0$ and $\epsilon_i = 1$, the subject belongs to the first category; if $T_i \leq t_0$ and $\epsilon_i = 2$, the subject belongs to the second category; otherwise the subject is in the third category of being "healthy." In practice there is often independent censoring C. Hence $X = \min(T; C)$ and the combined cause indicator $\eta = I(T \leq C)\epsilon$ are observed. Let z_1, a p-dimension vector, denote conventional predictors and let z_2, a q-dimension vector, denote new biomarkers. The data consist of $\{X_i, \eta_i, \boldsymbol{z}_{i1}, \boldsymbol{z}_{i2} | i = 1, ..., n\}$. In the sequel we denote the "old"

model with conventional markers as $\mathscr{M}_1$ and the "new" model with both conventional and new markers as $\mathscr{M}_2$.

An extension of the NRI in Li et al. (2013b) for a K-level categorical outcome D is:

$$S = \sum_{k=1}^{K} \omega_k P\left\{\hat{p}_k(\mathscr{M}_2) = \max \hat{\boldsymbol{p}}(\mathscr{M}_2), \hat{p}_k(\mathscr{M}_1) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_1) | D = k\right\}$$
$$- \sum_{k=1}^{K} \omega_k P\left\{\hat{p}_k(\mathscr{M}_2) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_2), \hat{p}_k(\mathscr{M}_1) = \max \hat{\boldsymbol{p}}(\mathscr{M}_1) | D = k\right\},$$

where ω_k is a weight function for the k-th category of the outcome, and $\sum_k \omega_k = 1$, and $\hat{p}_k(\mathscr{M}_m)$ is the estimated probability of the outcome from the k-th category based on the model $\mathscr{M}_m$ for $m = 1,2$. $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3)$. When there are only two categories $K = 2$, and the weights are $\omega_k = 1/2$ for $k = 1,2$, then the S is equivalent to the NRI given in Pencina et al. (2008). Li et al. (2013b) also proposed an extension of the IDI based on the relationship between the IDI and the increase in the coefficient of determination R^2 from the "old" multinomial logistic model to the "new" one with additional markers. That is, $R = \sum_{k=1}^{K} \omega_k \left\{R_k^2(\mathscr{M}_2) - R_k^2(\mathscr{M}_1)\right\}$, where ω_k is again a weight function for the k-th category of the outcome, and $R_k^2(\mathscr{M}_m)$ is the coefficient of determination from $\mathscr{M}_m$, $m = 1,2$. Again when $K = 2$ and $\omega_k = 1/2$, the multi-category IDI reduces to the original IDI in Pencina et al. (2008).

4.2.2 Net Reclassification Improvement for Competing Outcomes

Without loss of generality, we consider competing outcomes with three categories. For model $\mathscr{M}_m, m = 1,2$, define $p_k(\mathscr{M}_m, t_0) = P(T \leq t_0, \epsilon = k | \mathscr{M}_m)$, for $k = 1,2$, and $p_3(\mathscr{M}_m, t_0) = P(T > t_0 | \mathscr{M}_m)$. Three categories are defined in Section 4.2.1. A well-calibrated regression model such as the multi-state Cheng et al. (1998), Fine and Gray (1999), and Gerds et al. (2012) models can be used. For each subject i, we obtain the estimators $\hat{p}_{1i}(\mathscr{M}_m, t_0)$, $\hat{p}_{2i}(\mathscr{M}_m, t_0)$, $\hat{p}_{3i}(\mathscr{M}_m, t_0)$, and $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, \hat{p}_3)$. The NRI for multi-category outcomes can thus be extended to the competing-risk setting at any $t_0 > 0$:

$$\begin{aligned} S(t_0) = &\sum_{k=1}^{K} \omega_k P\left\{\hat{p}_k(\mathscr{M}_2, t_0) = \max \hat{\boldsymbol{p}}(\mathscr{M}_2, t_0), \hat{p}_k(\mathscr{M}_1, t_0) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_1, t_0) | D = k\right\} \\ &- \sum_{k=1}^{K} \omega_k P\left\{\hat{p}_k(\mathscr{M}_2, t_0) \neq \max \hat{\boldsymbol{p}}(\mathscr{M}_2, t_0), \hat{p}_k(\mathscr{M}_1, t_0) = \max \hat{\boldsymbol{p}}(\mathscr{M}_1, t_0) | D = k\right\}. \end{aligned} \tag{4.1}$$

One complication in estimating $S(t_0)$ with censored competing risks data is that not every subject status is available. For example, some subjects may have been censored before t_0, and hence their disease status cannot be determined. Therefore, those subjects whose disease status can be decided based on the observed pair $(X_i;\eta_i)$ should be properly weighted to account for those subjects with "missing" disease status due to censoring. Thus, we propose the following estimator of the NRI at any time point t_0 as:

$$\hat{S}(t_0) = \sum_{k=1,2}\left[\omega_k \frac{\sum_{i=1}^n (h^+_{i,k}(t_0) - h^-_{i,k}(t_0))}{\sum_{i=1}^n I\left\{X_i \le t_0, \eta_i = k\right\}/\hat{G}(X_i-)}\right] + \omega_3 \frac{\sum_{i=1}^n (h^+_{i,3}(t_0) - h^-_{i,3}(t_0))}{\sum_{i=1}^n I\left\{X_i > t_0\right\}/\hat{G}(t_0)},$$

$$h^+_{i,k}(t_0) = I\left\{\hat{p}_{1i}(\mathscr{M}_2,t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_2,t_0), \hat{p}_{1i}(\mathscr{M}_1,t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_1,t_0), X_i \le t_0, \eta_i = k\right\}/\hat{G}(X_i-),$$

$$h^-_{i,k}(t_0) = I\left\{\hat{p}_{1i}(\mathscr{M}_2,t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_2,t_0), \hat{p}_{1i}(\mathscr{M}_1,t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_1,t_0), X_i \le t_0, \eta_i = k\right\}/\hat{G}(X_i-),$$

$$h^+_{i,3}(t_0) = I\left\{\hat{p}_{3i}(\mathscr{M}_2,t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_2,t_0), \hat{p}_{3i}(\mathscr{M}_1,t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_1,t_0), X_i > t_0\right\}/\hat{G}(t_0),$$

$$h^-_{i,3}(t_0) = I\left\{\hat{p}_{3i}(\mathscr{M}_2,t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_2,t_0), \hat{p}_{3i}(\mathscr{M}_1,t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_1,t_0), X_i > t_0\right\}/\hat{G}(t_0),$$

where $\omega_k, k = 1,2,3$, are weight functions for the three disease categories and can be simply set to be $1/3$ if there is no prior on the categories, and $\hat{G}$ is the Kaplan-Meier estimator of the censoring survival function. For each category $k = 1,2,3$, $h^+_{i,k}(t_0)$ is an indicator function whether the "old" model $\mathscr{M}_1$ makes a wrong prediction on Category k for the i-th subject while the "new" $\mathscr{M}_2$ correctly identifies it. Conversely, $h^-_{i,k}(t_0)$ indicates whether the "new" model changes a right prediction from the "old" model.

The consistency and asymptotic normality of $\hat{S}(t_0)$ are given in Appendix A, and variance estimate can be obtained from the influence function provided by equation A.1. Estimated variance is inversely related with $\sqrt{N}$, where N is sample size.

4.2.3 Integrated Discrimination Improvement for Competing Outcomes

We first define the time-dependent IDI for competing risks outcomes by adapting its definition for multi-category outcomes proposed by Li et al. (2013b). The IDI for multi-category outcomes is defined to be a weighted sum of variability explained, which is increase of coefficient of determination $R^2_k(\mathscr{M}_m)$, $k = 1,2,3...$, by the "new" model, $m = 2$, over the "old" model, $m = 1$. $R^2_k(\mathscr{M}_m)$ is

closely connected to the probabilities of each category, so we extend the IDI for competing risks outcomes at time t_0 as:

$$R(t_0) = \sum_{k=1}^{K} \omega_k \left\{ R_k^2(\mathscr{M}_2, t_0) - R_k^2(\mathscr{M}_1, t_0) \right\}, \tag{4.2}$$

where ω_k are again some weight functions. The estimation of the IDI at time t_0 involves the evaluation of $R_k^2(\mathscr{M}_m, t_0)$, which is the proportion of variability in the k-th category that is explained by model $\mathscr{M}_m$, for $m = 1,2$ and $k = 1,2$. Without any covariates, we estimate the probability of falling into the k-th category by $\hat{\pi}_k(t_0)$, where $\hat{\pi}_k(t_0) = \hat{n}_k(t_0)/(\hat{n}_1(t_0) + \hat{n}_2(t_0) + \hat{n}_3(t_0))$, with $\hat{n}_k(t_0) = \sum_{i=1}^{n} I\left\{X_i \le t_0, \eta_i = k\right\} / \hat{G}(X_i-), k = 1,2$ and $\hat{n}_3(t_0) = \sum_{i=1}^{n} I\left\{X_i > t_0\right\} / \hat{G}(t_0)$. Hence the variance without any model is $\hat{\pi}_k(t_0)(1-\hat{\pi}_k(t_0))$. With model $\mathscr{M}_m, m = 1,2$, the variance can be estimated by $\frac{1}{n}\sum_{i=1}^{n}\left\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\right\}^2$, where $\overline{\hat{p}_k(\mathscr{M}_m, t_0)} = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_{ki}(\mathscr{M}_m, t_0)$. Therefore, we propose the following estimator of the IDI at time t_0:

$$\hat{R}(t_0) = \sum_{k=1}^{3} \frac{\omega_k}{n\hat{\pi}_k(t_0)\{1-\hat{\pi}_k(t_0)\}} \sum_{i=1}^{n} \left[\left\{\hat{p}_{ki}(\mathscr{M}_2, t_0) - \overline{\hat{p}_k(\mathscr{M}_2, t_0)}\right\}^2 - \left\{\hat{p}_{ki}(\mathscr{M}_1, t_0) - \overline{\hat{p}_k(\mathscr{M}_1, t_0)}\right\}^2 \right].$$

Appendix B shows the asymptotic property of $\hat{R}(t_0)$. Based on equation B.1, we can estimate the influence function $IF^{**}(X_i, \eta_i, M_i, t_0)$ to compute its variance estimate using the sample. However, the IDI estimator relies on the estimated probabilities from a particular competing-risk model, and asymptotic variance will change if another model is used. Some competing-risk models have well-defined influence functions $IF_{\hat{p}_{ki}}$, while others do not have explicit forms. As a result, it is difficult to obtain an explicit form of variance estimation for the IDI with various competing-risk models of choice. Bootstrap procedure provides an alternative way for inference purposes. Confidence interval can be constructed from bootstrap standard error assuming asymptotic normality or by selecting percentiles from bootstrapped sample, but the skewness and bias of bootstrap distribution may lead to misleading results. Thus, we propose to use a bias-corrected and accelerated (BCa) bootstrap procedure to obtain confidence intervals for the IDI, which correct the skewness and the bias of the bootstrap distribution (Efron, 1987; Efron and Tibshirani, 1993). Shi et al. (2014b) has shown BCa bootstrap performs better than the former two. By conducting BCa bootstrap, we first bootstrap the original sample to obtain difference between median of estimators from bootstrapped sample and original estimator from original sample. Then we use jackknife approach to calculate the acceleration factor for measuring skewness. By the end, percentiles are calculated

based on the bias-correction factor and acceleration parameter, so we can obtain the confidence intervals of the IDI estimator $\hat{R}(t_0)$ (Efron and Tibshirani, 1993).

As suggested by Blanche et al. (2013), independent censoring assumption is restrictive in practice, while assuming conditional independence of censoring given biomarkers may make the extended NRI and IDI more general, since risk of censoring can be correlated with biomarkers. Consequently, the Kaplan-Meier estimator $\hat{G}$ used for inverse probability of censoring weighting in equations (4.2.2) and (4.2.3) can be replaced with a conditional survival estimator (Blanche et al., 2013). However, asymptotic theory will become much more complex and BCa bootstrap can be used for inferential purposes.

There are certain limitations of the proposed NRI and IDI. The purpose is to evaluate the effect of new biomarkers on diagnostic accuracy rather than the competing-risk model itself. As a result, model diagnostics are important before applying the NRI and IDI. Pencina et al. (2011) suggested cross-validation (CV) can be applied to account for over-optimism of the model. The probabilities of each outcome is computed from cross-validated sample, which can then be used to calculate the NRI and IDI. In the simulation studies, we are going to examine how cross-validation impacts these two measures. Due to the complexity of asymptotic theory with CV, we propose to use BCa bootstrap for obtaining confidence intervals for proposed estimators.

4.3 SIMULATION STUDIES

In practice, we usually do not know the "right" model, and there is a chance that we could pick a reasonable yet incorrect model for our data. Thus, we need to evaluate the impact of model choices on the performance of accuracy improvement evaluation with new biomarkers included. Here, we first designed three different sets of data with respect to three popular competing-risk models, including multi-state, Fine and Gray, and multinomial logistic models, to examine the proposed estimators for the extended NRI and IDI in competing risks settings. Three covariates were used in all three designs, where Z_1 and Z_3 were generated from standard normal distribution, truncated at ± 3.5 to prevent extreme values, and Z_2 was generated from a Bernoulli (0.7) distribution. The three cases of data were simulated as follows:

Case 1. We simulated the event time from a Weibull model with three covariates,

$$\log(T) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \sigma W,$$

where W was generated from the standard extreme value distribution. This error distribution gives the proportional hazard interpretations for all covariates. We set $\beta_0 = 2.5$, $\beta_1 = 0.05$, $\beta_2 = -0.05$, $\beta_3 = 0.15$ and $\sigma = 0.2$. Since the coefficient for the new marker β_3 is three times the size of the coefficients β_1 and β_2 for two conventional predictors, we expect that the "new" model including Z_1, Z_2 and Z_3 would have improved predictive ability over the "old" model that only uses Z_1 and Z_2. The cause indicators, $k = 1, 2$, were generated with equal probability. The censoring time was simulated from a uniform $[2, 31]$ distribution, resulting in about 30% censoring, and from a uniform $[1, 21]$ with 50% censoring.

