

**JOINT MODELING OF TIME-TO-EVENT DATA  
AND MULTIPLE RATINGS OF A DISCRETE  
DIAGNOSTIC TEST WITHOUT A GOLD  
STANDARD**

by

**Seung Hyun Won**

B.S. in Statistics, Chung-Ang University, Korea, 2007

M.S. in Biostatistics, Columbia University, 2009

Submitted to the Graduate Faculty of

Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

**Doctor of Philosophy**

University of Pittsburgh

2014

UNIVERSITY OF PITTSBURGH  
DEPARTMENT OF BIOSTATISTICS  
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Seung Hyun Won

It was defended on

August 20, 2014

and approved by

Gong Tang, Ph.D., Associate Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Yu Cheng, Ph.D., Associate Professor,

Department of Statistics, School of Arts and Sciences, University of Pittsburgh

Jong-Hyeon Jeong, Ph.D., Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Ruosha Li, Ph.D., Assistant Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Gong Tang, Ph.D., Associate Professor,

Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

# JOINT MODELING OF TIME-TO-EVENT DATA AND MULTIPLE RATINGS OF A DISCRETE DIAGNOSTIC TEST WITHOUT A GOLD STANDARD

Seung Hyun Won, PhD

University of Pittsburgh, 2014

## ABSTRACT

Histologic tumor grade is a strong predictor of risk of recurrence in breast cancer. Nevertheless, tumor grade readings by pathologists are susceptible to intra- and inter-observer variability due to its subject nature. Because of this limitation, histologic tumor grade is not included in the breast cancer staging system. Latent class models have been considered for analysis of such discrete diagnostic tests with regarding the underlying truth as a latent variable. However, the model parameters in latent class models are only locally identifiable, that is, any permutation on the categories of the underlying truth can lead to the same likelihood value.

In many clinical practices, the underlying truth is known associated with the risk of a certain event in a trend. Here, we proposed a joint model with a Cox proportional hazards model for time-to-event data where the underlying truth is a latent predictor and a latent class model for multiple ratings of a discrete diagnostic test without a gold standard. With the known association between the underlying truth and the risk of an event in a trend, the proposed joint model not only fully identifies all model parameters but also provides valid assessment of the association between the diagnostic test result and the risk of an event.

The modified EM algorithm was used for estimation with employing the survey-weighted Cox model in the M-step. To test whether the known trend imposed on model parameters can be assumed, we applied the Union-Intersection principle for the proposed joint model. The

proposed method is illustrated in the analysis of data from the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14 sub-study and through simulation studies.

The proposed method is relevant to public health fields, such as chronic diseases and psychiatry, where some components of the initial diagnostics are subjective but have important implications in patient management. Application of our method leads to accurate assessment on the association between the diagnostic tests and the clinical outcomes and subsequently significant improvement in decision-making on treatment or patient management.

**Keywords:** Discrete diagnostic test, Misclassification, Latent class model, EM algorithm, Survey-weighted Cox model, Order restricted testing, Union-Intersection principle.

## TABLE OF CONTENTS

<b>1.0 INTRODUCTION</b> . . . . .	1
<b>2.0 LITERATURE REVIEW</b> . . . . .	7
2.1 Existing Methods for analysis of data on diagnostic tests without a gold standard . . . . .	7
2.1.1 A latent class model for discrete diagnostic tests without a gold standard	8
2.1.2 EM algorithm for the latent class model (2.1) . . . . .	12
2.1.3 Local identifiability . . . . .	15
2.1.4 Conditional independence in latent class models . . . . .	19
2.2 Joint modeling of time-to-event outcome and covariates with measurement errors . . . . .	23
2.3 Survey-Weighted Cox models . . . . .	27
<b>3.0 PROPOSED JOINT ANALYSIS OF TIME-TO-EVENT DATA AND DISCRETE DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD</b>	29
3.1 A new joint modeling approach . . . . .	29
3.2 An EM algorithm for parameter estimation . . . . .	30
3.2.1 E-step . . . . .	30
3.2.2 M-step . . . . .	31
3.3 Variance Estimation from the profile likelihood . . . . .	34
<b>4.0 STATISTICAL TESTING WITH ORDER RESTRICTED HYPOTHESIS</b> . . . . .	36
4.1 Likelihood-base approaches . . . . .	37
4.1.1 General two-sided alternative: $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$ . . . . .	37

4.1.2	Constrained one-sided alternative: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta} \in C$ . . . . .	38
4.2	The Union-Intersection Principle for Cox proportional hazards models . . . . .	39
4.3	The Union-Intersection Principle for the proposed joint model . . . . .	44
<b>5.0</b>	<b>SIMULATION STUDIES</b> . . . . .	49
<b>6.0</b>	<b>APPLICATION: ANALYSIS OF DATA FROM THE NSABP B-14</b>	
<b>DATA</b>	. . . . .	55
<b>7.0</b>	<b>CONCLUSION</b> . . . . .	60
<b>BIBLIOGRAPHY</b>	. . . . .	62

## LIST OF TABLES

1	The scoring in the modified Bloom-Richardson grading system . . . . .	2
2	The modified Bloom-Richardson grading system: Tumor grade . . . . .	2
3	Independent tumor grade readings from three pathologists in the NSABP B-14 trial. . . . .	6
4	Estimated prevalences and classification rates from the method in Dawid and Skene (1974) . . . . .	18
5	Weights for the mixture distribution for the UI-test statistic . . . . .	47
6	Simulation setup: Classification rates for simulation studies . . . . .	50
7	Simulation results with $(\gamma, \beta_1, \beta_2) = (0.3, 0.5, 2.0)$ . . . . .	53
8	Estimated prevalences and classification rates from the proposed joint model	56
9	Parameter estimates and standard errors of Cox parameter estimates from the joint model. . . . .	57
10	The conditions for calculating the U-I test statistic for the NSABP B-14 data.	58
11	Estimated weights for the mixture distribution for the U-I test statistic . . .	59

## 1.0 INTRODUCTION

In cancer treatment, it is critical to diagnose cancer at early stage and administer the most effective treatment option to obtain better prognosis. Diagnostic tests are often considered to determine disease status along with several clinical factors and help physicians and patients make decisions on treatment options. In clinical practice, however, the true disease status called the ‘gold standard’ sometimes remains unknown because of either the subjective nature of the diagnostic tests themselves or limited resources or ethical issues to operate a test on all patients [14].

In breast cancer research, the histologic tumor grade has been well-known as a strong predictor of breast cancer recurrence [9, 20, 22]. The histologic tumor grade is evaluated by pathologists with a patient’s paraffin-embedded tumor tissue. After surgical removal of breast tumor, patient’s tumor sample is embedded in a block of paraffin. Once the paraffin-embedded tissue is ready, a pathologist evaluates the patient’s biopsy sample under a microscope and determines patient’s tumor grade based on the Bloom-Richardson grading system. The Bloom-Richardson grading system consists of the three features: tubular differentiation, nuclear features, and mitotic counts (Table 1). Once each of these features is scored from 1 to 3, a primary tumor is graded into one of three categories based on the total sum of these scores: Well, Moderately, and Poorly differentiated (correspond to Grade 1, 2, and 3, respectively, See Table 2). A patient diagnosed with a lower tumor grade is more likely to have better prognosis of breast cancer compared to one having a higher tumor grade.

Despite of the strong association between a histologic tumor grade and clinical outcomes such as overall survival or disease-free survival, a histologic tumor grade has not been included in the current breast cancer staging system of the American Joint Committee on Cancer (AJCC) due to disagreements in tumor grade readings by different pathologists. Even

Table 1: The scoring in the modified Bloom-Richardson grading system

Tubule formation	Score	Mitotic counts	Score	Nuclear Grade	Score
75% or more	1	1-10	1	Low	1
10-75% or more	2	11-20	2	Intermediate	2
10% or less	3	Greater than 20	3	High	3

Table 2: The modified Bloom-Richardson grading system: Tumor grade

Total score	Histologic Tumor Grade
3 to 5	Well-differentiated tumor (Grade 1)
6 to 7	Moderately-differentiated tumor (Grade 2)
8 to 9	Poorly-differentiated tumor (Grade 3)

pathologists use the same tumor grading system, such as the modified Bloom-Richardson criteria, the subjective aspects of grading systems can cause substantial discrepancies in tumor grade readings [12].

In the sub-study of the National Surgical Adjuvant Breast and Bowel Project (NSABP) trial B-14, tumor grades from 668 tamoxifen-treated participants with node-negative, estrogen receptor-positive breast cancer were independently assessed by three experienced pathologists who followed the modified Bloom-Richardson grading criteria [20]. All three pathologists were blinded from any other clinical information except for participant’s tumor scan of paraffin-embedded tissues and independently evaluated participant’s tumor grade [9, 20]. The criteria has three categories of tumor grade: ‘well-differentiated’ grade (Grade=1) indicating the least aggressive and slowly growing tumors, ‘moderately-differentiated’ grade (Grade=2), and ‘poorly-differentiated’ grade (Grade=3) indicating highly aggressive and progressed tumors.

In Theissig et al. [30], the agreements of tumor grade readings from three pathologists based on the modified Bloom-Richardson criteria were reached 72.3% . In the sub-study of

the B-14 trial, however, the overall agreement in tumor grade readings from three pathologists was only 43.3% and the agreements in tumor grade readings from any two pathologists were ranged from 58.8% to 64.8 % [20]. The tumor grade reading data are shown in Table 3. Given that each tumor grade reading is subject to misclassification, it is hard to determine true tumor grade for each patient. Subsequently, the assessment of the association between tumor grade and risk of breast cancer recurrence based on the readings from one pathologist can be substantially biased. This result points out that the true tumor grade should be identified to integrate tumor grade readings with other prognostic factors or biomarkers so that the diagnostic test results can guide better treatment decision. With more accurate diagnosis, patients with well-differentiated tumor can avoid over-treatment along with cytotoxic chemotherapy regimens. At the same time, patients with poorly-differentiated tumor can get more intensive chemotherapy to improve their survival.