Case 2. We used a simulation design similar to the one proposed by Fine and Gray (1999) in this case. The subdistribution for cause 1 is defined by

$$F_1(t|\mathbf{Z}) = 1 - \left[1 - p\left\{1 - \exp\left(-(t/20)^5\right)\right\}\right]^{\exp(\beta_{11}Z_1 + \beta_{12}Z_2 + \beta_{13}Z_3)},$$

with a mass of $1 - p$ when t is at ∞ and all covariates are zeros. When a uniform random number exceeds $F_1(\infty|\mathbf{Z})$, subjects are assumed to experience the cause 2 event with the conditional probability

$$P(T \le t|\epsilon = 2, \mathbf{Z}) = 1 - \exp\left(-\exp(\beta_{21}Z_1 + \beta_{22}Z_2 + \beta_{23}Z_3)(t/20)^5\right).$$

We set $\beta_{11} = 0.2$, $\beta_{12} = -0.5$, $\beta_{13} = 1$, $\beta_{21} = 0.02$, $\beta_{22} = -0.05$, $\beta_{23} = 0.1$, and $p = 0.65$. Including Z_3 in the model, besides Z_1 and Z_2, is expected to improve prediction over the one not including Z_3. The censoring distribution follows uniform $[10, 37.5]$ and uniform $[7, 29.5]$ with 30% and 50% censoring.

Case 3. We considered a multinomial logistic regression model as suggested by Gerds et al. (2012) in this case. Define $F_k(t, \boldsymbol{Z}) = P(T \le t, \epsilon = k|\boldsymbol{Z})$, $k = 1, 2$. For cause k, logistic-transformed probabilities were set as

$$\log\{F_k(t)/(1 - F_1(t) - F_2(t))\} = \mu(t) + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3, \quad t > 0,$$

where $\mu(t)$ was set to be $t-11$, $\beta_1 = 0.5$, $\beta_2 = 0.5$ and $\beta_3 = 1$. Since β_3 is twice the size of β_1 and β_2, we expect the new model including Z_3 would have a better predictive ability than the old one using only Z_1 and Z_2. The event time was simulated by inverting the survival probability, and cause indicators were assigned with equal probabilities. Independent censoring time was simulated from a uniform $[0,32]$ distribution for 30% censoring and from a uniform $[0,29.2]$ distribution for 50% censoring.

In light of Demler et al. (2017), we also want to examine the robustness of the inferential procedures for the NRI and the IDI under the null, where adding the new biomarker into "old" model doesn't improve the predictive ability. Thus, we consider the following three scenarios:

Case 4. Similar to *Case 1*, we set $\beta_0 = 2.5$, $\beta_1 = 0.25$, $\beta_2 = -0.05$, $\beta_3 = 0$ and $\sigma = 0.2$. Censoring time was simulated from uniform [3, 32] for 30% censoring, and from uniform [1, 21] for 50% censoring.

Case 5. Similar to *Case 2*, we set $\beta_{11} = 0.2$, $\beta_{12} = -0.5$, $\beta_{13} = 0$, $\beta_{21} = 0.02$, $\beta_{22} = -0.05$, $\beta_{23} = 0$, and $p = 0.65$. The censoring distributions followed a uniform $[10,39]$ distribution for the 30% censoring case and from uniform $[7,30]$ for 50% censoring.

Case 6. Similar to *Case 3*, we set $\beta_1 = 3$, $\beta_2 = 1$ and $\beta_3 = 0$. Independent censoring time was simulated from uniform [0, 32] for 30% censoring and from uniform [0, 29.2] for 50% censoring.

Figure 12 shows the survival curves of each case. For *Cases 1, 2* and *3*, true NRI and IDI are difficult to obtain because probabilities depend on covariate values. 1,000 samples of size 1,000 without censoring is used to simulate true values. Under *Cases 4, 5* and *6*, we expect that the predictive ability of the "new" model would not be improved. Thus the true NRI and IDI are zero. For each simulation case, we generated 1,000 samples of size 400, and applied all three models without CV, Cox's proportional hazard model, Gerds' multinomial logistic risk regression, and Fine-Gray's subdistribution hazard model. Probabilities $\hat{p}$ for each cause $k = 1,2$ and survival $k = 3$ at chosen time points were obtained with each model and then used for the NRI and the IDI calculation. Cox regression and Fine-Gray's model estimate the CIF for each cause separately, while Gerds' model estimates CIFs for both causes simultaneously. We built confidence intervals (CIs) for the NRI based on (A.1) and and compare it with the CIs using biased-corrected accelerated (BCa) bootstrapping. For the IDI, we also calculated the CIs using BCa bootstrapping. Simulation was run in `R`, where packages `survival`, `lme4` and `cmprsk` were used for competing risks model-

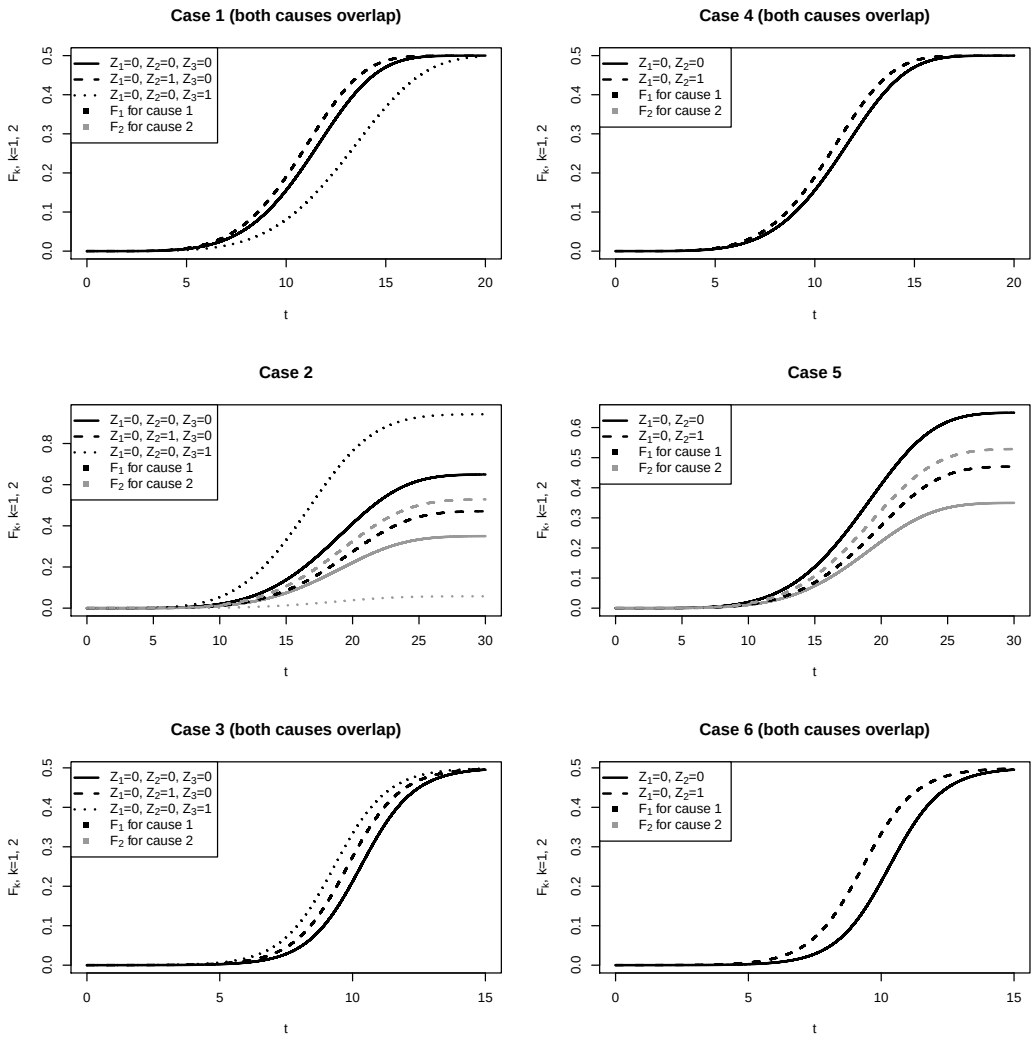

Figure 12: Probability of each cause for all six cases in simulation study

ing.The simulation results for the NRI and the IDI from cases 1, 2 and 3 under 30% censoring, in which model predictive ability should improve with the "new" marker, are shown in Table 3 and 4, respectively. Table 6 and 7 summarize the simulation results for the NRI and the IDI under cases 4, 5 and 6 with 30% censoring, when the added covariate does not improve prediction accuracy.

For the Tables 3, 4, 6 and 7, results from correct models are given in bold. Under alternative hypothesis when the added covariate improves predictability, true means S and R were calculated using 1,000 samples with size 1,000. Under null, true means S and R are 0. 1,000 samples with size 200 or 400 each was used to calculate the sample means $\hat{S}$, $\hat{R}$ and empirical standard error SE. Mean of estimated standard deviations $\text{SD}_{\hat{S}}$ was outputted by NRI formula provided. Coverage rate $\text{CR}_{\hat{S}}$=(count of true NRI entering the intervals $[\hat{S}-1.96\text{SD}_{\hat{S}}, \hat{S}+1.96\text{SD}_{\hat{S}}]$)/1,000. Each sample was bootstrapped 1,000 times, and the mean of 1,000 bootstrap standard deviations is denoted as BSD. Coverage rate BCR=(count of true value entering the 95% BCa bootstrap intervals)/1,000.

From Table 3 and 4, we first notice that, for both NRI and IDI, estimated $\hat{S}$ and $\hat{R}$ on average are very close to true values S and R, with the correct model for a specific data design. The average standard deviations of the estimated NRIs based on formula (A.1) approximate the empirical standard errors closely. The 95% CIs based on asymptotic normality and estimated standard deviation cover the true values about 95% times, though the coverage rates are a bit lower than 95% at some time. One possible reason is the use of approximation from Taylor's expansion, and our formula-based asymptotic variance could underestimate the true variance of the proposed NRI estimators in this situation. Nevertheless, when models are specified correctly, the results are very good in general. Similar to the NRI, IDI estimators are close to their true values when models are correctively specified, average bootstrap standard deviations are comparable to empirical standard errors, and coverage rates are around 95% using BCa bootstrap CIs. As the censoring rate increases from 30% to 50% and sample size decreases from 400 to 200 (results shown in Tables S1, S3 and S5 in the supplementary material), standard errors of both NRI and IDI estimators increase but similar coverage rates are observed.

However, despite the appealing interpretation of covariate effects on cumulative incidence functions, Fine and Gray's model does not guarantee the sum of all cause probabilities is equal to one. So, proposed standard deviation estimation for the NRI often underestimates. The underestimation is worse for the IDI estimators when Fine and Gray is mis-used in predicting event

probabilities. Table 5 presents the results from the IDI and incremental AUC (IAUC) that were proposed by Shi et al. (2014b) under the same settings as in Table 3 and 4. The methods from Shi and others (2014) lump the competing events with healthy controls together, and can lead to the wrong conclusion of no effect of the added covariate on competing outcomes as shown in Case 2, despite the fact that the new marker is clearly related to competing outcomes.

Similar conclusions are observed from Table 6 and 7. Even though the true underlying data are from the null and both NRI and IDI are degenerate, the probabilities of covering zero are high for both the NRI's formula-based CIs and the IDI's BCa bootstrap CIs. Demler et al. (2017) suggested to "un-nest" the models by including independent weak predictors in both models such that they are no longer nested. As a result, we added independent and non-informative noises from the standard normal distribution as additional covariates into both models. The coverage rates are improved for the NRI estimation by un-nesting the models, except for Fine and Gray's model for its design that the sum of all cause probabilities is not equal to one. However, by un-nesting the models, bias would be introduced into the IDI estimation, which might lead to lower coverage rates of CIs. Thus, we chose to simply use the original BCa bootstrapping procedure instead. Results for the null hypothesis under 50% censoring and with 200 sample size are summarized in Tables 9-20 from Appendix C. The same patterns are observed.

Results from cross-validation when models are corrected specified are shown in Tables 21 and 22 from Appendix C. Under the alternative, the true NRI values, except for Fine and Gray's model with improper probabilistic design, are smaller than the ones without cross-validation, which is consistent to Pencina et al. (2011). The coverage rates of cross-validated NRI with BCa bootstrap are relatively low.So, a better bootstrap procedure or explicit asymptotic theory can be further explored. The performance of BCa bootstrap of the IDI with CV is satisfactory, except for Gerds model, where B-spline technique is used for model approximation and later time points suffer from more censoring. As for Cox regression and Fine and Gray's model, the IDI is close to the one without CV. This means probabilities estimated from CV did not change much, while the small change made huge difference for outcomes to be recategorized.