Latent class models for independent ratings of a discrete diagnostic test without a gold standard have been widely studied [7, 11, 17, 34]. Mostly in these models, the underlying truth of the diagnostic test is considered as a latent variable and the multiple ratings given the underlying truth are assumed mutually independent. The Pearson chi-square goodness-of-fit test is usually employed to check whether the conditional independence assumption holds or not [17]. Under some regularity conditions, the model parameters in a latent class model are locally identifiable, that is, identifiable up to a permutation on the categories of the underlying truth [7, 11, 17]. In our motivating example on tumor grade readings, we can estimate the prevalence of three categories of tumor grade but cannot determine the actual level of tumor grade due to the local identifiability. Dawid and Skene [7] proposed to pick the set of estimates that yield higher classification rates for all raters. However, if some pathologists have difficulties in classifying tumor grade, it is hard to justify the approach proposed by Dawid and Skene.

In many circumstances, clinical outcomes that are associated with the underlying truth in a known trend may be available. Instead of relying on Dawid and Skene’s strategy to choose the set of estimates, we propose to incorporate auxiliary variables in a joint model and to utilize a known trend in the association between the underlying truth and the risk of a certain event for global identification. Under the condition that the multiple ratings given

the true status are independent of auxiliary variables, the parameters of interest in latent class models for a discrete diagnostic test can be fully identified.

Joint models have been widely studied for modeling time-to-event data and longitudinal outcomes or covariates measured with errors [13,15,16,31,35,36]. Typically, the time-to-event data are modeled by the Cox proportional hazards model and the longitudinal covariates with measurement errors are modeled by a random effects model [13,15,31,35,36]. These proposed joint models are flexible and enhance the performance of prediction in clinical outcomes by accounting for the measurement errors in covariates. However, these models require to replace the true values of covariates with predicted values from a separate sub-model and so this may entail additional assumptions for the specification of sub-models [13,15,16,31,35,36].

In this dissertation, the time-to-event data are also modeled by the Cox proportional hazards model with using the underlying true tumor grade as a predictor in addition to other classical clinical predictors. Instead of building a separate model with additional covariates to predict the latent truth as Larsen (2004, 2005) [15, 16], we incorporate the latent class model for a discrete diagnostic test [7, 11, 17] into our proposed method without requiring additional structures that ask an external information to predict the latent truth. The proposed joint modeling of multiple independent ratings and time-to-event data not only fully identifies the parameters in the latent class model but also provides valid assessment of the association between the diagnostic test and the risk of the event of interest.

The rest of the dissertation is organized as following. In Chapter 2, we review the existing methods for analyzing discrete diagnostic tests without a gold standard. After that, we discuss the issue of local identifiability in latent class models and review the application of the EM algorithm. This chapter also contains the reviews of the existing joint models for time-to-event data and covariates with measurement errors or misclassification. The overview of survey-weighted Cox model is also covered in this chapter. In Chapter 3, the proposed joint model is introduced with its details in procedures for parameter estimation. The profile likelihood approach for estimating standard errors of parameter estimates is also introduced in this chapter. In Chapter 4, the application of Union-Intersection method to provide a global test on the association between the underlying truth and the risk of an event is presented. The results of simulation studies for evaluating statistical properties of

the proposed joint model are discussed in Chapter 5. The analysis of the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14 data with the proposed joint model is illustrated in Chapter 6. In Chapter 7, we conclude with summary on the proposed method and discussion about future works.

Table 3: Independent tumor grade readings from three pathologists in the NSABP B-14 trial.

Pathologists				Pathologists			
1	2	3	Frequency $\{n_{klm}\}$	1	2	3	Frequency $\{n_{klm}\}$
W	W	W	49	M	P	W	2
W	W	M	27	M	P	M	38
W	M	W	7	M	P	P	12
W	M	M	23	P	W	W	1
W	M	P	1	P	W	M	5
M	W	W	55	P	W	P	2
M	W	M	82	P	M	M	29
M	W	P	3	P	M	P	15
M	M	W	17	P	P	W	1
M	M	M	189	P	P	M	44
M	M	P	15	P	P	P	51

<sup>1</sup>W = Well-differentiated; M = Moderately-differentiated; P = Poorly-differentiated

## 2.0 LITERATURE REVIEW

### 2.1 EXISTING METHODS FOR ANALYSIS OF DATA ON DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD

To evaluate the accuracy of a new diagnostic test, a gold standard is usually compared with the new test. Under the presence of a gold standard, the accuracy of tests can be measured by sensitivity and specificity. However, it is common that the true disease status cannot be obtained due to the complexity of the diagnostic test or insurmountable cost to perform the gold standard test. Even the gold standard test is free from the above limitations, the subjective natures of diagnostic procedures may prohibit the investigators from obtaining the true disease status. In past decades, the statistical methods for evaluating the performance of a diagnostic test with an unknown true disease status have been widely studied. Hui and Zhou [14] provided a comprehensive review of statistical methods for diagnostic tests without a gold standard under a variety of conditions. With the frequentist and bayesian perspectives, the authors discussed statistical methods to sensitivity and specificity of a new test and prevalence of disease with or without some known parameters. In addition, the authors covered a brief background of methods to estimate the parameters of interest when the assumption of conditional independence is not suitable. With pointing out the identifiability issues with an unknown truth, the authors introduced various methods for estimating the parameters of interest from multiple tests with a binary outcome in a single population or tests with a multinomial outcome in a multiple population. Among the variety of statistical methods for multiple tests without a gold standard, we focus on the classical latent class model and the local identifiability issues with providing a detailed review of works by McHugh [17], Goodman [11], and Dawid and Skene [7] in the following.

### 2.1.1 A latent class model for discrete diagnostic tests without a gold standard

Latent class models have been widely studied for analyzing multiple independent assessments with an unknown truth. A latent class simply means the underlying class membership that cannot be directly observed. With assuming the local independence among multiple assessments in a class membership, a latent class model aims to estimate the prevalence of class memberships in a latent class and the conditional probabilities of observed class memberships given a certain level of latent class.

McHugh [17] presented a latent class model for  $m$  observable Bernoulli random variables and a binary latent variable. In the paper, the sufficient conditions for local identifiability derived by Fisher information matrix and a chi-square goodness-of-fit test to determine the number of latent classes were discussed. Goodman [11] extended McHugh's latent class model to a latent class model for  $m$  polytomous observable variables with a 3-class latent variable. In addition, the author proposed an algorithm for obtaining maximum likelihood estimates of model parameters and generalized the sufficient conditions for the local identifiability in McHugh. A method to test overall model fit and the use of parameter constraints to achieve identifiability are also covered in Goodman [11].

For a subject  $i$ , we denote the multiple independent assessments from the  $s$ th rater by  $V_i^{(s)}$ , where  $i = 1, 2, \dots, n; s = 1, 2, \dots, S$ ,  $S$  is the total number of raters and each rating has  $K$  levels. Assuming that the latent class variable  $W_i$  has  $J$  levels, the latent class model is [11]:

$$P(W = j) = \pi_j, \quad P(V^{(s)} = k | W = j) = q_{jk}^{(s)}, \quad (2.1)$$

where  $j = 1, 2, \dots, J; k = 1, 2, \dots, K; s = 1, 2, \dots, S$  and with restrictions  $\sum_{j=1}^J \pi_j = \sum_{k=1}^K q_{jk}^{(s)} = 1; \pi_j \geq 0, q_{jk}^{(s)} \geq 0$ . In the equation (2.1), the  $\pi_j$  indicates the prevalence of the underlying truth at the  $j$ th level, and the  $q_{jk}^{(s)}$  indicates the classification rate for the  $s$ th rater when the level of rating is  $k$  given the level of the underlying truth is  $j$ . We assume that the multiple ratings  $V^{(s)}$  are conditionally independent given at a certain level of truth  $W$ . This assumption can be tested by using the Pearson chi-square goodness-of-fit test. The details are discussed in Section 2.1.4.

The likelihood function under the latent class model (2.1) is

$$L(\boldsymbol{\pi}, \mathbf{q}) = \prod_{i=1}^n \sum_{j=1}^J \left[ \pi_j \left\{ \prod_{s=1}^S \prod_{k=1}^K (q_{jk}^{(s)})^{I(v_i^{(s)}=k)} \right\} \right], \quad (2.2)$$

where  $I(v_i^{(s)} = k)$  is equal to 1, if the level of ratings for the  $i$ th subject evaluated by the  $s$ th pathologists is  $k$ , or 0, otherwise where  $s = 1, 2, \dots, S$ ;  $k = 1, 2, \dots, K$ .