Table 3: Simulation details for the NRI under alternative (30% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 1*)	S	**.114**	**.100**	**.126**	.114	.100	.126	.114	.100	.126
	$\hat{S}$	**.112**	**.105**	**.125**	.068	.105	.100	.108	.104	.121
	$SE_{\hat{S}}$	**.027**	**.030**	**.031**	.027	.029	.033	.025	.030	.031
	$SD_{\hat{S}}$	**.024**	**.030**	**.030**	.016	.028	.026	.024	.030	.030
	$CR_{\hat{S}}$	**.935**	**.944**	**.943**	.346	.926	.754	.930	.945	.937
	$BCR_{\hat{S}}$	**.918**	**.898**	**.899**	.621	.869	.855	.910	.896	.904
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 2*)	S	.109	.127	.123	**.109**	**.127**	**.123**	.109	.127	.123
	$\hat{S}$	.114	.131	.127	**.115**	**.128**	**.126**	.121	.132	.126
	$SE_{\hat{S}}$	.035	.033	.028	**.038**	**.035**	**.030**	.035	.033	.030
	$SD_{\hat{S}}$	.032	.029	.025	**.033**	**.032**	**.026**	.032	.028	.025
	$CR_{\hat{S}}$	.919	.911	.902	**.911**	**.918**	**.909**	.907	.886	.873
	$BCR_{\hat{S}}$	.908	.900	.905	**.910**	**.898**	**.914**	.882	.914	.912
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 3*)	S	.209	.189	.169	.209	.189	.169	**.209**	**.189**	**.169**
	$\hat{S}$	.208	.193	.176	.065	.143	.165	**.205**	**.186**	**.172**
	$SE_{\hat{S}}$	.027	.031	.031	.029	.028	.029	**.028**	**.031**	**.031**
	$SD_{\hat{S}}$	.025	.027	.030	.015	.020	.027	**.026**	**.027**	**.030**
	$CR_{\hat{S}}$	.924	.897	.933	0	.435	.933	**.929**	**.920**	**.937**
	$BCR_{\hat{S}}$	.914	.918	.897	.019	.757	.894	**.912**	**.898**	**.889**

Table 4: Simulation details for the IDI under alternative (30% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 1*)	R	**.109**	**.106**	**.099**	.109	.106	.099	.109	.106	.099
	$\hat{R}$	**.110**	**.108**	**.100**	.018	.025	.034	.101	.092	.079
	$\text{SE}_{\hat{R}}$	**.016**	**.016**	**.015**	.005	.006	.008	.017	.016	.016
	$\text{BSD}_{\hat{R}}$	**.017**	**.016**	**.016**	.007	.008	.010	.018	.017	.016
	$\text{BCR}_{\hat{R}}$	**.953**	**.952**	**.946**	0	0	.008	.935	.873	.810
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 2*)	R	.151	.158	.165	**.151**	**.158**	**.165**	.151	.158	.165
	$\hat{R}$	.128	.137	.144	**.148**	**.155**	**.162**	.035	.045	.055
	$\text{SE}_{\hat{R}}$	.020	.021	.022	**.020**	**.021**	**.021**	.014	.015	.016
	$\text{BSD}_{\hat{R}}$	.021	.022	.023	**.021**	**.021**	**.022**	.015	.017	.017
	$\text{BCR}_{\hat{R}}$	.661	.730	.781	**.946**	**.946**	**.950**	.012	.002	.002
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 3*)	R	.288	.271	.251	.288	.271	.251	**.288**	**.271**	**.251**
	$\hat{R}$	.266	.251	.234	.034	.045	.058	**.289**	**.272**	**.252**
	$\text{SE}_{\hat{R}}$	.026	.024	.023	.007	.008	.010	**.020**	**.018**	**.018**
	$\text{BSD}_{\hat{R}}$	.027	.024	.023	.010	.011	.012	**.023**	**.020**	**.019**
	$\text{BCR}_{\hat{R}}$	.512	.606	.732	0	0	0	**.947**	**.958**	**.949**

Table 5: Simulation details for the IDI and IAUC from Shi et al. (2014b) when the added covariate improves predictability (30% censoring). Results for each case were obtained with correct models specified. 1,000 samples with size 400 each was used to calculate the empirical standard error *SE* and sample means $\widehat{\text{IDI}}$ and $\widehat{\text{IAUC}}$.

		Weibull (*Case 1*)			Fine Gray (*Case 2*)			Gerds (*Case 3*)		
	time	11	12	13	20	21	22	9	10	11
Cause 1	$\widehat{\text{IDI}}$	.097	.099	.096	-.00007	-.00003	-.00008	.267	.253	.236
	$\text{SE}_{\widehat{\text{IDI}}}$	.032	.033	.034	.002	.001	.002	.041	.039	.038
	$\widehat{\text{IAUC}}$	.359	.382	.411	.0007	-.0002	.0005	.688	.679	.674
	$\text{SE}_{\widehat{\text{IAUC}}}$	.080	.084	.099	.017	.012	.015	.053	.058	.067
	time	11	12	13	20	21	22	9	10	11
Cause 2	$\widehat{\text{IDI}}$	.098	.101	.097	-.00006	-.000003	-.000003	.269	.253	.235
	$\text{SE}_{\widehat{\text{IDI}}}$	.031	.032	.034	.001	.001	.002	.042	.041	.041
	$\widehat{\text{IAUC}}$	.362	.387	.413	-.0001	.0004	-.001	.690	.682	.671
	$\text{SE}_{\widehat{\text{IAUC}}}$	.077	.085	.098	.009	.011	.016	.053	.059	.069

Table 6: Simulation details for the NRI under null (30% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 4*)	$\hat{S}$	**.005**	**.005**	**.005**	.005	.004	.003	.004	.004	.004
	$\text{SE}_{\hat{S}}$	**.015**	**.015**	**.015**	.013	.015	.015	.015	.015	.016
	$\text{SD}_{\hat{S}}$	**.013**	**.014**	**.014**	.010	.013	.015	.014	.014	.015
	$\text{CR}_{\hat{S}}$	**.915**	**.919**	**.939**	.845	.918	.940	.931	.949	.935
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	**.949**	**.941**	**.945**	.913	.933	.957	.940	.927	.933
	$\text{BCR}_{\hat{S}}$	**.855**	**.845**	**.834**	.835	.860	.868	.842	.817	.859
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 5*)	$\hat{S}$	.005	.004	.002	**.005**	**.004**	**.002**	.006	.004	.002
	$\text{SE}_{\hat{S}}$	.016	.015	.012	**.015**	**.015**	**.011**	.016	.016	.011
	$\text{SD}_{\hat{S}}$	.014	.014	.010	**.013**	**.013**	**.010**	.014	.014	.010
	$\text{CR}_{\hat{S}}$	.915	.934	.891	**.913**	**.913**	**.897**	.916	.919	.918
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.939	.946	.951	**.946**	**.955**	**.952**	.937	.942	.959
	$\text{BCR}_{\hat{S}}$	.902	.908	.882	**.903**	**.917**	**.898**	.889	.872	.901
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 6*)	$\hat{S}$	.005	.005	.005	.004	.007	.005	**.005**	**.006**	**.006**
	$\text{SE}_{\hat{S}}$	.016	.016	.016	.010	.015	.016	**.018**	**.017**	**.016**
	$\text{SD}_{\hat{S}}$	.015	.015	.015	.006	.011	.014	**.015**	**.015**	**.015**
	$\text{CR}_{\hat{S}}$	.915	.922	.933	.709	.856	.907	**.769**	**.783**	**.786**
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.938	.945	.946	.856	.889	.935	**.945**	**.947**	**.945**
	$\text{BCR}_{\hat{S}}$	.818	.836	.827	.732	.831	.868	**.819**	**.836**	**.857**

Table 7: Simulation details for the IDI under null (30% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 4*)	$\hat{R}$	**.002**	**.002**	**.002**	.002	.002	.002	.002	.002	.002
	$\text{SE}_{\hat{R}}$	**.002**	**.003**	**.003**	.003	.003	.003	.002	.002	.002
	$\text{BSD}_{\hat{R}}$	**.004**	**.004**	**.005**	.004	.005	.005	.004	.004	.004
	$\text{BCR}_{\hat{R}}$	**.962**	**.947**	**.953**	.897	.903	.909	.955	.957	.955
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 5*)	$\hat{R}$	.002	.003	.003	**.002**	**.002**	**.002**	.0004	.0007	.0009
	$\text{SE}_{\hat{R}}$	.003	.003	.003	**.002**	**.002**	**.003**	.0006	.0008	.001
	$\text{BSD}_{\hat{R}}$	.004	.005	.005	**.004**	**.004**	**.004**	.001	.001	.002
	$\text{BCR}_{\hat{R}}$	.864	.857	.842	**.776**	**.773**	**.755**	.865	.872	.870
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 6*)	$\hat{R}$	.003	.003	.003	.002	.002	.002	**.002**	**.002**	**.002**
	$\text{SE}_{\hat{R}}$	.003	.003	.003	.003	.003	.004	**.002**	**.002**	**.003**
	$\text{BSD}_{\hat{R}}$	.005	.005	.005	.004	.005	.005	**.004**	**.004**	**.004**
	$\text{BCR}_{\hat{R}}$	.953	.927	.932	.910	.907	.905	**.960**	**.943**	**.935**

4.4 APPLICATION TO THE MULTICENTER AIDS COHORT STUDY

We applied the NRI and IDI methods to data obtained by the MACS. It is an ongoing study of homosexual and bisexual men at risk for or infected with HIV, recruited from four institutions in Baltimore, Chicago, Pittsburgh and Los Angeles (Kingsley et al., 1987; Kaslow et al., 1987). The data used for this analysis were gathered between April 2,1984 and April 8, 2017. Each participant underwent a clinical examination semi-annually, and neuropsychological testing approximately every two years (however, see Miller et al. (1990); Becker et al. (2014) for details) until they drop out of the study voluntarily or die. The current analysis utilizes the data from a substudy of the legacy effect of HIV on cognitive impairment, which contains 2,783 HIV seropositive men (Farinpour et al., 2003).

Individuals with HIV disease have historically been at risk for cognitive impairment. The MACS measured cognitive functions over time with a battery of neuropsychological (NP) tests which were summarized by T scores in six cognitive domains: working memory & attention, learning, motor speed & coordination, executive functioning, speed of information processing, and memory. We adopted the Multivariate Normative Comparison (MNC) method to define abnormality in cognition as in Huizenga et al. (2007) and Wang et al. (2019). Time to impairment was defined as the interval between study entry and the first visit where the six domain scores were deemed abnormal by the MNC method. Those subjects who were impaired at their first visit were excluded from the analyses. Though cognitive impairment and death could be thought of as semi-competing risks where death may censor impairment but not vice versa, we treated them as competing risks data by defining the events as cognitive impairment and death without impairment. If a subject died after the last complete NP visit and no cognitive impairment was detected, his time to impairment was competing-rick censored by death. Otherwise, subjects were censored at their last visit.

In the presence of competing-risk censoring, techniques such as Cox regression, Gerds' model, and Fine-Gray's model can be used to identify potential risk factors affecting cognition after the onset of HIV infection. However, these methods do not directly quantify the relative importance a factor is in predicting who might develop impairment, who might die, or who might be alive and disease free after a fixed time interval. Here we apply the NRI and the IDI treating CD4+

cell count as the "new" biomarker (with both linear and quadratic terms to account for nonlinearity when modeling cognitive impairment) to examine whether the inclusion of this variable will yield a better prediction. In the Legacy substudy (Popov et al., 2019), three other predictors – age, center for epidemiologic studies depression scale, and recruitment cohort (before or after 2001) – were found to be significantly related to cognitive impairment and were treated here as conventional predictors. All four predictors were measured at study entry. Final analysis included 1,972 seropositive subjects who had at least one visit with complete cognitive tests and the information on four predictors.

Within this subsample, 553 men were classified with cognitive impairment using the MNC method (28.0%), 597 died during follow-up without any cognitive impairment (30.3%), and 822 were censored by the "end" (at the data freeze) of the study (41.7%). Time to event or time in the study ranged from 5 months to 33 years. We examined the performance of CD4+ cell count and its quadratic transformation as the "new" biomarkers in predicting health status at 10 and 12 years since the start of the study with a proportional hazard model, Fine-Gray's model, and Gerds' model, using both NRI and IDI. The two events, cognitive impairment and death without cognitive impairment, were again modeled separately with Cox's model or Fine-Gray's model, and they were modeled simultaneously in the Gerds model. 5-fold cross-validation was used to compute the probabilities of both events and survival at selected time. Based on the predicted probabilities of both events $\hat{p}_1$ and $\hat{p}_2$ and predicted survival $\hat{p}_3$ that were calculated from the three models at 10 and 12 years, we computed the values of the NRI and IDI. For IDI, 10,000 bootstrap samples were used to produce 95% BCa bootstrap CIs. In order to select the most suitable regression model, we also computed cause-specific Brier scores (Schoop et al., 2011a) from in-sample probabilities, which also uses the method of inverse probability of censoring weighting on competing risk. The results are summarized in Table 8.

In Table 8 we can see that the estimated NRI and IDI and their 95% CIs are comparable across the three different models. Among the three models, Cox regression has the lowest Brier scores for both events at 10 and 12 years, suggesting that Cox regression is the most suitable competing risks model for our data. Moreover, from the Cox-Snell residual plots shown in Fig. 1, we can see that Cox regression provides a good overall fit to the MACS data, as the cumulative hazards of Cox-Snell residuals for both events go through a straight line with slope 1.

From the Cox regression model the estimated NRIs at 10 and 12 years since the start of the study are .037 and .084 with 95% CIs [.023, .052] and [.068, .101] respectively. The estimated IDIs are .049 and .060 with 95% BCa CIs [.039, .065] and [.048, .078]. Because the 95% CIs of both NRI and IDI do not include zero, we conclude that including the CD4+ cell counts in competing risks models increases the accuracy of predicting cognitive impairment and death after 10 and 12 years in the study. More specifically, the probabilities of correctly predicting health status (impairment, death, or neither) for a subject after 10 and 12 years of observation improves by 3.7% and 8.4%, by simply incorporating CD4+ cell counts with its quadratic transformation into the model. Also, the variability explained by the predictive model is increased by 4.9% and 6.0% for events at 10 and 12 years with the addition of the CD4+ counts.

However, some participants withdrew from the legacy study and died many years afterwards. If a subject died more than 4 years after his last NP visit, he may have experienced cognitive impairment between his last NP visit and death. As a sensitivity analysis, we censored such subjects four years after their last NP visit, assuming cognition stayed relatively stable over two consecutive NP visits (about 4 years as scheduled). In this way, 553 men were classified with cognitive impairment using the MNC method (28.0%), 425 died within 4 years after the last NP visit without any cognitive impairment (21.6%), and 994 were censored either at their last study visit or 4 years following the last NP exam, whichever was first (50.4%). Using the Cox regression model, the estimated NRIs at 10 and 12 years since the start of the study are .101 and .109 with 95% confidence intervals [.083, .119] and [.091, .127] respectively. The estimated IDIs are .095 and .100 with 95% BCa confidence intervals of [.084, .137] and [.086, .143]. Again, these findings suggest that including CD4+ cell counts in competing risks models can increase prediction accuracy of death and cognitive impairment after 10 and 12 years in the study.

4.5 DISCUSSION

We have demonstrated here the good practical performance of the extended NRI and IDI in competing risks settings. Although a CI for the IDI can be efficiently constructed based on the asymptotic linear representation for a well-studied regression model, the BCa bootstrap method serves as a

flexible alternative when a model is relatively new and its theoretical properties are less known. When the added variables have no effect on the events and models to be compared are nested, Demler et al. (2017) showed that the theory based on U-statistics fails. Still, the CIs for the NRI based on asymptotic normality and the BCa bootstrap CIs for the IDI seem to have satisfactory coverage as demonstrated by simulations. After un-nesting the models, the CI for the NRI is improved.

In this work we have considered three reasonable competing risks models. However, one can use any other semiparametric or parametric models such as Scheike et al. (2008) and Cheng (2009). The limitation of the extended NRI and IDI is that they are model dependent and are not robust against model mis-specification. As a result, it remains important to select a proper predictive model before examining diagnostic accuracy improvement over the course of variables' addition. Metrics, such as the Brier score, are useful in choosing the most appropriate model for the data.

Competing endpoints are common in biomedical research, although they are often neglected in analysis. The extended NRI and IDI for competing events provide alternative and straightforward interpretations of the importance of new biomarkers on top of conventional factors. They also serve as more unifying metrics than model coefficients such as hazards ratio or odds ratio, since the latter depend on the types and the scales of covariates. Moreover, this is in line with recent debate about moving away from statistical significance of 0.05 level (Wasserstein et al., 2019). Instead of simply looking at p values for the added variables in a regression model, one can assess the contribution of additional risk factors in prediction through interval estimates of the IDI and NRI. Thus, the extended NRI and IDI for multiple competing endpoints might be useful in screening and selecting covariates in high dimensional settings.

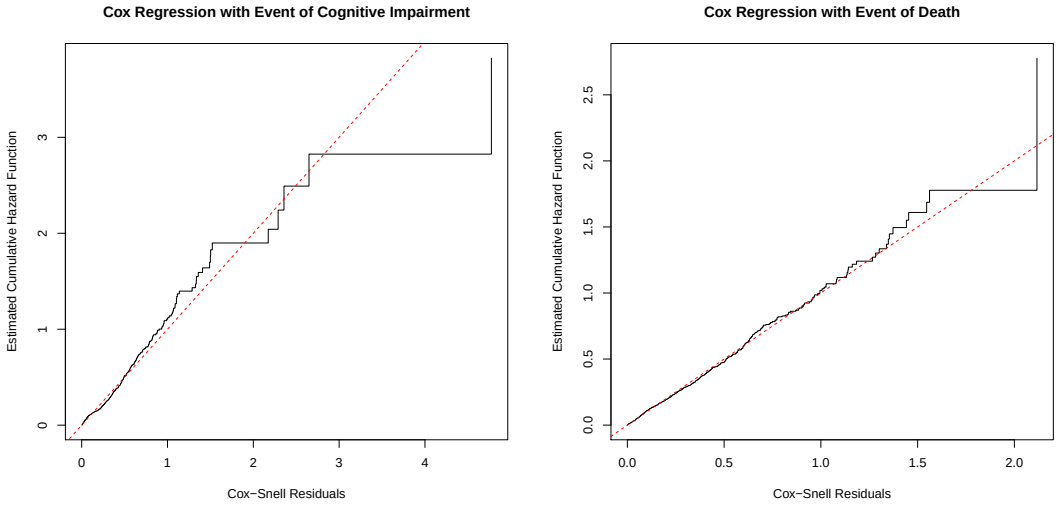

Figure 13: Cox-Snell residual plots for the MACS data with Cox regression

Table 8: NRI and IDI results for the MACS data at times 10 and 12 years. Competing risk censoring by death occurred when subjects died without cognitive impairment.

Model	Time	NRI	IDI	Brier Score with Event	
				Cognitive Impairment	Death
Cox Regression	t=10	.037 [.023, .052]	.049 [.039, .065]	.096	.132
	t=12	.084 [.068, .101]	.060 [.048, .078]	.107	.154
Gerds	t=10	.024 [.011, .038]	.050 [.040, .068]	.100	.166
	t=12	.070 [.054, .086]	.055 [.044, .074]	.113	.202
Fine and Gray	t=10	.024 [.014, .034]	.030 [.022, .041]	.096	.133
	t=12	.069 [.055, .084]	.038 [.028, .051]	.107	.156

APPENDIX A. THE ASYMPTOTIC THEORY OF NRI

We now establish the asymptotic normality of $\hat{S}(t_0)$, starting with introducing more notation.

$$\hat{S}(t_0) = \omega_1 \frac{\sum_{i=1}^n (h_{i,1}^+(t_0) - h_{i,1}^-(t_0))}{\sum_{i=1}^n I\left\{X_i \le t_0, \eta_i = 1\right\} / \hat{G}(X_i-)} + \omega_2 \frac{\sum_{i=1}^n (h_{i,2}^+(t_0) - h_{i,2}^-(t_0))}{\sum_{i=1}^n I\left\{X_i \le t_0, \eta_i = 2\right\} / \hat{G}(X_i-)} + \omega_3 \frac{\sum_{i=1}^n (h_{i,3}^+(t_0) - h_{i,3}^-(t_0))}{\sum_{i=1}^n I\left\{X_i > t_0\right\} / \hat{G}(t_0)}.$$

For $k = 1, 2$

$$h_{i,k}^+(t_0) = I\left\{\hat{p}_{1i}(\mathscr{M}_2, t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{1i}(\mathscr{M}_1, t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i \le t_0, \eta_i = k\right\} / \hat{G}(X_i-),$$

$$h_{i,k}^-(t_0) = I\left\{\hat{p}_{1i}(\mathscr{M}_2, t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{1i}(\mathscr{M}_1, t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i \le t_0, \eta_i = k\right\} / \hat{G}(X_i-),$$

$$h_{i,3}^+(t_0) = I\left\{\hat{p}_{3i}(\mathscr{M}_2, t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{3i}(\mathscr{M}_1, t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i > t_0\right\} / \hat{G}(t_0),$$

$$h_{i,3}^-(t_0) = I\left\{\hat{p}_{3i}(\mathscr{M}_2, t_0) \ne \max\hat{\mathbf{p}}(\mathscr{M}_2, t_0), \hat{p}_{3i}(\mathscr{M}_1, t_0) = \max\hat{\mathbf{p}}(\mathscr{M}_1, t_0), X_i > t_0\right\} / \hat{G}(t_0).$$

For each category $k = 1, 2, 3$, $h_{i,k}^+(t_0)$ is an indicator function of whether the "old" model $\mathscr{M}_1$ makes a wrong prediction on category k for i-th subject while the "new" $\mathscr{M}_2$ correctly identifies it. Conversely, $h_{i,k}^-(t_0)$ indicates whether the "new" model changes a right prediction from the "old" model. Let $Q_X(t) = P(X_i > t)$, and define the martingale of the censoring time C as $M_{C_i}(t) = I\{\eta_i = 0, X_i \le t\} - \int_0^t I\{X_i \ge u\} d\Lambda_C(u)$, where $\Lambda_C(\cdot)$ is the cumulative hazard function of C. For $k = 1, 2$ and the third "healthy" category, we define $f_{i,k}(t_0) = I\left\{X_i \le t_0, \eta_i = k\right\} / G(X_i-)$ and $f_{i,3}(t_0) = I\left\{X_i > t_0\right\} / G(t_0)$. For $k = 1, 2, 3$, define $h_k^{+/-}(t_0) = E h_{i,k}^{+/-}(t_0)$, $f_k(t_0) = E f_{i,k}(t_0)$, where the expectation is with respect to T, C, and covariates Z. Let $\hat{M}_C$ be the estimator defined by plugging in the usual Nelson-Aalen estimator of the cumulative hazard function of the censoring time C and let $\hat{h}_{i,k}^{+/-}$, $\hat{f}_{i,k}$ be defined by plugging in the Kaplan-Meier estimator $\hat{G}(\cdot)$, if applicable. Define

$$\hat{h}_k^{+/-}(t_0) = \frac{1}{n}\sum_{i=1}^n \hat{h}_{i,k}^{+/-}(t_0), \quad \hat{f}_k(t_0) = \frac{1}{n}\sum_{i=1}^n \hat{f}_{i,k}(t_0), \quad \hat{S}(t_0) = \sum_{k=1}^3 \omega_k \frac{\hat{h}_k^+(t_0) - \hat{h}_k^-(t_0)}{\hat{f}_k(t_0)}.$$

Since $G(t_0)$ in $h_{i,3}^{+/-}(t_0)$ and $f_{i,3}(t_0)$ will be canceled out, we redefine $h_{i,3}^{+/-}(t_0)$ and $f_{i,3}(t_0)$ by multiplying $G(t_0)$. Following Blanche et al. (2013), the Martingale representation of the Kaplan-Meier estimator of the censoring survival function (Hung and Chin-Tsang, 2010; Andersen et al., 1993) entails that:

$$\sup_t \left| \sqrt{n}(\hat{G}(t) - G(t)) + \frac{G(t)}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{dM_{C_i}(u)}{Q_X(u)} \right| = o_p(1).$$

By Taylor's expansion (Serfling, 1980),

$$\sup_{t_0} \left| \sqrt{n}\big(\hat{h}_k^+(t_0) - h_k^+(t_0)\big) - \left[\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} h_{i,k}^+(t_0) \left\{ 1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)} \right\} - h_k^+(t_0) \right] \right| = o_p(1),$$

for $k = 1, 2$. Similar results hold for $\hat{h}_k^-(t_0), k = 1, 2$. Moreover,

$$\sup_{t_0} \left| \sqrt{n}\big(\hat{f}_k(t_0) - f_k(t_0)\big) - \left[\frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} f_{i,k}(t_0) \left\{ 1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)} \right\} - F_k(t_0) \right] \right| = o_p(1).$$

When $k = 3$, again we have

$$\sup_{t_0} \left| \sqrt{n}\big(\hat{h}_3^+(t_0) - h_3^+(t_0)\big) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(h_{i,3}^+(t_0) - h_3^+(t_0)\big) \right| = 0,$$
$$\sup_{t_0} \left| \sqrt{n}\big(\hat{f}_3(t_0) - f_3(t_0)\big) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big(f_{i,3}(t_0) - f_3(t_0)\big) \right| = 0.$$

Then $\hat{S}(t_0)$ can be further formulated using Taylor's expansion:

$$\sup_t \left| \sqrt{n}\big(\hat{S}(t_0) - S(t_0)\big) - \frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \Psi_{ij}(t_0) \right| = o_p(1),$$

where

$$\begin{aligned}
\Psi_{ij}(t_0) &= \sum_{k=1}^{2} \omega_k \left\{ \frac{IF_{h_k^+(t_0)} - IF_{h_k^-(t_0)}}{f_k(t_0)} - \frac{h_k^+(t_0) - h_k^-(t_0)}{f_k^2(t_0)} IF_{f_k(t_0)} \right\} \\
&= \sum_{k=1}^{2} \omega_k \left\{ \left[h_{i,k}^+(t_0) - h_{i,k}^-(t_0) - \frac{f_{i,k}(t_0)}{f_k(t_0)} \big(h_k^+(t_0) - h_k^-(t_0)\big) \right] \left(1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)}\right) \right\} \Big/ f_k(t_0) \\
&\quad + \omega_3 \left\{ h_{i,3}^+(t_0) - h_{i,3}^-(t_0) - \frac{f_{i,3}(t_0)}{f_3(t_0)} \big(h_3^+(t_0) - h_3^-(t_0)\big) \right\} \Big/ f_3(t_0),
\end{aligned}$$

and IF denotes the influence function (Hampel et al., 1986) of each estimator respectively. By Hájek's projection principle, the following Hoeffding's decomposition (van der Vaart, 1998; Serfling, 1980) holds:

$$\frac{\sqrt{n}}{n(n-1)} \sum_{i \neq j}^{n} \Psi_{ij}(t_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(X_i, \eta_i, t_0) + o_p(1).$$