To obtain maximum likelihood estimates for latent class models, iterative numerical procedures have been considered [7, 8, 11, 17]. McHugh [17] optimized the likelihood function by using the scoring system with first approximations on  $(\boldsymbol{\pi}, \mathbf{q})$ . The scoring system consists of the score and information functions for the parameters of interest  $(\boldsymbol{\pi}, \mathbf{q})$  in the latent class model with binary observed variables. The information functions are equivalent to the variances and covariances of the scores. With an initial value  $(\boldsymbol{\pi}^{(0)}, \mathbf{q}^{(0)})$ , the iterative algorithm with the scoring system from McHugh [17] is

$$\begin{aligned} \hat{\boldsymbol{\pi}}^{(t)} &= \boldsymbol{\pi}^{(t-1)} + \boldsymbol{\Delta}_{\hat{\boldsymbol{\pi}}^{(t-1)}} \\ \hat{\mathbf{q}}^{(t)} &= \mathbf{q}^{(t-1)} + \boldsymbol{\Delta}_{\hat{\mathbf{q}}^{(t-1)}}, \end{aligned}$$

where  $t = 1, 2, \dots$  and  $\boldsymbol{\Delta}_{\hat{\boldsymbol{\pi}}^{(t-1)}}$ ,  $\boldsymbol{\Delta}_{\hat{\mathbf{q}}^{(t-1)}}$  are calculated from the corresponding values for score and information functions. Here, the length of  $\hat{\boldsymbol{\pi}}^{(t-1)}$  is  $(J-1)$  with  $\hat{\pi}_j^{(t-1)} = 1 - \sum_{j=1}^{J-1} \hat{\pi}_j^{(t-1)}$ , where  $\sum_{j=1}^J \hat{\pi}_j^{(t-1)} = 1$ . First, the score values for parameters  $(\hat{\boldsymbol{\pi}}, \hat{\mathbf{q}}) = (\hat{\boldsymbol{\pi}}^{(t-1)}, \hat{\mathbf{q}}^{(t-1)})$  are calculated by [17]

$$\begin{aligned} S_{\hat{\pi}_j} &= \frac{\partial}{\partial \hat{\pi}_j} \log L(\boldsymbol{\pi}, \mathbf{q}), j = 1, 2, \dots, J-1 \\ S_{\hat{q}_{jk}^{(s)}} &= \frac{\partial}{\partial \hat{q}_{jk}^{(s)}} \log L(\boldsymbol{\pi}, \mathbf{q}); j = 1, 2, \dots, J, k = 1, 2; s = 1, 2, \dots, S. \end{aligned}$$

Second, the information values for parameters  $(\hat{\boldsymbol{\pi}}, \hat{\mathbf{q}}) = (\hat{\boldsymbol{\pi}}^{(t-1)}, \hat{\mathbf{q}}^{(t-1)})$  are calculated by [17]

$$\begin{aligned} I_{\hat{\pi}_j \hat{\pi}_{j'}} &= E(S_{\hat{\pi}_j} S_{\hat{\pi}_{j'}}), j, j' = 1, 2, \dots, J-1 \\ I_{\hat{q}_{jk}^{(s)} \hat{q}_{j'k'}^{(s')}} &= E(S_{\hat{q}_{jk}^{(s)}} S_{\hat{q}_{j'k'}^{(s')}}), j, j' = 1, 2, \dots, J; k, k' = 1, 2; s, s' = 1, 2, \dots, S \\ I_{\hat{\pi}_{j'} \hat{q}_{jk}^{(s)}} &= E(S_{\hat{\pi}_{j'}} S_{\hat{q}_{jk}^{(s)}}), j' = 1, 2, \dots, J-1; j = 1, 2, \dots, J; k = 1, 2; s = 1, 2, \dots, S. \end{aligned}$$

With the computed score and information values,  $\Delta_{\hat{\pi}}, \Delta_{\hat{q}}$ , where  $(\hat{\pi}, \hat{q}) = (\hat{\pi}^{(t-1)}, \hat{q}^{(t-1)})$  are calculated by [17]

$$\begin{aligned}\Delta_{\hat{\pi}_j} &= \sum_{j'=1}^{J-1} \tilde{I}_{\hat{\pi}_j \hat{\pi}_{j'}} \tilde{S}_{\hat{\pi}_{j'}} + \sum_{j'=1}^J \sum_{k=1}^2 \sum_{s=1}^S \tilde{I}_{\hat{\pi}_j \hat{q}_{j'k}^{(s)}} \tilde{S}_{\hat{q}_{j'k}^{(s)}} \\ \Delta_{\hat{q}_{jk}^{(s)}} &= \sum_{j'=1}^{J-1} \tilde{I}_{\hat{\pi}_{j'} \hat{q}_{jk}^{(s)}} \tilde{S}_{\hat{\pi}_{j'}} + \sum_{j'=1}^J \sum_{k'=1}^2 \sum_{s'=1}^S \tilde{I}_{\hat{q}_{jk}^{(s)} \hat{q}_{j'k'}^{(s')}} \tilde{S}_{\hat{q}_{j'k'}^{(s')}}\end{aligned}$$

where  $\tilde{S}, \tilde{I}$  indicate the score and information values at the  $(t-1)$ th iteration, respectively. This algorithm proposed by McHugh [17] is dealt only with the latent class model for 2-class observed variables. Goodman [11] introduced the iterative procedure which can be applied for a latent class model with  $m$ -class observed variables,  $m \geq 2$ . The iterative procedure calculates the maximum likelihood estimates in a  $m$ -class latent class model with an iterative proportional fitting algorithm. With an initial value  $(\boldsymbol{\pi}^{(0)}, \mathbf{q}^{(0)})$ , the iterative procedure is as follows [11]:

**Step 1:** With the current estimates  $\{\pi_j^{(t-1)}, q_{jk}^{(1,t-1)}, q_{jl}^{(2,t-1)}, q_{jm}^{(3,t-1)}\}$ ,  $j = 1, 2, \dots, J; k = 1, 2, \dots, K; l = 1, 2, \dots, L; m = 1, 2, \dots, M; t = 1, 2, \dots$ , first calculate  $\hat{p}_{jklm}^{(t)}$ , where

$$\hat{p}_{jklm}^{(t)} = \hat{\pi}_j^{(t-1)} q_{jk}^{(1,t-1)} q_{jl}^{(2,t-1)} q_{jm}^{(3,t-1)}.$$

With  $\hat{p}_{jklm}^{(t)}$ , calculate  $\hat{p}_{klm}^{(t)}$ , which is the estimated probability with  $(V^{(1)} = k, V^{(2)} = l, V^{(3)} = m)$  and  $\hat{p}_{j|klm}^{(t)}$ , which is the estimated conditional probability that the level of unknown truth is  $j$  given that  $(V^{(1)} = k, V^{(2)} = l, V^{(3)} = m)$ ,  $J; k = 1, 2, \dots, K; l = 1, 2, \dots, L; m = 1, 2, \dots, M; t = 1, 2, \dots$ :

$$\begin{aligned}\hat{p}_{klm}^{(t)} &= Pr(V^{(1)} = k, V^{(2)} = l, V^{(3)} = m) \\ &= \sum_{j=1}^J \hat{p}_{jklm}^{(t)} \\ \hat{p}_{j|klm}^{(t)} &= Pr[W = j | V^{(1)} = k, V^{(2)} = l, V^{(3)} = m] \\ &= \frac{\hat{p}_{jklm}^{(t)}}{\sum_{j=1}^J \hat{p}_{jklm}^{(t)}}.\end{aligned}$$

**Step 2:** Update the maximum likelihood estimates for  $(\boldsymbol{\pi}, \mathbf{q})$  with the following equations, where  $\tilde{p}_{klm}$  is the observed proportion of  $(V^{(1)} = k, V^{(2)} = l, V^{(3)} = m)$ ,  $j = 1, 2, \dots, J; k = 1, 2, \dots, K; l = 1, 2, \dots, L; m = 1, 2, \dots, M$ .

$$\hat{\pi}_j^{(t)} = \sum_{k,l,m} \tilde{p}_{klm} \hat{p}_{j|klm}^{(t)} \quad (2.3)$$

$$\hat{q}_{jk}^{(1,t)} = \frac{\sum_{l,m} \tilde{p}_{klm} \hat{p}_{j|klm}^{(t)}}{\hat{\pi}_j^{(t)}} \quad (2.4)$$

$$\hat{q}_{jl}^{(2,t)} = \frac{\sum_{k,m} \tilde{p}_{klm} \hat{p}_{j|klm}^{(t)}}{\hat{\pi}_j^{(t)}} \quad (2.5)$$

$$\hat{q}_{jm}^{(3,t)} = \frac{\sum_{k,l} \tilde{p}_{klm} \hat{p}_{j|klm}^{(t)}}{\hat{\pi}_j^{(t)}}, \quad (2.6)$$

The numerator in Equation (2.4) - (2.6) indicates the estimated probability that the level of rating from the  $s$ th rater ( $s = 1, 2, 3$ ) is  $k$  (or  $l, m$ ) and the level of the underlying truth is  $j$ . For example, the numerator in Equation (2.4) can be written as  $Pr(V^{(1)} = k, W = j)$ , that is, the probability that the level of rating from the first rater is  $k$  and the level of the underlying truth is  $j$ . Without loss of generality,  $\sum_{j=1}^J \hat{\pi}_j = 1$ ,  $\sum_{k=1}^K \hat{q}_{jk}^{(s)} = 1$ ,  $\hat{\pi}_j \geq 0$ ,  $\hat{q}_{jk}^{(s)} \geq 0$ ,  $j = 1, 2, \dots, J; s = 1, 2, \dots, S$ . These steps will be repeated until a certain convergence criteria is met. Goodman's iterative proportional fitting algorithm is straightforward and easy to be applied for latent class models with  $m$ -class observed variables. The iterative algorithms proposed by McHugh [17] and Goodman [11] require relatively few iterations to converge, but convergence is not guaranteed. Dawid and Skene [7] proposed the EM algorithm that was initially proposed by Dempster et al. [8] in the wider context of missing data to estimate classification rates and prevalences in latent class models with regarding the unknown truth as missing. We cover the basic aspects of the EM algorithm and discuss more details in the application of EM algorithm to a latent class model in the next section 2.1.2.