Given Martingale's properties (Kalbfleisch and Prentice, 2002), we also know that

$$E\Big[IF(X_i, \eta_i, t_0)\Big] = 0.$$

Let $\hat{M}_C$ be the estimator by plugging in the usual Nelson-Aalen estimator of the cumulative hazard function of the censoring time C. $\hat{h}_k^{+/-}$, $\hat{F}_k$, $\hat{h}_{i,k}^{+/-}$, and $\hat{f}_{i,k}$ were defined as above. $\hat{Q}_X$ is the estimate of Q using the empirical distribution of X. Plugging in these estimators to estimate Ψ_{ij}, we compute $IF(X_i, \eta_i, t_0)$ as

$$\hat{IF}(X_i, \eta_i, t_0) = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \big[\hat{\Psi}_{ij}(t_0) + \hat{\Psi}_{ji}(t_0)\big]. \tag{A.1}$$

APPENDIX B. THE ASYMPTOTIC THEORY OF IDI

The IDI at time t_0 is estimated as:

$$\hat{R}(t_0) = \sum_{k=1}^{3} \frac{\omega_k}{n\hat{\pi}_k(t_0)\{1-\hat{\pi}_k(t_0)\}} \sum_{i=1}^{n} \left[\left\{ \hat{p}_{ki}(\mathscr{M}_2, t_0) - \overline{\hat{p}_k(\mathscr{M}_2, t_0)} \right\}^2 - \left\{ \hat{p}_{ki}(\mathscr{M}_1, t_0) - \overline{\hat{p}_k(\mathscr{M}_1, t_0)} \right\}^2 \right].$$

Without any covariates, we estimate the probability of falling into the k-th category by $\hat{\pi}_k(t_0)$, where $\hat{\pi}_k(t_0) = \hat{n}_k(t_0)/(\hat{n}_1(t_0) + \hat{n}_2(t_0) + \hat{n}_3(t_0))$, with $\hat{n}_k(t_0) = \sum_{i=1}^{n} I\{X_i \le t_0, \eta_i = k\}/\hat{G}(X_i-), k = 1,2$ and $\hat{n}_3(t_0) = \sum_{i=1}^{n} I\{X_i > t_0\}/\hat{G}(t_0)$. Hence the variance without any model is $\hat{\pi}_k(t_0)(1-\hat{\pi}_k(t_0))$. With model $\mathscr{M}_m, m = 1,2$, the variance can be estimated by $\frac{1}{n}\sum_{i=1}^{n} \left\{ \hat{p}_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)} \right\}^2$, where $\overline{\hat{p}_k(\mathscr{M}_m, t_0)} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{ki}(\mathscr{M}_m, t_0)$.

Define $\pi_k = E\hat{n}_k/n$, for $k = 1,2,3$. Analogous to the arguments in Section (A), we have, for $k = 1,2$,

$$\sup_{t_0} \left| \sqrt{n}\Big(\hat{\pi}_k(t_0) - \pi_k(t_0)\Big) - \left\{ \frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} n_{i,k}(t_0)\Big(1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)}\Big) - \pi_k(t_0) \right\} \right| = o_p(1),$$

$$\sup_{t_0} \left| \sqrt{n}\Big(\hat{\pi}_3(t_0) - \pi_3(t_0)\Big) - \left\{ \frac{\sqrt{n}}{n(n-1)} \sum_{i=1}^{n}\sum_{j\neq i}^{n} n_{i,3}(t_0)\Big(1 + \int_0^{t_0} \frac{dM_{C_j}(u)}{Q_X(u)}\Big) - \pi_3(t_0) \right\} \right| = o_p(1).$$

Let

$$Q_{ij,k}(t_0) = n_{i,k}(t_0)\left(1 + \int_0^{X_i} \frac{dM_{C_j}(u)}{Q_X(u)}\right) - \pi_k(t_0), k = 1,2$$

$$Q_{ij,3}(t_0) = n_{i,3}(t_0)\left(1 + \int_0^{t_0} \frac{dM_{C_j}(u)}{Q_X(u)}\right) - \pi_3(t_0).$$

When covariates are involved, the variance that is explained by model $\mathscr{M}_m, m = 1,2$, is given as $\hat{D}_{k(m)} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\right\}^2$. $\hat{D}_{k(m)}$ can be rewritten as

$$\begin{aligned}\hat{D}_{k(m)} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0) + p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\right\}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\Bigg[\left\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0)\right\}^2 + \left\{p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\right\}^2 \\ &\quad + 2\left\{\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_{ki}(\mathscr{M}_m, t_0)\right\}\left\{p_{ki}(\mathscr{M}_m, t_0) - \overline{\hat{p}_k(\mathscr{M}_m, t_0)}\right\}\Bigg].\end{aligned}$$

Let $D_{k(m)} = E\hat{D}_{k(m)}$. By taking Taylor's expansion, it's easy to get the asymptotic linear representation for $\hat{D}_{k(m)}$. For $k = 1,2,3$:

$$\begin{aligned}\sup_{t_0}\Bigg|\sqrt{n}(\hat{D}_{k(m)} - D_{k(m)}) - \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\Bigg[&2\big(p_{ki}(\mathscr{M}_l, t_0) - p_k(\mathscr{M}_m, t_0)\big)IF_{\hat{p}_{ki}} \\ &+\big(p_{ki}(\mathscr{M}_m, t_0) - p_k(\mathscr{M}_m, t_0)\big)^2 - D_{k(m)}\Bigg]\Bigg| = o_p(1), \quad m = 1,2,\end{aligned}$$

where $IF_{\hat{p}_{ki}}$ is the influence function (Hampel et al., 1986) that is specific to the estimated CIF from a particular competing risks model, and will be discussed again in the following paragraph. Denote $B_{ki(m)}(t_0) = \big(\hat{p}_{ki}(\mathscr{M}_m, t_0) - p_k(\mathscr{M}_m, t_0)\big)^2 - D_{k(m)}$. By Taylor's expansion:

$$\sup_{t_0}\left|\sqrt{n}\big(\hat{R}(t_0) - R(t_0)\big) - \frac{\sqrt{n}}{n(n-1)}\sum_{i\neq j}^{n}\Psi_{ij}^{**}(t_0)\right| = o_p(1),$$

where

$$\Psi_{ij}^{**}(t_0) = \sum_{k=1}^{3}\omega_k\left\{\frac{B_{ki(2)} - B_{ki(1)}}{\pi_k(t_0)\big(1-\pi_k(t_0)\big)} + \frac{Q_{ij,k}(t_0)\big(D_{k(2)} - D_{k(1)}\big)\big(2\pi_k(t_0) - 1\big)}{\pi_k^2(t_0)\big(1-\pi_k(t_0)\big)^2}\right\}.$$

By Hájek's projection principle , the following Hoeffding decomposition (van der Vaart, 1998; Serfling, 1980) holds:

$$\frac{\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Psi_{ij}^{**}(t_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}IF^{**}(X_i, \eta_i, M_i, t_0) + o_p(1),$$

where

$$IF^{**}(X_i, \eta_i, M_i, t_0) = E\Big(\Psi_{ij}^{**}(t_0) + \Psi_{ji}^{**}(t_0)\Big|(X_i, \eta_i, M_i)\Big) \tag{B.1}$$

and $E\Big[IF^{**}(X_i, \eta_i, M_i, t_0)\Big] = 0$. Using the same procedure from the previous proof, we can also estimate $IF^{**}(X_i, \eta_i, M_i, t_0)$ using the sample.

APPENDIX C. SUPPLEMENTAL TABLES FOR SECTION 4.3

For the supplemental tables, results from correct models are given in bold. Under alternative hypothesis when the added covariate improves predictability, true means S and R were calculated using 1,000 samples with size 1,000. Under null, true means S and R are 0. 1,000 samples with size 200 or 400 each was used to calculate the sample means $\hat{S}$, $\hat{R}$ and empirical standard error SE. Mean of estimated standard deviations $\text{SD}_{\hat{S}}$ was outputted by NRI formula provided. Coverage rate $\text{CR}_{\hat{S}}$=(count of true NRI entering the intervals $[\hat{S}-1.96\text{SD}_{\hat{S}}, \hat{S}+1.96\text{SD}_{\hat{S}}]$)/1,000. Each sample was bootstrapped 1,000 times, and the mean of 1,000 bootstrap standard deviations is denoted as BSD. Coverage rate BCR=(count of true value entering the 95% BCa bootstrap intervals)/1,000.

Table 9: Simulation details for the NRI under alternative (50% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 1*)	S	**.114**	**.100**	**.126**	.114	.100	.126	.114	.100	.126
	$\hat{S}$	**.113**	**.106**	**.124**	.089	.107	.119	.109	.105	.117
	$SE_{\hat{S}}$	**.030**	**.036**	**.038**	.027	.033	.040	.030	.035	.038
	$SD_{\hat{S}}$	**.028**	**.035**	**.036**	.022	.033	.034	.027	.035	.036
	$CR_{\hat{S}}$	**.930**	**.933**	**.933**	.725	.945	.903	.925	.945	.909
	$BCR_{\hat{S}}$	**.908**	**.888**	**.905**	.867	.892	.910	.918	.908	.897
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 2*)	S	.109	.127	.123	**.109**	**.127**	**.123**	.109	.127	.123
	$\hat{S}$	.116	.131	.126	**.112**	**.125**	**.125**	.120	.131	.126
	$SE_{\hat{S}}$	.041	.038	.034	**.042**	**.040**	**.036**	.038	.037	.034
	$SD_{\hat{S}}$	.038	.035	.030	**.039**	**.037**	**.032**	.038	.033	.030
	$CR_{\hat{S}}$	.915	.915	.908	**.933**	**.935**	**.906**	.934	.918	.887
	$BCR_{\hat{S}}$	.892	.899	.930	**.912**	**.921**	**.934**	.896	.907	.934
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 3*)	S	.209	.189	.169	.209	.189	.169	**.209**	**.189**	**.169**
	$\hat{S}$	.206	.189	.176	.126	.156	.159	**.208**	**.187**	**.172**
	$SE_{\hat{S}}$	.031	.036	.037	.028	.029	.034	**.032**	**.035**	**.034**
	$SD_{\hat{S}}$	.028	.031	.036	.022	.026	.033	**.030**	**.032**	**.036**
	$CR_{\hat{S}}$	.911	.923	.940	.110	.756	.928	**.917**	**.936**	**.950**
	$BCR_{\hat{S}}$	.914	.912	.912	.357	.880	.918	**.927**	**.921**	**.916**

Table 10: Simulation details for the IDI under alternative (50% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 1*)	R	**.109**	**.106**	**.099**	.109	.106	.099	.109	.106	.099
	$\hat{R}$	**.110**	**.108**	**.101**	.035	.048	.066	.096	.087	.075
	$\text{SE}_{\hat{R}}$	**.018**	**.018**	**.018**	.010	.013	.017	.018	.018	.017
	$\text{BSD}_{\hat{R}}$	**.021**	**.020**	**.018**	.012	.014	.016	.021	.019	.018
	$\text{BCR}_{\hat{R}}$	**.965**	**.959**	**.928**	.002	.081	.622	.911	.859	.825
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 2*)	R	.151	.158	.165	**.151**	**.158**	**.165**	.151	.158	.165
	$\hat{R}$	.131	.140	.148	**.146**	**.153**	**.160**	.037	.046	.055
	$\text{SE}_{\hat{R}}$	.023	.024	.025	**.024**	**.024**	**.025**	.016	.016	.017
	$\text{BSD}_{\hat{R}}$	.026	.027	.028	**.025**	**.025**	**.025**	.019	.020	.020
	$\text{BCR}_{\hat{R}}$	.757	.816	.856	**.921**	**.925**	**.924**	.021	.008	.006
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 3*)	R	.288	.271	.251	.288	.271	.251	**.288**	**.271**	**.251**
	$\hat{R}$	.282	.267	.250	.089	.113	.143	**.288**	**.270**	**.251**
	$\text{SE}_{\hat{R}}$	.028	.026	.025	.017	.020	.023	**.022**	**.021**	**.021**
	$\text{BSD}_{\hat{R}}$	.033	.031	.029	.020	.022	.023	**.028**	**.025**	**.023**
	$\text{BCR}_{\hat{R}}$	.634	.802	.911	0	0	.031	**.911**	**.956**	**.957**

Table 11: Simulation details for the NRI under null (50% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 4*)	$\hat{S}$	**.006**	**.005**	**.006**	.006	.005	.005	.006	.005	.006
	$\text{SE}_{\hat{S}}$	**.018**	**.017**	**.018**	.017	.017	.018	.017	.018	.018
	$\text{SD}_{\hat{S}}$	**.015**	**.016**	**.017**	.013	.016	.017	.016	.017	.017
	$\text{CR}_{\hat{S}}$	**.897**	**.934**	**.922**	.882	.915	.926	.920	.927	.921
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	**.943**	**.950**	**.945**	.916	.943	.938	.937	.940	.964
	$\text{BCR}_{\hat{S}}$	**.857**	**.884**	**.871**	.840	.891	.881	.831	.809	.851
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 5*)	$\hat{S}$	.007	.006	.004	**.005**	**.005**	**.007**	.006	.003	.002
	$\text{SE}_{\hat{S}}$	.020	.019	.014	**.018**	**.020**	**.014**	.020	.020	.014
	$\text{SD}_{\hat{S}}$	.018	.018	.013	**.017**	**.018**	**.013**	.018	.018	.013
	$\text{CR}_{\hat{S}}$	.907	.933	.893	**.919**	**.911**	**.882**	.904	.921	.895
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.929	.958	.952	**.941**	**.959**	**.952**	.933	.953	.948
	$\text{BCR}_{\hat{S}}$	.857	.888	.903	**.891**	**.888**	**.899**	.879	.874	.892
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 6*)	$\hat{S}$	.005	.006	.005	.006	.008	.007	**.005**	**.005**	**.006**
	$\text{SE}_{\hat{S}}$	.019	.018	.018	.016	.017	.018	**.019**	**.018**	**.018**
	$\text{SD}_{\hat{S}}$	.016	.016	.017	.011	.014	.016	**.017**	**.017**	**.017**
	$\text{CR}_{\hat{S}}$	.904	.920	.923	.808	.878	.909	**.911**	**.922**	**.904**
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.948	.950	.946	.900	.920	.942	**.929**	**.954**	**.948**
	$\text{BCR}_{\hat{S}}$	.863	.860	.859	.803	.825	.867	**.846**	**.863**	**.879**