While the iterative procedures for obtaining maximum likelihood estimates have been used for latent class models, no procedure can guarantee of finding a global maxima. When the model parameters are not fully identifiable but locally identifiable, multiple sets of parameter estimates optimize the likelihood function. Goodman [11] extended the sufficient

conditions for the local identifiability of latent class models with dichotomously observed variables discussed by McHugh [17] to polytomous observed variables. The method to check local identifiability is relied on the Fisher information matrix, which is the matrix of second partial derivatives of the logarithm of complete-data likelihood function with respect to the model parameters. If the rank of this matrix is less than full rank, the latent class model is not even locally identifiable. We discuss more details in Section 2.1.3.

### 2.1.2 EM algorithm for the latent class model (2.1)

The Expectation-Maximization (EM) algorithm is an iterative procedure to compute maximum likelihood estimates in a model including quantities which can be viewed as missing data [8]. This iterative algorithm consists of two steps at each iteration: E-step and M-step. With current estimates for model parameters  $\theta \in \Theta$ , the EM algorithm calculates the conditional expectation of the logarithm of the complete-data likelihood given observed data (E-step) and maximizes the conditional expectation in terms of the model parameter (M-step).

We denote the complete data as  $Z = (X, Y)$ , where  $X$  is a matrix of incomplete or unobserved data and  $Y$  is a matrix of fully observed data. At step  $t$ ,  $t = 1, 2, \dots$ , the conditional expectations of the logarithm of the complete-data log-likelihood function given the observed data and the current estimates are computed in E-step.

**E-step:** Calculate

$$Q(\theta; \theta^{(t)}) = E_{\theta^{(t)}}[l_{com}(\theta; X, Y)|Y],$$

where  $l_{com}(\theta; X, Y) = \log L_{com}(\theta; X, Y) = \log f(X, Y; \theta)$ .

If the distribution of  $(X, Y)$  follows an exponential family such as: Gaussian, Binomial, Multinomial, Exponential, etc.,  $l_{com}(\theta; X, Y)$  is a linear function of the sufficient statistics. The E-step is equivalent to calculating the conditional expectation of the sufficient statistics given the observed data and the current estimates of  $\theta$ ,  $\theta^{(t)}$ .

**M-step:** Find  $\theta^{(t+1)}$  such that

$$\theta^{(t+1)} = \operatorname{argmax}_{\theta} Q(\theta; \theta^{(t)})$$

The E- and M-steps are repeated until the sequence  $\{\theta^{(t)}\}$  converges.

Dawid and Skene [7] presented the application of the EM algorithm to latent class models. Since a latent class variable represents the unknown or unobserved data, the EM algorithm can be applied in a straightforward manner. For the latent class model in Equation (2.1), the complete data are  $D_{com} = \{W_i, V_i^{(s)}\}$ , where  $W_i, i = 1, 2, \dots, n$  is the underlying truth, which cannot be observed for each subject  $i$  and  $V_i^{(s)}, i = 1, 2, \dots, n; s = 1, 2, \dots, S$  is the ratings evaluated by the  $s$ th rater for a subject  $i$ . Hence, the observed data are  $D_{obs} = \{V_i^{(s)}, i = 1, 2, \dots, n; s = 1, 2, \dots, S$ , where  $n$  and  $S$  are the total number of subjects and raters, respectively. The set of parameters in the equation (2.1) is  $\Theta = \{\pi_j, q_{jk}^{(s)}, j = 1, 2, \dots, J; k = 1, 2, \dots, K; s = 1, 2, \dots, S$ , where  $J$  and  $K$  are the number of categories in the latent class variable  $W_i$  and the ratings  $V_i^{(s)}$ , respectively. Here,  $\pi_j$  is the prevalence of the underlying truth at the level  $j$  and  $q_{jk}^{(s)}$  is the probability that the level of rating from the  $s$ th rater is  $k$  given the underlying truth at the level  $j$ . The complete-data log-likelihood function with  $D_{com}$  is

$$l_{com}(\boldsymbol{\pi}, \mathbf{q}) = \sum_{i=1}^n \left[ \sum_{j=1}^J I(W_i = j) \left\{ \log(\pi_j) + \sum_{s=1}^S \sum_{k=1}^K I(V_i^{(s)} = k) \log(q_{jk}^{(s)}) \right\} \right].$$

In the E-step, the conditional expectation of the complete-data log-likelihood  $E[l_{com}(\boldsymbol{\pi}, \mathbf{q})|\theta^{(t)}]$  is computed with the current estimates  $\theta^{(t)} = (\boldsymbol{\pi}^{(t)}, \mathbf{q}^{(t)})$ . The complete-data log-likelihood function is a linear function of  $I(W_i = j)$ , which is the indicator function of the latent truth for the  $i$ th subject at level  $j$ . With the observed data  $D_{obs} = \{V_i^{(s)}, i = 1, 2, \dots, n; s = 1, 2, \dots, S$ , the conditional expectation of the form  $E[h(\tilde{W}_i)|D_{obs}; \theta^{(t)}]$  can be written as

$$E[I(W_i = j)|D_{obs}; \theta^{(t)}] = Pr[W_i = j|D_{obs}, \theta^{(t)}] = \frac{\pi_j^{(t)} [\prod_{s=1}^S \prod_{k=1}^K (q_{jk}^{(s,t)})^{I(v_i^{(s)}=k)}]}{\sum_{j=1}^J \pi_j^{(t)} [\prod_{s=1}^S \prod_{k=1}^K (q_{jk}^{(s,t)})^{I(v_i^{(s)}=k)}]},$$

where  $i = 1, 2, \dots, n; j = 1, 2, \dots, J; k = 1, 2, \dots, K; s = 1, 2, \dots, S$ .

In the M-step, the conditional expectation of the complete-data log-likelihood function is maximized with replacing the underlying truth  $W_i$  by the sufficient statistics  $E[I(W_i = j)|D_{obs}; \theta^{(t)}]$ . Then the maximum likelihood estimates for the latent class model (2.1) are updated by

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n E[I(W_i = j)|D_{obs}; \theta^{(t)}]}{n} \quad (2.7)$$

$$q_{jk}^{(s,t+1)} = \frac{\sum_{i=1}^n I(V_i^{(s)} = k) E[I(W_i = j)|D_{obs}; \theta^{(t)}]}{\sum_{i=1}^n E[I(W_i = j)|D_{obs}; \theta^{(t)}]}. \quad (2.8)$$

If there is no closed-form of solutions for model parameters, the additional iterative procedure, such as Newton-Raphson algorithm may be necessary. The E- and M-steps are repeated until the difference in the complete-data log-likelihood between  $\theta^{(t)}$  and  $\theta^{(t+1)}$  is arbitrarily small.

Denote the complete data as  $(\mathbf{V}_i, W_i)$  for a subject  $i$ , where  $\mathbf{V}_i = (V_i^{(1)}, V_i^{(2)}, V_i^{(3)})$  are the tumor grade readings from three pathologists and  $W_i$  is the unknown true tumor grade. The counts of tumor grade readings from three pathologists in Table 3 can be denoted by  $n_{klm} = \# \{i : V_i^{(1)} = k, V_i^{(2)} = l, V_i^{(3)} = m\}$ ,  $k, l, m = 1, 2, 3$ . Then, the observed data can be written as  $n_{+klm} = \sum_{j=1}^3 n_{jklm}$ , where  $n_{jklm} = \# \{i : V_i^{(1)} = k, V_i^{(2)} = l, V_i^{(3)} = m, W_i = j\}$ ,  $j, k, l, m = 1, 2, 3$ .

In E-step at  $t$ th iteration, the conditional expectations with the current estimates  $\theta^{(t)} = \{\pi_j^{(t)}, q_{jk}^{(1,t)}, q_{jl}^{(2,t)}, q_{jm}^{(3,t)}\}$ ,  $j, k, l, m = 1, 2, 3$ ;  $t = 1, 2, \dots$  can be calculated by

$$\begin{aligned} E[n_{jklm}|n_{+klm}, j, k, l, m = 1, 2, 3; \theta^{(t)}] &= n_{+klm} Pr[W = j|V^{(1)} = k, V^{(2)} = l, V^{(3)} = m] \\ &= n_{+klm} \frac{Pr[W = j, V^{(1)} = k, V^{(2)} = l, V^{(3)} = m]}{Pr[V^{(1)} = k, V^{(2)} = l, V^{(3)} = m]}. \end{aligned}$$

Under the conditional independence assumption,  $Pr[V^{(1)} = k, V^{(2)} = l, V^{(3)} = m]$  can be written as

$$\begin{aligned} Pr[V^{(1)} = k, V^{(2)} = l, V^{(3)} = m] &= Pr[V^{(1)} = k, V^{(2)} = l, V^{(3)} = m|W = j] Pr[W = j] \\ &= q_{jk}^{(1,t)} q_{jl}^{(2,t)} q_{jm}^{(3,t)} \pi_j^{(t)}. \end{aligned}$$

Hence,  $E[n_{jklm}|n_{+klm}, j, k, l, m = 1, 2, 3; \theta^{(t)}]$  is

$$E[n_{jklm}|n_{+klm}, j, k, l, m = 1, 2, 3; \theta^{(t)}] = n_{+klm} \frac{q_{jk}^{(1,t)} q_{jl}^{(2,t)} q_{jm}^{(3,t)} \pi_j^{(t)}}{\sum_{j=1}^3 q_{jk}^{(1,t)} q_{jl}^{(2,t)} q_{jm}^{(3,t)} \pi_j^{(t)}}.$$

In M-step, the maximum likelihood estimates for the parameters  $(\boldsymbol{\pi}, \mathbf{q})$  are updated by

$$\begin{aligned} \pi_j^{(t+1)} &= \frac{\sum_{k,l,m} E[n_{jklm}|n_{+klm}; \theta^{(t)}]}{n}, \\ q_{jk}^{(1,t+1)} &= \frac{\sum_{l,m} E[n_{jklm}|n_{+klm}; \theta^{(t)}]}{\sum_{k,l,m} n_{jklm}^{(t)}}, \\ q_{jl}^{(2,t+1)} &= \frac{\sum_{k,m} E[n_{jklm}|n_{+klm}; \theta^{(t)}]}{\sum_{k,l,m} n_{jklm}^{(t)}}, \\ q_{jm}^{(3,t+1)} &= \frac{\sum_{k,l} E[n_{jklm}|n_{+klm}; \theta^{(t)}]}{\sum_{k,l,m} n_{jklm}^{(t)}}, \end{aligned}$$

where  $j, k, l, m = 1, 2, 3$  and  $n = n_{++++}$ .

### 2.1.3 Local identifiability

In statistical models, the model parameters are fully identifiable if different sets of parameter estimates strictly correspond to different probability distributions. By definition, the parameter  $\phi$  of a distribution family  $F_\phi(\cdot)$  is fully identifiable if  $F_\phi(\cdot)$  and  $F_{\phi'}(\cdot)$  are different cumulative distribution functions for any  $\phi \neq \phi'$ . It is a rule of thumb that the global identifiability can be guaranteed if the number of parameters in the model does not exceed the degree of freedom in the data. However, a model can still suffer from the lack of global identifiability despite of a sufficient degree of freedom. Especially, latent class models cannot be fully identifiable due to its ‘‘label-switching’’ phenomenon, which means that the  $J!$  permuted sets of parameter estimates, where  $J$  is the number of classes, can reach the maximum of the log-likelihood function.

For example, when we estimate parameters for a latent class model with a 3-class latent variable, total  $3! = 6$  permuted sets of estimates can generate the same value of the complete-data likelihood. Given  $\phi = \{\pi_1, \pi_2, \pi_3, q_{1k}^{(s)}, q_{2k}^{(s)}, q_{3k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$ ,  $\sum_{j=1}^3 \pi_j =$

$\sum_{k=1}^3 q_{jk}^{(s)} = 1$ , one of the permuted sets  $\phi^1 = \{\pi_1, (1 - \pi_1 - \pi_2), \pi_2, q_{1k}^{(s)}, (1 - q_{1k}^{(s)} - q_{2k}^{(s)}), q_{2k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$  can generate the same likelihood value,  $\log L_{com}(\phi; D_{obs}) = \log L_{com}(\phi^1; D_{obs})$ . With this “label-switching” phenomenon, global identification for latent class models cannot be achievable. However, under certain conditions, local identification for latent class models can be obtained even global identification is not guaranteed. The general definition of local identifiability for a distribution  $F_\phi$  at the parameter  $\phi$  is that there exists some neighborhood  $\Phi$  of  $\phi$  such that  $F_{\phi'} = F_\phi$  if and only if  $\phi' = \phi$ , for all  $\phi, \phi' \in \Phi$ . The lack of full identification does not hinder statistical inferences, but the parameters cannot be uniquely estimated [2]. Therefore, the verification of model identifiability must be considered before making statistical inferences for latent class models.

McHugh [17] showed the sufficient conditions for the local identifiability of latent class models (2.1) with binary observed variables and a 2-class latent variable. The specific conditions are described as below [17]:

- (i)  $2^s \geq J - 1 + SJ$
- (ii)  $\sum_{k=1}^2 \{\sum_{j=1}^2 \pi_j \prod_{s=1}^S q_{jk}^{(s)}\} = 1$
- (iii)  $\sum_{k=1}^2 \pi_j \prod_{s=1}^S q_{jk}^{(s)} > 0$ , for all  $j, k, s$ .
- (iv)  $\sum_{k=1}^2 \pi_j \prod_{s=1}^S q_{jk}^{(s)}$ , for all  $j, k, s$  are continuous functions of  $\pi_j, q_{jk}^{(s)}$  and its continuous first and second derivatives are existed.
- (v) There are at least  $(2S + 1)$  of the expressions  $\sum_{k=1}^2 \pi_j \prod_{s=1}^S q_{jk}^{(s)}$ , which are linearly independent.

If all of five conditions are met, the model parameters are locally identifiable. For example, if we have three observed binary variables, the degree of freedom is 7 df. The equation (2.1) with a 2-class latent variable requires to estimate seven parameters (1 prevalence and 3 classification rates for each latent class). In this case, the number of estimated parameters is equal to the degree of freedom, so the equation (2.1) with a 2-class latent variable and three observed variables is locally identifiable.

Goodman [11] extended the sufficient conditions showed by McHugh (1956) to polytomous observed variables with a  $m$ -class latent variable,  $m \geq 2$ . In the paper, the author pointed out that if the rank of the Jacobian matrix of the latent class model is equal to the

number of estimated parameters in the model, then the model can be locally identifiable [11]. The sufficient conditions for the local identifiability in McHugh (1956) can be generalized as below [11, 17]:

- (i)  $J^s \geq K + SK - 1$
- (ii) Let  $h(\mathbf{v}) = \sum_{j=1}^J \pi_j \left\{ \prod_{s=1}^S q_{jv^{(s)}} \right\}$  with  $\mathbf{v} = (v^{(1)}, \dots, v^{(S)})$ , then  $\sum_{\mathbf{v}} h(\mathbf{v}) = \sum_{\mathbf{v}} pr[\mathbf{V} = \mathbf{v}] = 1$ .
- (iii)  $h(\mathbf{v}) = pr[\mathbf{V} = \mathbf{v}] > 0$ , for all  $\mathbf{v} = (v^{(1)}, \dots, v^{(S)})$ .
- (iv) Let  $\pi = (\pi_1, \pi_2, \dots, \pi_J)$ ,  $\mathbf{q} = \left\{ q_{jk}^{(s)}, j = 1, 2, \dots, J; k = 1, 2, \dots, K; s = 1, 2, \dots, S \right\}$ ,  $h(\mathbf{v})$ s are continuous functions of  $(\pi, \mathbf{q})$  and have continuous first and second derivatives.
- (v) There exist  $\{\mathbf{v}_1, \dots, \mathbf{v}_T\}$  with  $T \geq K - 1 + SK$ , such that  $\{h(\mathbf{v}_1), \dots, h(\mathbf{v}_T)\}$  are linearly independent as functions of the model parameters.

When the model parameters are non-identifiable due to the violation of the five conditions, restrictions on the parameters enable the model to be locally identifiable. Goodman [11] suggested the restricted polytomous latent class models which are brought by imposing various restrictions on the model parameters.

With the equation (2.1) and assuming four observed variables, one of the scenarios described in Goodman (1974) is below:

- Models in the  $m$  latent classes can be partitioned into  $\alpha$  mutually exclusive and exhaustive subsets  $F_1^A, \dots, F_\alpha^A$ , where  $\alpha \leq m$  and/or into  $\beta$  mutually exclusive and exhaustive subsets  $F_1^B, \dots, F_\beta^B$ , where  $\beta \leq m$  and/or into  $\gamma$  mutually exclusive and exhaustive subsets  $F_1^C, \dots, F_\gamma^C$ , where  $\gamma \leq m$ , and/or into  $\eta$  mutually exclusive and exhaustive subsets  $F_1^D, \dots, F_\eta^D$ , where  $\eta \leq m$  such that  $q_{jk^A}^{(1)} = q_{jk'^A}^{(1)}(k, k' \in F_1^A)$ ,  $q_{jk^B}^{(2)} = q_{jk'^B}^{(2)}(k, k' \in F_1^B)$ ,  $q_{jk^C}^{(3)} = q_{jk'^C}^{(3)}(k, k' \in F_1^C)$ ,  $q_{jk^D}^{(4)} = q_{jk'^D}^{(4)}(k, k' \in F_1^D)$ .

With the above restriction, the number of estimated parameters will be reduced from  $J - 1 + 4J(K - 1)$  to  $J - 1 + (K - 1)\alpha + (K - 1)\beta + (K - 1)\gamma + (K - 1)\eta$ . Hence, if the rank of the Jacobian matrix with the restrictions is equal to the number of estimated parameter  $J - 1 + (K - 1)\alpha + (K - 1)\beta + (K - 1)\gamma + (K - 1)\eta$ , the parameters in the restricted latent class model will be locally identifiable [11].

Table 4: Estimated prevalences and classification rates from the method in Dawid and Skene (1974)

True grade	$\pi$	Rater#1			Rater#2			Rater#3		
		W	M	P	W	M	P	W	M	P
W	0.26	0.47	0.52	0.01	0.87	0.13	0	0.67	0.33	0
M	0.51	0.08	0.86	0.06	0.20	0.70	0.09	0.04	0.91	0.05
P	0.24	0	0.19	0.81	0.02	0.23	0.74	0.01	0.47	0.52

Under the sufficient conditions for local identification, the parameters in the equation (2.1) can be uniquely identifiable within the neighborhood of the true values of these parameters and there exists only one set of maximum likelihood estimates within the neighborhood [11, 17]. Due to the lack of global identifiability,  $J!$  sets of estimates, which are the permutations on the set of  $\{1, 2, \dots, J\}$ , reach the maximum of the likelihood function. To deal with the local identification, Dawid and Skene [7] suggested choosing one set of estimates that leads to the largest conditional probabilities for all observed variables given a certain level of latent class. This guideline is applied to the previously mentioned sub-study of the NSABP B-14 trial. The maximum likelihood estimates for the prevalence of tumor grade and the classification rates of the three pathologists are shown in Table 4. Among  $3!$  sets of estimates, one set of estimates which yields the best classification rates for all raters is chosen. One of cautions about the strategy proposed by Dawid and Skene is that if the classification rates in two adjacent categories within a rater are closed to each other, it is hard to select a set of estimates. In Table 4, we can see that the classification rates in two adjacent categories for Rater #1 and #3 are close to each other. Therefore, more options should be considered to handle the issue of local identifiability in latent class models. In Chapter 3, we discuss the proposed joint modeling and how it does handle the local identifiability of the parameters in a latent class model.