Table 12: Simulation details for the IDI under null (50% censoring, 400 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 4*)	$\hat{R}$	**.003**	**.003**	**.003**	.003	.003	.004	.002	.003	.003
	$SE_{\hat{R}}$	**.003**	**.003**	**.004**	.004	.004	.005	.003	.003	.003
	$BSD_{\hat{R}}$	**.006**	**.006**	**.007**	.006	.007	.008	.005	.006	.006
	$BCR_{\hat{R}}$	**.964**	**.957**	**.947**	.943	.941	.927	.958	.958	.953
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 5*)	$\hat{R}$	.003	.004	.004	**.003**	**.003**	**.004**	.0006	.0008	.001
	$SE_{\hat{R}}$	.003	.004	.004	**.003**	**.004**	**.004**	.0007	.001	.001
	$BSD_{\hat{R}}$	.006	.007	.007	**.005**	**.006**	**.006**	.002	.002	.003
	$BCR_{\hat{R}}$	.905	.883	.874	**.823**	**.807**	**.785**	.865	.872	.870
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 6*)	$\hat{R}$	.003	.004	.004	.003	.003	.004	**.002**	**.003**	**.003**
	$SE_{\hat{R}}$	.004	.004	.004	.004	.004	.005	**.003**	**.003**	**.004**
	$BSD_{\hat{R}}$	.008	.008	.008	.008	.007	.008	**.005**	**.006**	**.006**
	$BCR_{\hat{R}}$	.964	.946	.948	.960	.959	.946	**.969**	**.961**	**.948**

Table 13: Simulation details for the NRI under alternative (30% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 1*)	S	**.114**	**.100**	**.126**	.114	.100	.126	.114	.100	.126
	$\hat{S}$	**.111**	**.105**	**.122**	.073	.104	.093	.108	.106	.117
	$SE_{\hat{S}}$	**.038**	**.042**	**.043**	.037	.041	.046	.039	.043	.046
	$SD_{\hat{S}}$	**.035**	**.041**	**.042**	.024	.038	.036	.035	.041	.041
	$CR_{\hat{S}}$	**.914**	**.945**	**.932**	.546	.922	.759	.908	.948	.905
	$BCR_{\hat{S}}$	**.906**	**.895**	**.903**	.772	.852	.848	.900	.892	.895
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 2*)	S	.109	.127	.123	**.109**	**.127**	**.123**	.109	.127	.123
	$\hat{S}$	.118	.130	.127	**.114**	**.127**	**.124**	.121	.131	.126
	$SE_{\hat{S}}$	.048	.043	.040	**.050**	**.048**	**.043**	.046	.042	.037
	$SD_{\hat{S}}$	.044	.041	.036	**.046**	**.044**	**.038**	.044	.040	.035
	$CR_{\hat{S}}$	.915	.917	.908	**.915**	**.920**	**.908**	.928	.935	.918
	$BCR_{\hat{S}}$	.878	.899	.921	**.889**	**.903**	**.911**	.882	.906	.933
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 3*)	S	.209	.189	.169	.209	.189	.169	**.209**	**.189**	**.169**
	$\hat{S}$	.209	.192	.180	.081	.145	.168	**.208**	**.191**	**.175**
	$SE_{\hat{S}}$	.040	.044	.043	.044	.040	.041	**.039**	**.043**	**.044**
	$SD_{\hat{S}}$	.036	.038	.042	.022	.030	.037	**.036**	**.039**	**.042**
	$CR_{\hat{S}}$	.917	.905	.933	.061	.645	.920	**.931**	**.923**	**.933**
	$BCR_{\hat{S}}$	.915	.898	.892	.226	.816	.874	**.924**	**.909**	**.979**

Table 14: Simulation details for the IDI under alternative (30% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 1*)	R	**.109**	**.106**	**.099**	.109	.106	.099	.109	.106	.099
	$\hat{R}$	**.115**	**.113**	**.106**	.021	.028	.037	.100	.092	.080
	$\text{SE}_{\hat{R}}$	**.025**	**.024**	**.023**	.008	.010	.013	.026	.025	.024
	$\text{BSD}_{\hat{R}}$	**.027**	**.026**	**.025**	.013	.014	.016	.028	.026	.025
	$\text{BCR}_{\hat{R}}$	**.961**	**.965**	**.969**	.004	.019	.137	.937	.916	.898
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 2*)	R	.151	.158	.165	**.151**	**.158**	**.165**	.151	.158	.165
	$\hat{R}$	.131	.141	.149	**.148**	**.156**	**.163**	.124	.135	.146
	$\text{SE}_{\hat{R}}$	.029	.030	.032	**.030**	**.030**	**.031**	.027	.028	.030
	$\text{BSD}_{\hat{R}}$	.034	.035	.036	**.031**	**.031**	**.032**	.030	.031	.033
	$\text{BCR}_{\hat{R}}$	.854	.877	.893	**.959**	**.950**	**.947**	.800	.855	.897
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 3*)	R	.288	.271	.251	.288	.271	.251	**.288**	**.271**	**.251**
	$\hat{R}$	.288	.272	.254	.038	.049	.062	**.294**	**.277**	**.258**
	$\text{SE}_{\hat{R}}$	.037	.035	.033	.011	.014	.016	**.029**	**.027**	**.026**
	$\text{BSD}_{\hat{R}}$	.049	.044	.041	.018	.020	.021	**.035**	**.032**	**.031**
	$\text{BCR}_{\hat{R}}$	.789	.822	.873	0	0	0	**.971**	**.976**	**.961**

Table 15: Simulation details for the NRI under null (30% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 4*)	$\hat{S}$	**.007**	**.006**	**.006**	.006	.006	.006	.007	.008	.007
	$\text{SE}_{\hat{S}}$	**.023**	**.022**	**.022**	.020	.023	.024	.022	.024	.023
	$\text{SD}_{\hat{S}}$	**.019**	**.020**	**.021**	.016	.020	.023	.020	.021	.022
	$\text{CR}_{\hat{S}}$	**.886**	**.893**	**.908**	.838	.889	.937	.901	.899	.927
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	**.964**	**.931**	**.941**	.903	.929	.962	.938	.939	.946
	$\text{BCR}_{\hat{S}}$	**.834**	**.855**	**.840**	.822	.842	.868	.853	.826	.857
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 5*)	$\hat{S}$	.009	.007	.006	**.009**	**.006**	**.005**	.009	.008	.006
	$\text{SE}_{\hat{S}}$	.024	.024	.019	**.024**	**.024**	**.019**	.027	.025	.020
	$\text{SD}_{\hat{S}}$	.022	.023	.017	**.021**	**.021**	**.016**	.023	.024	.018
	$\text{CR}_{\hat{S}}$	.899	.923	.903	**.890**	**.902**	**.851**	.904	.923	.890
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.939	.954	.953	**.923**	**.952**	**.942**	.937	.944	.939
	$\text{BCR}_{\hat{S}}$	.873	.880	.882	**.863**	**.862**	**.862**	.849	.869	.875
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 6*)	$\hat{S}$	.007	.007	.007	.006	.008	.007	**.007**	**.008**	**.008**
	$\text{SE}_{\hat{S}}$	.025	.024	.022	.017	.021	.022	**.024**	**.024**	**.023**
	$\text{SD}_{\hat{S}}$	.020	.021	.021	.010	.015	.020	**.021**	**.021**	**.021**
	$\text{CR}_{\hat{S}}$	.874	.898	.913	.659	.803	.876	**.866**	**.890**	**.896**
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.813	.837	.861	.640	.844	.857	**.925**	**.917**	**.924**
	$\text{BCR}_{\hat{S}}$	.914	.918	.897	.019	.757	.894	**.827**	**.840**	**.834**

Table 16: Simulation details for the IDI under null (30% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 4*)	$\hat{R}$	**.005**	**.005**	**.005**	.004	.004	.005	.004	.004	.005
	$\text{SE}_{\hat{R}}$	**.005**	**.006**	**.006**	.005	.006	.007	.005	.005	.005
	$\text{BSD}_{\hat{R}}$	**.010**	**.011**	**.011**	.009	.010	.011	.008	.009	.009
	$\text{BCR}_{\hat{R}}$	**.974**	**.963**	**.960**	.927	.922	.917	.884	.912	.918
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 5*)	$\hat{R}$	.005	.005	.006	**.004**	**.004**	**.005**	.005	.005	.005
	$\text{SE}_{\hat{R}}$	.005	.005	.006	**.004**	**.004**	**.005**	.005	.005	.006
	$\text{BSD}_{\hat{R}}$	.009	.010	.011	**.007**	**.008**	**.009**	.009	.009	.009
	$\text{BCR}_{\hat{R}}$	.896	.892	.883	**.800**	**.793**	**.774**	.832	.834	.828
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 6*)	$\hat{R}$	.006	.006	.006	.004	.004	.005	**.004**	**.004**	**.004**
	$\text{SE}_{\hat{R}}$	.007	.007	.007	.005	.006	.007	**.004**	**.004**	**.004**
	$\text{BSD}_{\hat{R}}$	.013	.012	.013	.009	.010	.011	**.009**	**.009**	**.009**
	$\text{BCR}_{\hat{R}}$	.976	.972	.971	.923	.916	.909	**.909**	**.906**	**.910**

Table 17: Simulation details for the NRI under alternative (50% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 1*)	S	**.114**	**.100**	**.126**	.114	.100	.126	.114	.100	.126
	$\hat{S}$	**.111**	**.107**	**.117**	.090	.105	.111	.109	.108	.117
	$\text{SE}_{\hat{S}}$	**.043**	**.049**	**.052**	.039	.045	.055	.044	.050	.054
	$\text{SD}_{\hat{S}}$	**.040**	**.048**	**.051**	.032	.045	.047	.039	.048	.050
	$\text{CR}_{\hat{S}}$	**.930**	**.939**	**.928**	.794	.940	.877	.912	.937	.906
	$\text{BCR}_{\hat{S}}$	**.908**	**.892**	**.905**	.881	.888	.888	.911	.893	.897
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 2*)	S	.109	.127	.123	**.109**	**.127**	**.123**	.109	.127	.123
	$\hat{S}$	.122	.131	.128	**.109**	**.123**	**.124**	.123	.130	.126
	$\text{SE}_{\hat{S}}$	.054	.055	.049	**.056**	**.056**	**.053**	.054	.049	.044
	$\text{SD}_{\hat{S}}$	.051	.050	.044	**.053**	**.053**	**.048**	.051	.047	.042
	$\text{CR}_{\hat{S}}$	.929	.920	.918	**.924**	**.923**	**.924**	.930	.940	.927
	$\text{BCR}_{\hat{S}}$	.883	.885	.909	**.891**	**.900**	**.917**	.878	.927	.931
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 3*)	S	.209	.189	.169	.209	.189	.169	**.209**	**.189**	**.169**
	$\hat{S}$	.208	.194	.180	.135	.159	.164	**.210**	**.193**	**.177**
	$\text{SE}_{\hat{S}}$	.047	.049	.051	.041	.042	.049	**.045**	**.049**	**.053**
	$\text{SD}_{\hat{S}}$	.041	.045	.051	.032	.038	.046	**.041**	**.045**	**.050**
	$\text{CR}_{\hat{S}}$	.924	.925	.942	.399	.836	.929	**.923**	**.921**	**.927**
	$\text{BCR}_{\hat{S}}$	.908	.891	.902	.700	.878	.892	**.909**	**.896**	**.889**

Table 18: Simulation details for the IDI under alternative (50% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 1*)	R	**.109**	**.106**	**.099**	.109	.106	.099	.109	.106	.099
	$\hat{R}$	**.114**	**.113**	**.106**	.040	.053	.071	.098	.090	.079
	$\text{SE}_{\hat{R}}$	**.029**	**.028**	**.028**	.016	.019	.026	.029	.028	.027
	$\text{BSD}_{\hat{R}}$	**.035**	**.033**	**.032**	.021	.023	.027	.033	.031	.029
	$\text{BCR}_{\hat{R}}$	**.966**	**.968**	**.966**	.130	.374	.800	.933	.913	.899
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 2*)	R	.151	.158	.165	**.151**	**.158**	**.165**	.151	.158	.165
	$\hat{R}$	.135	.145	.154	**.149**	**.156**	**.164**	.126	.138	.150
	$\text{SE}_{\hat{R}}$	.033	.035	.036	**.034**	**.034**	**.036**	.031	.033	.035
	$\text{BSD}_{\hat{R}}$	.044	.045	.048	**.037**	**.037**	**.040**	.036	.038	.042
	$\text{BCR}_{\hat{R}}$	.899	.927	.942	**.941**	**.950**	**.951**	.853	.897	.934
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 3*)	R	.288	.271	.251	.288	.271	.251	**.288**	**.271**	**.251**
	$\hat{R}$	.305	.290	.273	.096	.119	.149	**.292**	**.276**	**.257**
	$\text{SE}_{\hat{R}}$	.043	.040	.038	.026	.029	.034	**.034**	**.031**	**.031**
	$\text{BSD}_{\hat{R}}$	.044	.040	.035	.033	.034	.037	**.047**	**.042**	**.040**
	$\text{BCR}_{\hat{R}}$	.764	.830	.882	.001	.013	.239	**.956**	**.967**	**.965**