### 2.1.4 Conditional independence in latent class models

The critical assumption on basic latent class models is that observed variables or ratings are mutually independent given the latent truth. This is known as conditional independence or local independence which can be written as  $Pr[V_i^{(s)} = k | W_i = j] \perp Pr[V_i^{(s)} = k' | W_i = j]$ , for all  $k \neq k'$  in the equation (2.1). Sometimes, observed variables, such as assessments for related symptoms to determine a disease status, can be correlated within a true disease status. In this case, latent class models which fail to account for the correlation structure among observed variables will lead to biased estimates. There are various strategies to test whether the assumption of conditional independence in a latent class model is met.

#### Method 1: Goodness of fit test statistic

The goodness-of-fit test statistic can be used to detect the violation of conditional independence in a latent class model. The conventional goodness-of-fit test statistics, such as Pearson chi-square statistic, the likelihood ratio statistic, and empirical likelihood ratio statistics known as the Cressie-Read power-divergence statistic (Cressie and Read, 1984), compare the sets of observed frequencies with the sets of expected frequencies and asymptotically follow chi-square distribution [4,6,29]. If the assumption is violated, the goodness-of-fit test statistic tends to have a large value. The chi-square statistics is defined as

$$\chi_{df}^2 = \sum_r \frac{(O_r - E_r)^2}{E_r},$$

where  $O_r$  is the observed frequency in the contingency table for the observed data or ratings and  $E_r$  is the corresponding quantity estimated by the latent class model.  $R$  is the total number of response patterns defined by all possible combinations of ratings, such as  $\{V^{(1)} = k, V^{(2)} = l, V^{(3)} = m, k, l, m = 1, 2, 3\}$  in Table 3. The degree of freedom  $df$  can be computed by  $df = R - p - 1$ , where  $p$  is the number of model parameters. If the p-value of the test statistic is low (conventionally, less than 0.05), the model is said to not fit and so the assumption of conditional independence is violated.

The likelihood ratio statistic is defined as

$$G^2 = 2 \sum_r^R O_r \log \left( \frac{O_r}{E_r} \right).$$

The Cressie-Read power-divergence statistic is defined as [33]

$$CR = 1.8 \sum_r^R O_r \left[ \left( \frac{O_r}{E_r} \right)^{\frac{2}{3}} - 1 \right].$$

The degree of freedom of two statistics are the same as the chi-square test [33]. All three test statistics are reasonable for situation where the sample size is large enough, and the number of observed variables is small enough. However, when data are sparse and the expected frequencies are very low, the chi-square approximation is not acceptable. In this case, the parametric bootstrapping goodness-of-fit statistic proposed by Aitkin et al.(1981) can be considered as an alternative method of the three goodness-of-fit tests [1, 29, 33].

## Method 2: Graphical procedures

In addition to goodness-of-fit tests with a chi-square approximation, a graphical approach can be considered to check the violation of conditional independence. Qu et al. [21] proposed the pairwise correlation residual plot for latent class models with binary observed variables and a 2-class latent variable. With  $S$  binary observed variables, total  $S(S - 1)/2$  pairwise correlations will be calculated. The correlation between observed variables ( $V_i^{(s)} = 1, V_i^{(s')} = 1$ ), where can be computed by [21]

$$corr_{ss'} = \frac{P(V_i^{(s)} = 1, V_i^{(s')} = 1) - \mu_s \mu_{s'}}{\sqrt{\mu_s(1 - \mu_s)\mu_{s'}(1 - \mu_{s'})}},$$

where  $\mu_s = P(V_i^{(s)} = 1)$ .

The observed correlation,  $\mu_s$  and  $P(V_i^{(s)} = 1, V_i^{(s')} = 1)$  can be computed by

$$\begin{aligned}\mu_s &= \frac{1}{n} \sum_{i=1}^n I(V_i^{(s)} = 1) \\ P(V_i^{(s)} = 1, V_i^{(s')} = 1) &= \frac{1}{n} \sum_{i=1}^n I(V_i^{(s)} = 1)I(V_i^{(s')} = 1).\end{aligned}$$

The expected correlation,  $E(\mu_s)$  and the estimated probability of  $(V_i^{(s)} = 1, V_i^{(s')} = 1)$  can be computed by

$$\begin{aligned}\mu_s &= \pi_0 q_{00}^{(s)} + \pi_1 q_{11}^{(s)} \\ P(V_i^{(s)} = 1, V_i^{(s')} = 1) &= \sum_{j=0}^1 \pi_j \prod_{s=1}^S q_{j1}^{(s)},\end{aligned}$$

where  $\pi_j, j = 0, 1$  is the estimated prevalence of the underlying truth at level  $j$ , and  $q_{j1}^{(s)}, j = 0, 1; s = 1, 2, \dots, S$  is the estimated classification rate that the level of the  $s$ th observed variable is 1 given the level of underlying truth is 1. With the difference between the observed and expected correlations under the assumption of conditional independence, the pairwise correlation plot is obtained by plotting  $S(S-1)/2$  pairwise correlation coefficients. If the assumption is valid, all residual correlations should be randomly distributed around zero [21].

Another way of graphical procedures to detect the lack of fit is proposed by Garrett and Zeger [10], which is known as the Log-Odds Ratio Check (LORC) plot. First, the observed and expected pairwise log-odds ratios from two-way cross-classification frequency tables for all possible  $\binom{S}{2}$  combinations of observed variables are calculated with a continuity correction of 0.5 in the case of zero cells. Then, 95% confidence intervals of the observed pairwise log-odds ratios are constructed with the standard errors of the observed log-odds ratio  $\sqrt{(1/a + 1/b + 1/c + 1/d)}$ , where  $a, b, c, d$  denote the four frequencies of a two-way table. By plotting 95% confidence intervals for the observed log-odds ratios and the expected log-odds ratios on the same graph, the differences between the observed and expected log-odds ratios can be detected. If the assumption of conditional independence is valid, all relative magnitudes in differences between the observed and expected log-odds ratios should be small enough [10].

When the conditional independence between observed variables does not be guaranteed in latent class models, the parameter estimates will be biased. The simulation studies for examining the robustness of latent class models under the violation of conditional independence and exploring the performance of the methods for checking conditional independence were presented by Subtil et al. [29]. The authors considered a latent class model with a 2-class latent variable and four binary observed variables. The data were simulated with admitting conditional dependence between first and second observable variables  $(V_i^{(1)}, V_i^{(2)})$  within the latent class of  $W_i = 1$ . The model for generating data with conditional dependence is formulated by [29]

$$P(\mathbf{V} = \mathbf{v}) = \pi_1 \{ p_{11}^{v_1} (1 - p_{11})^{1-v_1} p_{21}^{v_2} (1 - p_{21})^{1-v_2} + (-1)^{v_1-v_2} \sigma_{12|W=1} \} \prod_{s=1}^S \{ p_{s1}^{v_s} (1 - p_{s1})^{1-v_s} \} (1 - \pi_1) \prod_{s=1}^S \{ p_{s0}^{v_s} (1 - p_{s0})^{1-v_s} \},$$

where  $\mathbf{V} = (V^{(1)}, \dots, V^{(s)})'$ ;  $\mathbf{v} = (v^{(1)}, \dots, v^{(s)})'$ ;  $v_s = 0, 1, s = 1, \dots, S$ ; and  $\sigma_{12|W=1} = cov(V^{(1)}, V^{(2)}|W = 1)$ . The degree of conditional dependence can be defined by  $\sigma_{12|W=1}$ . In Subtil et al. [29], the four types of goodness-of-fit test statistic (likelihood ratio, chi-square, CR, parametric bootstrap), the pairwise correlation residual plot, and the LORC plot were considered with four scenarios: conditional independence, conditional dependence ranging from weak to strong correlation. The performance of each method was measured in the proportions of detection of the violation of conditional independence. The simulation results in Subtil et al. (2012) pointed out that all methods showed similar performance under conditional independence, however, the correlation residual plot and the LORC plot showed unsatisfactory performances under conditional dependence between  $(V_i^{(1)}, V_i^{(2)})$ . With strong correlation between  $(V_i^{(1)}, V_i^{(2)})$ , the LORC plot even yielded a poor performance in terms of identifying the true pair of observed variables having conditional dependence. For example, the correlation residual plot and the LORC plot identified the wrong pair of observed variables, such as  $(V_i^{(3)}, V_i^{(4)})$ , instead of the true pair  $(V_i^{(1)}, V_i^{(2)})$  forced to admit conditional dependence.

Since the expected pairwise correlations and the expected log-odds ratios can be impacted by biased estimates, these two plots may not be able to show better performance in detecting the violation of conditional independence and identifying the true pair of observed variables having conditional dependence. Accordingly, it may be problematic to use the correlation residual plot or the LORC plot only to check the assumption of conditional independence in a certain case with strong association between observed variables [29].

## 2.2 JOINT MODELING OF TIME-TO-EVENT OUTCOME AND COVARIATES WITH MEASUREMENT ERRORS

Joint modeling analyses of time-to-event data and covariates with measurement errors have been widely studied. Due to random errors in a covariate or the presence of misclassification, the true status of a covariate may not be achievable [15, 16, 35]. Wulfsohn and Tsiatis [35] proposed the joint model of time-to-event data and a longitudinal continuous covariate with measurement errors. For the time-to-event data, the Cox proportional hazards model with an unspecified baseline hazard was considered. As an alternative of a two-stage modeling for longitudinal covariates with measurement errors, the authors incorporated a linear growth curve model with random intercept and random slope to account for the heterogeneity in the CD4 counts among patients caused by different progresses on HIV disease [35]. Larsen [15, 16] proposed the joint analysis of time-to-event data and a latent covariate with adapting the estimation procedures in Wulfsohn and Tsiatis (1997). The Cox proportional hazards model was also used to model the hazard functions for time-to-event with a latent covariate in the regression component. In Larsen [15, 16], a generalized logit model was considered as a sub-model for a latent class covariate [15] and a two-parameter logistic item response model was considered as a sub-model for a latent continuous covariate [16].

In Wulfsohn and Tsiatis [35], a covariate  $\mathbf{Z}_i$  is measured over time and the measurement time  $m_i$  could be different for each subject  $i$ . The observable data for each subject  $i$  is denoted as  $(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i)$ . Here, the observable event time is  $X_i = \min(T_i, C_i)$ , where  $T_i$  and  $C_i$  are the survival time and a potential right censoring time, respectively. The failure

indicator is denoted as  $\Delta_i = I(T_i \leq C_i)$ .  $\mathbf{t}_i = (t_{ij} : t_{ij} \leq X_i)$  is the time for a measurement, where  $t_{ij}, j = 1, 2, \dots, m_i$  is the time from randomization for subject  $i$ .  $\mathbf{Z}_i = (Z_{ij} : t_{ij} \leq X_i)$  is the value of covariate at time  $t_{ij}$ . Due to the heterogeneity in the CD4 counts among patients, the observed CD4 counts  $Z_{ij}$  at time  $t_{ij}$  is modeled by [35]

$$Z_{ij} = \theta_{0i} + \theta_{1i}t_{ij} + e_{ij}, \quad (2.9)$$

where  $e_{ij} \sim N(0, \sigma_e^2)$ ,  $e_{ij} \perp e_{ij'}, j \neq j'$ , and error  $e_{ij}$  is independent of random intercept and slope,  $\theta_{0i}$  and  $\theta_{1i}$ , respectively. The distribution of random intercept and slope is assumed to be a bivariate normal distribution [35].

Assuming that the true covariate value  $Z_{ij}$  can be derived by using the equation (2.9), which is the growth curve model with random effects, the part of time-to-event can be modeled as

$$\lambda(t|\theta_i, \mathbf{Z}_i, \mathbf{t}_i) = \lambda_0(t) \exp\{\beta(\theta_{0i} + \theta_{1i}t)\} \quad (2.10)$$

With the observed data for each subject  $i$ ,  $(X_i, \Delta_i, \mathbf{Z}_i, \mathbf{t}_i)$ , the observed data likelihood is given by [35]

$$\prod_{i=1}^n \left[ \int_{-\infty}^{\infty} \left\{ \prod_{j=1}^{m_i} f(z_{ij}|\theta_i, \sigma_e^2) \right\} f(\theta_i|\boldsymbol{\theta}, \mathbf{V}) f(X_i, \Delta_i|\theta_i, \boldsymbol{\lambda}_0, \beta) d\theta_i \right], \quad (2.11)$$

where

$$f(z_{ij}|\theta_i, \sigma_e^2) = (2\pi\sigma_e^2)^{-1/2} \exp\{-(z_{ij} - \theta_{0i} - \theta_{1i}t_{ij})^2/2\sigma_e^2\},$$

$$f(\theta_i|\boldsymbol{\theta}, \mathbf{V}) = (2\pi|\mathbf{V}|)^{-1/2} \exp\{-(\boldsymbol{\theta}_i - \boldsymbol{\theta})' \mathbf{V}^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\theta})/2\},$$

$$f(X_i, \Delta_i|\theta_i, \boldsymbol{\lambda}_0, \beta) = [\lambda_0(X_i) \exp\{\beta(\theta_{0i} + \theta_{1i}X_i)\}]^{\Delta_i} \exp\left[-\int_0^{X_i} \lambda_0(\mu) \exp\{\beta(\theta_{0i} + \theta_{1i}\mu)\} d\mu\right].$$

In order to estimate the parameters with unknown random effects  $\boldsymbol{\theta}$ , Wulfsohn and Tsiatis [35] adapted the EM algorithm introduced by Dempster et al. [8]. Since there are no closed-form of solutions to estimate the baseline hazard  $\lambda_0$  and the parameter  $\beta$  in the joint model,  $\lambda_0$  and  $\beta$  are updated via a one-step Newton-Raphson algorithm in M-step [35]. For variance estimation of the Cox model parameters, the authors employed the profile score with using the restricted maximum likelihood estimates [35].

In Larsen [15],  $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})^t$ , is a random vector of  $J$  binary indicators for the  $i$ th subject. With assuming that the population consists of  $K$  sub-populations, a latent class variable  $C_i$ , which may represent the unobserved status of health or group membership, takes one of the values  $1, \dots, K$ . The probability of each binary indicator  $Y_{ij}$  would be  $y$  given the latent class  $C_i = c$  is modeled by [15]

$$Pr(Y_{ij} = y | C_i = c) = \pi_{cj}^y (1 - \pi_{cj})^{1-y}, y = 0, 1,$$

where  $\boldsymbol{\pi}_c = (\pi_{c1}, \dots, \pi_{cJ})^t, c = 1 \dots K$ . Under the conditional independence,  $Pr(\mathbf{Y}_i = \mathbf{y}_i | C_i = c) = \prod_{j=1}^J Pr(Y_{ij} = y | C_i = c)$ .

Larsen [15] included additional covariates to predict the latent class  $C_i$  for each subject  $i$ . To build the predictive model for the latent class  $C_i$  with additional covariates, a generalized logit model is considered [15]:

$$Pr(C_i = c | x_i) = \frac{\exp(\mathbf{x}_i \boldsymbol{\kappa}_c)}{\sum_{k=1}^K \exp(\mathbf{x}_i \boldsymbol{\kappa}_k)},$$

where  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$  is a vector of covariates for the  $i$ th subject. Then the joint distribution of  $(\mathbf{Y}_i, C_i)$  is defined by [15]

$$\begin{aligned} Pr(\mathbf{Y}_i = \mathbf{y}_i, C_i = c | \mathbf{x}_i) &= Pr(C_i = c | \mathbf{x}_i) \times Pr(\mathbf{Y}_i = \mathbf{y}_i | C_i = c, \mathbf{x}_i) \\ &= \frac{\exp(\mathbf{x}_i \boldsymbol{\kappa}_c)}{\sum_{k=1}^K \exp(\mathbf{x}_i \boldsymbol{\kappa}_k)} \times \left\{ \prod_{j=1}^J \pi_{cj}^{y_{ij}} (1 - \pi_{cj})^{1-y_{ij}} \right\}. \end{aligned}$$

For the sub-model for time-to-event data, Larsen [15] also used the Cox proportional hazards model. The observable event time is denoted by  $U_i = \min(T_i, V_i)$ , where  $T_i$  is the failure time and  $V_i$  is the possible censoring time. The event indicator is denoted by  $\Delta_i = I(T_i \leq V_i)$ . The density of the event time  $T_i$  is defined by [15]

$$P(t | \mathbf{z}_i, c_i) = \lambda_0(t) \exp(\mathbf{z}_i \boldsymbol{\beta} + \nu_{c_i}) \exp \{ -\Lambda_0(t) \exp(\mathbf{z}_i \boldsymbol{\beta} + \nu_{c_i}) \},$$

where  $\Lambda_0(t) = \int_0^t \lambda_0(s) ds$ .  $\nu$  represents the effect of  $c_i$  on the hazard with defining  $\nu_1 = 0$  for identification.