Table 19: Simulation details for the NRI under null (50% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$	$S(11)$	$S(12)$	$S(13)$
Weibull (*Case 4*)	$\hat{S}$	**.008**	**.007**	**.008**	.009	.009	.006	.008	.007	.007
	$\text{SE}_{\hat{S}}$	**.024**	**.025**	**.026**	.023	.026	.029	.027	.026	.027
	$\text{SD}_{\hat{S}}$	**.022**	**.023**	**.025**	.020	.023	.027	.024	.025	.027
	$\text{CR}_{\hat{S}}$	**.871**	**.898**	**.902**	.848	.884	.911	.877	.923	.923
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	**.952**	**.950**	**.950**	.923	.923	.944	.935	.946	.952
	$\text{BCR}_{\hat{S}}$	**.855**	**.864**	**.866**	.846	.863	.853	.836	.853	.867
		$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$	$S(20)$	$S(21)$	$S(22)$
Fine Gray (*Case 5*)	$\hat{S}$	.011	.011	.008	**.011**	**.006**	**.007**	.012	.009	.008
	$\text{SE}_{\hat{S}}$	.030	.032	.026	**.029**	**.029**	**.024**	.032	.031	.026
	$\text{SD}_{\hat{S}}$	.028	.029	.022	**.027**	**.028**	**.023**	.028	.029	.022
	$\text{CR}_{\hat{S}}$	.896	.903	.866	**.903**	**.922**	**.885**	.889	.893	.875
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.935	.953	.948	**.931**	**.949**	**.944**	.935	.937	.939
	$\text{BCR}_{\hat{S}}$	.864	.876	.877	**.872**	**.874**	**.885**	.870	.863	.886
		$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$	$S(9)$	$S(10)$	$S(11)$
Gerds (*Case 6*)	$\hat{S}$	.007	.008	.008	.009	.009	.008	**.008**	**.009**	**.008**
	$\text{SE}_{\hat{S}}$	.029	.028	.027	.023	.026	.025	**.028**	**.027**	**.026**
	$\text{SD}_{\hat{S}}$	.024	.024	.025	.016	.020	.023	**.025**	**.024**	**.025**
	$\text{CR}_{\hat{S}}$	.860	.877	.897	.761	.832	.867	**.870**	**.882**	**.902**
	$\text{CR}_{\hat{S}}^{\text{unnest}}$	.952	.944	.955	.892	.923	.947	**.929**	**.942**	**.939**
	$\text{BCR}_{\hat{S}}$	.814	.826	.856	.778	.812	.852	**.822**	**.843**	**.851**

Table 20: Simulation details for the IDI under null (50% censoring, 200 sample size)

		Cox Regression			Fine Gray			Gerds		
		$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$	$R(11)$	$R(12)$	$R(13)$
Weibull (*Case 4*)	$\hat{R}$	**.006**	**.007**	**.007**	.006	.007	.008	.006	.006	.006
	$\text{SE}_{\hat{R}}$	**.007**	**.008**	**.009**	.008	.008	.009	.006	.007	.007
	$\text{BSD}_{\hat{R}}$	**.015**	**.016**	**.017**	.013	.014	.015	.012	.013	.014
	$\text{BCR}_{\hat{R}}$	**.987**	**.987**	**.983**	.934	.921	.916	.867	.881	.900
		$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$	$R(20)$	$R(21)$	$R(22)$
Fine Gray (*Case 5*)	$\hat{R}$	.007	.008	.008	**.006**	**.007**	**.008**	.006	.006	.006
	$\text{SE}_{\hat{R}}$	.007	.008	.009	**.006**	**.007**	**.008**	.006	.007	.007
	$\text{BSD}_{\hat{R}}$	.014	.015	.016	**.011**	**.012**	**.013**	.012	.012	.015
	$\text{BCR}_{\hat{R}}$	.931	.921	.918	**.851**	**.841**	**.824**	.792	.796	.803
		$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$	$R(9)$	$R(10)$	$R(11)$
Gerds (*Case 6*)	$\hat{R}$	.008	.008	.008	.006	.007	.008	**.005**	**.006**	**.006**
	$\text{SE}_{\hat{R}}$	.008	.009	.009	.009	.010	.011	**.006**	**.006**	**.007**
	$\text{BSD}_{\hat{R}}$	.023	.022	.023	.014	.015	.016	**.013**	**.013**	**.014**
	$\text{BCR}_{\hat{R}}$	.990	.986	.987	.945	.930	.920	**.887**	**.894**	**.911**

Table 21: Simulation details under cross-validation for the NRI and IDI under alternative with all results are from the correct models (400 sample size)

30% Censoring		Weibull (*Case 1*)			Fine Gray (*Case 2*)			Gerds (*Case 3*)		
		$S(11)$	$S(12)$	$S(13)$	$S(20)$	$S(21)$	$S(22)$	$S(9)$	$S(10)$	$S(11)$
NRI	S	.107	.099	.128	.114	.128	.124	.187	.173	.164
	$\hat{S}$	.100	.107	.126	.122	.128	.130	.181	.164	.159
	$\text{SE}_{\hat{S}}$	.028	.034	.034	.037	.036	.032	.032	.031	.032
	$\text{BSD}_{\hat{S}}$	.031	.036	.036	.040	.039	.035	.034	.034	.035
	$\text{BCR}_{\hat{S}}$	.842	.828	.840	.881	.872	.858	.774	.805	.813
		$R(11)$	$R(12)$	$R(13)$	$R(20)$	$R(21)$	$R(22)$	$R(9)$	$R(10)$	$R(11)$
IDI	R	.109	.107	.099	.151	.158	.164	.281	.267	.240
	$\hat{R}$	.111	.109	.101	.148	.155	.162	.292	.262	.225
	$\text{SE}_{\hat{R}}$	.017	.016	.016	.021	.021	.022	.021	.021	.026
	$\text{BSD}_{\hat{R}}$	.017	.017	.016	.021	.021	.022	.023	.023	.025
	$\text{BCR}_{\hat{R}}$	.955	.954	.937	.942	.934	.931	.951	.903	.816
50% Censoring		Weibull (*Case 1*)			Fine Gray (*Case 2*)			Gerds (*Case 3*)		
		$S(11)$	$S(12)$	$S(13)$	$S(20)$	$S(21)$	$S(22)$	$S(9)$	$S(10)$	$S(11)$
NRI	S	.107	.099	.128	.114	.128	.124	.187	.173	.164
	$\hat{S}$	.096	.106	.124	.126	.128	.128	.181	.163	.159
	$\text{SE}_{\hat{S}}$	.032	.039	.040	.045	.042	.038	.038	.038	.039
	$\text{BSD}_{\hat{S}}$	.036	.043	.044	.048	.048	.045	.039	.041	.043
	$\text{BCR}_{\hat{S}}$	.826	.828	.846	.814	.879	.886	.774	.778	.803
		$R(11)$	$R(12)$	$R(13)$	$R(20)$	$R(21)$	$R(22)$	$R(9)$	$R(10)$	$R(11)$
IDI	R	.109	.107	.099	.151	.158	.164	.281	.267	.240
	$\hat{R}$	.112	.110	.103	.147	.154	.161	.290	.259	.222
	$\text{SE}_{\hat{R}}$	.019	.018	.018	.023	.024	.024	.024	.024	.027
	$\text{BSD}_{\hat{R}}$	.021	.020	.019	.025	.025	.025	.030	.027	.028
	$\text{BCR}_{\hat{R}}$	.958	.965	.932	.935	.944	.946	.934	.870	.826

Table 22: Simulation details under cross-validation for the NRI and IDI under null with all results are from the correct models (400 sample size)

30% Censoring		Weibull (*Case 4*)			Fine Gray (*Case 5*)			Gerds (*Case 6*)		
		$S(11)$	$S(12)$	$S(13)$	$S(20)$	$S(21)$	$S(22)$	$S(9)$	$S(10)$	$S(11)$
NRI	$\hat{S}$	.001	.0008	.001	-.0006	-.0008	-.001	.0008	-.0002	-.0005
	$\text{SE}_{\hat{S}}$	.017	.018	.019	.016	.016	.012	.019	.020	.020
	$\text{BSD}_{\hat{S}}$	.019	.020	.021	.020	.021	.017	.022	.022	.021
	$\text{BCR}_{\hat{S}}$	.747	.750	.789	.811	.811	.807	.743	.689	.713
		$R(11)$	$R(12)$	$R(13)$	$R(20)$	$R(21)$	$R(22)$	$R(9)$	$R(10)$	$R(11)$
IDI	$\hat{R}$	.003	.003	.003	.002	.003	.003	-.004	-.003	-.002
	$\text{SE}_{\hat{R}}$	.002	.002	.003	.002	.002	.003	.021	.021	.020
	$\text{BSD}_{\hat{R}}$	.004	.004	.005	.004	.004	.004	.026	.025	.025
	$\text{BCR}_{\hat{R}}$	.888	.871	.878	.606	.581	.543	.743	.750	.772
50% Censoring		Weibull (*Case 4*)			Fine Gray (*Case 5*)			Gerds (*Case 6*)		
		$S(11)$	$S(12)$	$S(13)$	$S(20)$	$S(21)$	$S(22)$	$S(9)$	$S(10)$	$S(11)$
NRI	$\hat{S}$	.0007	.0002	.0007	-.0005	-.002	-.002	.002	.003	.003
	$\text{SE}_{\hat{S}}$	.019	.020	.021	.021	.021	.016	.003	.003	.003
	$\text{BSD}_{\hat{S}}$	.023	.024	.025	.026	.028	.024	.004	.004	.005
	$\text{BCR}_{\hat{S}}$	.758	.760	.774	.776	.824	.845	.862	.844	.850
		$R(11)$	$R(12)$	$R(13)$	$R(20)$	$R(21)$	$R(22)$	$R(9)$	$R(10)$	$R(11)$
IDI	$\hat{R}$	.004	.004	.004	.003	.004	.005	.003	.003	.004
	$\text{SE}_{\hat{R}}$	.003	.004	.004	.003	.003	.004	.003	.003	.003
	$\text{BSD}_{\hat{R}}$	.006	.007	.007	.005	.006	.006	.006	.006	.007
	$\text{BCR}_{\hat{R}}$	.908	.885	.874	.622	.578	.537	.874	.868	.898

BIBLIOGRAPHY

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.*, 6(4):701–726.

Abrahamowicz, M., Mackenzie, T., and Esdaile, J. M. (1996). Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, 91(436):1432–1439.

Aharoni, E. and Rosset, S. (2013). Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794.

Andersen, P. K., Abildstrom, S. Z., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, 11(2):203–215.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

Antinori, A., Arendt, G., Becker, J. T., Brew, B. J., Byrd, D. A., Cherner, M., Clifford, D. B., Cinque, P., Epstein, L. G., Goodkin, K., Gisslen, M., Grant, I., Heaton, R. K., Joseph, J., Marder, K., Marra, C. M., McArthur, J. C., Nunn, M., Price, R. W., Pulliam, L., Robertson, K. R., Sacktor, N., Valcour, V., and Wojna, V. E. (2007). Updated research nosology for hiv-associated neurocognitive disorders. *Neurology*, 69(18):1789–1799.

Becker, J., Kingsley, L., Molsberry, S., Reynolds, S., Aronow, A., Levine, A., Martin, E., Miller, E., Munro, C., Ragin, A., Sacktor, N., and Selnes, O. (2014). Cohort Profile: Recruitment cohorts in the neuropsychological substudy of the Multicenter AIDS Cohort Study. *International Journal of Epidemiology*, 44(5):1506–1516.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, 19(10):1319–1328.

Beyersmann, J., Schumacher, M., and Allignol, A. (2012). *Competing risks and multistate models with R*. Springer, New York.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890–896.

Black, M. A. (2004). A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(2):297–304.

Blanche, P., Dartigues, J., and Jacqmin−Gadda, H. (2013). Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine*, 32:5381–5397.

Blanche, P., Proust-Lima, C., LoubÃĺre, L., Berr, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*, 71(1):102–113.

Bloxom, B. (1985). A constrained spline estimator of a hazard function. *Psychometrika*, 50(3):301–321.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 2(3):437–453.

Cai, T., Pepe, M. S., Zheng, Y., Lumley, T., and Jenny, N. S. (2006). The sensitivity and specificity of markers for event times. *Biostatistics*, 7(2):182–197.

Cheng, S., Fine, J., and Wei, L. (1998). Prediction of cumulative incidence function under the proportional hazards model. *Biometrics*, 54(1):219–228.

Cheng, Y. (2009). Modeling cumulative incidences of dementia and dementia-free death using a novel three-parameter logistic function. *The International Journal of Biostatistics*, 5(1):1557–4679.

Cheng, Y. and Fine, J. P. (2012). Cumulative incidence association models for bivariate competing risks data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):183–202.

Cheng, Y., Fine, J. P., and K., M. R. (2007). Nonparametric association analysis of bivariate competing-risks data. *Journal of the American Statistical Association*, 102(480):1407–1415.

Cheng, Y. and Li, J. (2015). Time-dependent diagnostic accuracy analysis with censored outcome and censored predictor. *Journal of Statistical Planning and Inference*, 156:90 – 102.

Chipman, J. and Braun, D. (2017). Simpson's paradox in the integrated discrimination improvement. *Statistics in Medicine*, 36(28):4468–4481.

Cook, N. R., Demler, O. V., and Paynter, N. P. (2017). Clinical risk reclassification at 10 years. *Statistics in Medicine*, 36(28):4498–4502.

Cortese, G., Gerds, T. A., and Andersen, P. K. (2013). Comparing predictions among competing risks models with time-dependent covariates. *Statistics in Medicine*, 32(18):3089–3101.

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

De Boor, C. (2001). *A practical guide to splines*. Springer, New York.

Demler, O. V., Pencina, M. J., Cook, N. R., and D'Agostino, R. B. (2017). Asymptotic distribution of δAUC, NRIs, and IDI based on theory of U-statistics. *Statistics in medicine*, 36(21):3334–3360.

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000). Comparing three−class diagnostic tests by three−way ROC analysis. *Medical Decision Making*, 20(3):323–331.

Dunn, O. J. (1959). Estimation of the medians for dependent variables. *Annals of Mathematical Statistics*, 30:192–197.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association,*, 56:52–64.

Durrleman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8(5):551–561.

Edwards, D. C., Metz, C. E., and Kupinski, M. A. (2004). Ideal observers and optimal roc hypersurfaces in n-class classification. *IEEE Transactions on Medical Imaging*, 23(7):891–895.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Springer, New York.

Fang, H., Tian, G., Xiong, X., and Tan, M. (2006). A multivariate random-effects model with restricted parameters: application to assessing radiation therapy for brain tumours. *Statistics in Medicine*, 25(11):1948–1959.

Farinpour, R., Miller, E., P., S., Selnes, O., Cohen, B., Becker, J., Skolasky, R., and Visscher, B. (2003). Psychosocial risk factors of HIV morbidity and mortality: Findings from the Multi-

center AIDS Cohort Study (MACS). *Journal of Clinical and Experimental Neuropsychology*, 25(5):654–670.

Fieuws, S. and Verbeke, G. (2004). Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach. *Statistics in Medicine*, 23(20):3093–3104.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62(2):424–431.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, 94(446):496–509.

Foster, D. P. and Stine, R. A. (2008). α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444.

Foucher, Y., Giral, M., Soulillou, J. P., and Daures, J. P. (2010). Time−dependent ROC analysis for a three−class prognostic with application to kidney transplantation. *Statistics in Medicine*, 29(30):3079–3087.

Geman, S. and Hwang, C. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414.

Gerds, T. and Schumacher, M. (2007). Efron-type measures of prediction error for survival analysis. *Biometrics*, 63(4):1283–1287.

Gerds, T. A., Scheike, T. H., and Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in Medicine*, 31(29):3921–3930.

Gerds, T. A. and Schumacher, M. (2006). Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6):1029–1040.

Gisslén, M., Price, R. W., and Nilsson, S. (2011). The definition of hiv-associated neurocognitive disorders: are we overestimating the real prevalence? *BMC Infectious Diseases*, 11(1):356.

Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18(6):695–706.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17âĂŘ18):2529–2545.

Graw, F., Gerds, T. A., and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis*, 15(2):241–255.

Greenland, P. and O'Malley, P. G. (2005). When is a new prediction marker useful? A consideration of lipoprotein-associated phospholipase a2 and c-reactive protein for stroke risk. *Archives of Internal Medicine*, 165(21):2454–2456.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. M., and Strahel, W. A. (1986). *Robust Statistics: The approach based on influence functions*. John Wiley & Sons, New York.

Harrell, F. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, New York.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time−dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344.

Heagerty, P. J. and Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, 61(1):92–105.

Heitjan, D. F. and Sharma, D. (1997). Modeling repeated-series longitudinal data. *Statistics in Medicine*, 16(4):347–355.

Herndon II, J. E. and Harrell Jr., F. E. (1990). The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods*, 19(2):639–663.

Huizenga, H. M., Smeding, H., Grasman, R. P. P. P., and Schmand, B. (2007). Multivariate normative comparisons. *Neuropsychologia*, 45(11):2534 – 2542.

Hung, H. and Chin-Tsang, C. (2010). Estimation methods for time-dependent auc models with survival data. *The Canadian Journal of Statistics*, 38(1):8–26.

Inácio, V., Turkman, A. A., Nakas, C. T., and Alonzo, T. A. (2011). Nonparametric bayesian estimation of the three-way receiver operating characteristic surface. *Biometrical Journal*, 53(6):1011–1024.

Janes, H. (2013). Letter to the editor. *Biostatistics*, 14(4):807–808.

Jeong, J. and Fine, J. (2006). Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55(2):187–200.

Jeong, J. and Fine, J. P. (2007). Parametric regression on cumulative incidence function. *Biostatistics*, 8(2):184–196.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of Failure time data*. John-Wiley & Sons, New York, 2nd edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kaslow, R., Ostrow, D., Detels, R., Phair, J., Polk, F., and Rinaldo, C.R., J. (1987). The multicenter AIDS cohort study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology*, 126(2):310–318.

Kingsley, L., Kaslow, R., Rinaldo, C. J. R., Detre, K., Odaka, N., and Vanraden, M. (1987). Risk factors for seroconversion to human immunodeficiency virus among male homosexuals. *The Lancet*, 329(8529):345–349.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015). Micro-randomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(Suppl):1220–1228.

Klein, J. P. (2006). Modelling competing risks in cancer studies. *Statistics in Medicine*, 25:1015–1034.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*. Springer, New York.

Leening, M. J., Steyerberg, E. W., Van Calster, B., D'Agostino Sr., R. B., and Pencina, M. J. (2014). Net reclassification improvement and integrated discrimination improvement require calibrated models: relevance from a marker and model perspective. *Statistics in Medicine*, 33(19):3415–3418.

Li, C. (2016). The fine-gray model under interval censored competing risks data. *Journal of Multivariate Analysis*, 143:327 – 344.

Li, J. and Fine, J. P. (2008). ROC analysis with multiple classes and multiple tests: methodology and its application in microarray studies. *Biostatistics*, 9(3):566–576.

Li, J., Jiang, B., and Fine, J. P. (2013a). Authors' response. *Biostatistics*, 14(4):809–810.

Li, J., Jiang, B., and Fine, J. P. (2013b). Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. *Biostatistics*, 14(2):382–394.

Li, J. and Zhou, X. (2009). Nonparametric and semiparametric estimation of the three way receiver operating characteristic surface. *Journal of Statistical Planning and Inference*, 139(12):4133 – 4142.

Meyer, M. C. and Habtzghi, D. (2011). Nonparametric estimation of density and hazard rate functions with shape restrictions. *Journal of Nonparametric Statistics*, 23(2):455–470.

Miller, E. N., Seines, O. A., McArthur, J. C., Satz, P., Becker, J. T., Cohen, B. A., Sheridan, K., Machado, A. M., Gorp, W. V., and Visscher, B. (1990). Neuropsychological performance in HIV-1-infected homosexual men. *Neurology*, 40(2):197–197.

Mossman, D. (1999). Three-way ROCs. *Medical Decision Making*, 19(1):78–89.

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., and Murphy, S. A. (2018). Just-in-time adaptive interventions (jitais) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, 52(6):446–462.

Pencina, M. J., D' Agostino, R. B., D' Agostino, R. B., and Vasan, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172.

Pencina, M. J., D'Agostino, R. B., and Steyerberg, E. W. (2011). Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Statistics in Medicine*, 30(1):11–21.

Pencina, M. J., Fine, J. P., and D'Agostino, R. B. (2017a). Discrimination slope and integrated discrimination improvement – properties, relationships and impact of calibration. *Statistics in Medicine*, 36(28):4482–4490.

Pencina, M. J., Steyerberg, E. W., and D'Agostino, R. B. (2017b). Net reclassification index at event rate: properties and relationships. *Statistics in Medicine*, 36(28):4455–4467.

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9):882–890.

Pepe, M. S., Zheng, Y., Jin, Y., Huang, Y., Parikh, C. R., and Levy, W. C. (2008). Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, 14(1):86–113.

Popov, M., Molsberry, S., Lecci, F., Junker, B., Kingsley, L., Levine, A., Martin, E., Miller, E., Munro, C., Ragin, A., Seaberg, E., Sacktor, N., and Becker, J. (2019). Brain structural correlates of trajectories to cognitive impairment in men with and without hiv disease. *Brain Imaging and Behavior*.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

Reinsel, G. (1982a). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77(377):190–195.

Reinsel, G. (1982b). Multivariate repeated-measurement or growth curve models with multivariate random-effects covariance structure. *Journal of the American Statistical Association*, 77(377):190–195.

Romano, J. P. and Wolf, M. (2005a). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Romano, J. P. and Wolf, M. (2005b). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282.

Rutherford, M. J., Crowther, M. J., and Lambert, P. C. (2015). The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: a simulation study. *Journal of Statistical Computation and Simulation*, 85(4):777–793.

Saha, P. and Heagerty, P. J. (2010). Time−dependent predictive accuracy in the presence of competing risks. *Biometrics*, 66(4):999–1011.

Scheike, T. H., Zhang, M., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression. *Biometrika*, 95(1):205–220.

Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. (2011a). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112.

Schoop, R., Beyersmann, J., Schumacher, M., and Binder, H. (2011b). Quantifying the predictive accuracy of time-to-event models in the presence of competing risks. *Biometrical Journal*, 53(1):88–112.

Schoop, R., Graf, E., and Schumacher, M. (2008). Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates. *Biometrics*, 64(2):603–610.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, New York.

Shen, X. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika*, 85(1):165–177.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615.

Shi, H., Cheng, Y., and Jeong, J. (2014a). Constrained parametric model for simultaneous inference of two cumulative incidence functions. *Biometrical Journal*, 55(1):82–96.

Shi, H., Cheng, Y., and Li, J. (2014b). Assessing diagnostic accuracy improvement for survival or competing−risk censored outcomes. *Canadian Journal of Statistics*, 42(1):109–125.

Shumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, Cambridge.

Smeden, M. and Moons, K. G. M. (2017). Event rate net reclassification index and the integrated discrimination improvement for studying incremental value of risk markers. *Statistics in Medicine*, 36(28):4495–4497.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Stone, C. J. (1986). Comment: generalized additive models. *Statistical Science*, 1(3):312–314.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Su, T., Schouten, J., Geurtsen, G. J., Wit, F. W., Stolte, I. G., Prins, M., Portegies, P., Caan, M. W. A., Reiss, P., Majoie, C. B., and Schmand, B. A. (2015). Multivariate normative comparison, a novel method for more reliably detecting cognitive impairment in hiv infection. *AIDS*, 29(5):547–557.

Tsiatis, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proceedings of the National Academy of Sciences*, 72(1):20–22.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117.

Uno, H., Tian, L., Cai, T., Kohane, I. S., and Wei, L. J. (2013). A unified inference procedure for a class of measures to assess improvement in risk prediction systems with survival data. *Statistics in Medicine*, 32(14):2430–2442.

van den Hout, A., Fox, J., and Muniz-Terrera, G. (2015). Longitudinal mixed-effects models for latent cognitive function. *Statistical Modelling*, 15(4):366–387.

van der Vaart, A. (1998). *Asymptoptic statistics*. Cambridge University Press, Cambridge.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23(1):42–59.

Wang, Z., Molsberry, S., Cheng, Y., Kingsley, L., Levine, A., Martin, E., Munro, C., A., R., L.H., R., Sacktor, N., Seaberg, E., and J.T., B. (2019). Cross-sectional analysis of cognitive function using multivariate normative comparisons in men with HIV disease. *AIDS*, 33(14):2115–2124.

Ware, J. H. (2006). The limitations of risk factors as prognostic tools. *New England Journal of Medicine*, 355(25):2615–2617.

Wasserstein, R., Schirm, A., and Lazar, N. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, 73:sup1:1–19.

Whittemore, A. S. and Keller, J. B. (1986). Survival estimation using splines. *Biometrics*, 42(3):495–506.

Wu, Y. and Chiang, C. (2013). Optimal receiver operating characteristic manifolds. *Journal of Mathematical Psychology*, 57(5):237 – 248. Special Issue: A Discussion of Publication Bias and the Test for Excess Significance.

Zhang, Y., Hua, L., and Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37(2):338–354.

Zhang, Y. and Li, J. (2011). Combining multiple markers for multi−category classification: An roc surface approach. *Australian & New Zealand Journal of Statistics*, 53(1):63–78.

Zheng, Y., Cai, T., Jin, Y., and Feng, Z. (2012). Evaluating prognostic accuracy of biomarkers under competing risk. *Biometrics*, 68(2):388–396.

Zheng, Y. and Heagerty, P. J. (2004). Semiparametric estimation of timeâĂŘdependent roc curves for longitudinal marker data. *Biostatistics*, 5(4):615–632.

Zheng, Y. and Heagerty, P. J. (2007). Prospective accuracy for longitudinal markers. *Biometrics*, 63(2):332–341.

Zheng, Y., Parast, L., Cai, T., and Brown, M. (2013). Evaluating incremental values from new predictors with net reclassification improvement in survival analysis. *Lifetime Data Analysis*, 19(3):350–370.