The time-to-event data can be modeled as [15]

$$Pr(\mu_i, \Delta_i | \mathbf{z}_i, c_i) \propto \lambda_0(\mu_i) \exp(\mathbf{z}_i \boldsymbol{\beta} + \nu_{c_i})^{\delta_i} \exp \{ -\Lambda_0(\mu) \exp(\mathbf{z}_i \boldsymbol{\beta} + \nu_{c_i}) \}.$$

Then the joint distribution of  $(U_i, \Delta_i, \mathbf{Y}_i, C_i)$  is defined by

$$Pr(\mu_i, \Delta_i, \mathbf{y}_i, c_i | \mathbf{x}_i, \mathbf{z}_i) = pr(c_i | \mathbf{x}_i) pr(\mathbf{y}_i | c_i) pr(\mu_i, \Delta_i | c_i, \mathbf{z}_i).$$

By integrating out  $C_i$ , the marginal distribution of the observable data,  $(U_i, \Delta_i, \mathbf{Y}_i)$  becomes [15]

$$\begin{aligned} Pr(\mu_i, \Delta_i, \mathbf{y}_i | \mathbf{x}_i, \mathbf{z}_i) &= \sum_{c=1}^K Pr(c | \mathbf{x}_i) Pr(\mathbf{y}_i | c) Pr(\mu_i, \Delta_i | c, \mathbf{z}_i) \\ &= \sum_{c=1}^K \left[ \frac{\exp(\mathbf{x}_i \kappa_c)}{\sum_{k=1}^K \exp(\mathbf{x}_i \kappa_k)} \times \left\{ \prod_{j=1}^J \pi_{c_j}^{y_{ij}} (1 - \pi_{c_j})^{1-y_{ij}} \right\} \right] \\ &\times \lambda_0(\mu_i) \exp(\mathbf{z}_i \beta + \nu_c)^{\Delta_i} \exp\{-\Lambda_0(\mu_i) \exp(\mathbf{z}_i \beta + \nu_c)\}. \end{aligned}$$

With the complete data  $(\mathbf{u}, \Delta, \mathbf{y}, \mathbf{c})$ , the complete-data log-likelihood function is defined as following [15]

$$\begin{aligned} l_{com}(\theta; \mathbf{u}, \Delta, \mathbf{y}, \mathbf{c}) &= \sum_{i=1}^N l_{i,com}(\theta; \mu_i, \Delta_i, y_i, c_i) \\ &= \mathbf{x}_i \kappa_{c_i} - \log \left\{ \sum_{k=1}^k \exp(\mathbf{x}_i \kappa_k) \right\} + \sum_{j=1}^J y_{ij} \log(\pi_{c_{ij}}) + (1 - y_{ij}) \log(1 - \pi_{c_{ij}}) \\ &+ \Delta_i [\log \lambda_0(\mu_i) + \mathbf{z}_i \beta + \nu_{c_i}] - \Lambda_0(\mu_i) \exp(\mathbf{z}_i \beta + \nu_{c_i}). \end{aligned}$$

Then, the observed data log-likelihood is given by [15]

$$l(\theta; \mathbf{u}, \Delta, \mathbf{y}) = \log \sum_{c \in \{1, \dots, K\}^N} \exp\{l_{com}(\theta; \mathbf{u}, \Delta, \mathbf{y}, \mathbf{c})\}.$$

For parameter estimation, Larsen [15] also used the EM algorithm with a one-step Newton-Raphson algorithm in the M-step, that is similar to the procedure in Wulfsohn and Tsiatis (1997).

### 2.3 SURVEY-WEIGHTED COX MODELS

The Cox proportional hazards model is commonly used to investigate the association between a set of covariates and the risk of disease recurrence. The Cox model parameters  $\boldsymbol{\beta}$  can be estimated via the partial likelihood function with independent observations. Denote the observed failure time and the censoring indicator as  $X_i = \min(T_i, C_i)$  and  $\delta_i = I(T_i \leq C_i)$ , where  $T_i$  and  $C_i$  are the event time and a possible censoring time, respectively. Time-varying covariates and the risk set at time  $x_i$  are denoted as  $Z_i(x_i)$  and  $R(x_i)$ , respectively. The risk set  $R(x_i)$  can be defined as the set of subjects available for the event at time  $x_i$ . Then, the partial likelihood function for the Cox proportional hazards model can be defined by

$$\prod_{i=1}^N \left[ \frac{\lambda_0(x_i) \exp\{z'_i(x_i)\boldsymbol{\beta}\}}{\sum_{j \in R(x_i)} \lambda_0(x_i) \exp\{z'_j(x_i)\boldsymbol{\beta}\}} \right]^{\delta_i}.$$

The estimates for  $\boldsymbol{\beta}$  can be obtained by determining  $\hat{\boldsymbol{\beta}}$  which maximize the partial likelihood score function so that

$$\sum_{i=1}^N \delta_i \left[ z'_i(x_i) - \frac{S^{(1)}(x_i, \boldsymbol{\beta})}{S^{(0)}(x_i, \boldsymbol{\beta})} \right] = 0, \quad (2.12)$$

where

$$\begin{aligned} S^{(0)}(x_i, \boldsymbol{\beta}) &= \frac{1}{N} \sum_{j \in R(x_i)} \exp\{z'_j(x_i)\boldsymbol{\beta}\} \\ S^{(1)}(x_i, \boldsymbol{\beta}) &= \frac{1}{N} \sum_{j \in R(x_i)} z_j(x_i) \exp\{z'_j(x_i)\boldsymbol{\beta}\}. \end{aligned}$$

However, when the sample has been drawn from a complex designed sampling, such as a stratified random sampling, the observations within a strata can be correlated. The partial likelihood function for the Cox model cannot take account for the design of sampling and so it can lead to misleading results. Moreover, if parameters for stratified sampling are related to the risk of disease recurrence, estimating  $\boldsymbol{\beta}$  without considering the survey design can result in incorrect inference.

Binder [3] proposed a method for fitting the partial likelihood function with the sampling weights  $\omega_i, i = 1, 2, \dots, n$ . We assume that the sample of size  $n$  is drawn from a finite population of size  $N$  via a survey design and the sampling weights  $\omega_i, i = 1, 2, \dots, n$  are scaled so that  $\sum_n \omega_i = 1$ . With the weights  $\omega_i$ , the partial likelihood score function for  $\beta$  can be defined by replacing the summations in the equation (2.12) with weighted sums [3]:

$$\sum_{i=1}^N \omega_i \delta_i \left[ z'_i(x_i) - \frac{S^{(1)}(x_i, \beta)}{S^{(0)}(x_i, \beta)} \right] = 0, \quad (2.13)$$

where

$$S^{(0)}(x_i, \beta) = \frac{1}{N} \sum_{j \in R(x_i)} \omega_j \exp\{z'_j(x_i)\beta\}$$

$$S^{(1)}(x_i, \beta) = \frac{1}{N} \sum_{j \in R(x_i)} \omega_j z_j(x_i) \exp\{z'_j(x_i)\beta\}.$$

The survey-weighted Cox model can be fitted by using the R package `survey`. The survey design should be specified with the R function `svydesign` before fitting the survey-weighted Cox model by using the R function `svycoxph`.

**Step 1:** Specify the survey design with the R function `svydesign`.

```
grade.design <- svydesign(data=, ids=, strata=, variables=, weight=)
```

- `data`: Call data frame.
- `id`: Specify the level of cluster.  $\sim 1$  or  $\sim 0$  is for no cluster.
- `strata`: Specify the strata. `NULL` is for no strata.
- `variables`: Specify the variables measured in the survey.
- `weights`: Specify the sampling weights.

**Step 2:** Fit the survey-weighted Cox model with the R function `svycoxph`.

```
grade.model <- svycoxph(Model, design=)
```

- `Model`: Specify a Cox model. Use the same model statement for `coxph`.
- `design`: Specify the survey design which is defined with the R function `svydesign`.

Except for specifying the survey design, the outputs are similar to fitting a Cox model with the R function `coxph`. The details can be found in the document for R package `survey` '<http://cran.r-project.org/web/packages/survey/survey.pdf>'.

### 3.0 PROPOSED JOINT ANALYSIS OF TIME-TO-EVENT DATA AND DISCRETE DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD

#### 3.1 A NEW JOINT MODELING APPROACH

We propose a new joint modeling approach for time-to-event data and multiple ratings with an unknown truth. For each subject  $i$ , suppose multiple independent ratings on a discrete diagnostic test and time-to-event data are observed. Typically, such a diagnostic test is associated with the risk of the event in a known trend based on prior biological knowledge. Let  $T_i$  and  $C_i$  be the event time and a right censoring time on subject  $i$ . The observed event time and the censoring indicator are denoted by  $X_i = \min(T_i, C_i)$  and  $\Delta_i = I\{T_i \leq C_i\}$ , which equals to 1 if  $T_i \leq C_i$  and 0 otherwise, respectively. We assume that the censoring is random or non-informative. For simplicity, denote  $Z_i$  as another predictor of the risk of the event besides the diagnostic test under consideration. The multiple ratings are denoted by  $\{V_i^{(s)}, s = 1, 2, \dots, S\}$  with  $S = 3$  without loss of generality. The unobservable and underlying truth of the test for subject  $i$  is denoted by  $W_i$ . Therefore, the complete data for the  $i$ th subject is  $\mathbf{D}_{com} = \{X_i, \Delta_i, Z_i, V_i^{(s)}, W_i\}$  and the observed data for the  $i$ th subject is  $\mathbf{D}_{obs} = \{X_i, \Delta_i, Z_i, V_i^{(s)}\}$ , where  $i = 1, 2, \dots, n; s = 1, 2, \dots, S$ . Here, we propose to model the multiple independent ratings data by a latent class model as in Section 2.1 and the time-to-event data by a Cox proportional hazards model [5] with the latent truth as a predictor besides  $Z$ :

$$\lambda(t|Z_i, \tilde{W}_i) = \lambda_0(t) \exp(\boldsymbol{\beta} \tilde{W}_i + \gamma Z_i), \quad (3.1)$$

where  $\tilde{W}_i = (W_{i1}, W_{i2}) = (I(W_i = 2), I(W_i = 3))$ ,  $I(\cdot)$  is an indicator function;  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  and  $\gamma$  are the regression coefficients;  $\lambda_0(t)$  represents the baseline hazard function. Because

the more aggressive and progressed the tumor has the higher risk of recurrence, we would impose the following restriction on the Cox regression parameters:  $0 < \beta_1 < \beta_2$ .

In many circumstances, the assessment of a diagnostic test is blinded from the time-to-event data. Therefore, conditioned on the underlying truth  $W$ , the multiple ratings are independent from both the event time  $T$  and censoring time  $C$ . With the observed data  $\mathbf{D}_{obs} = \{X_i, \Delta_i, Z_i, V_i^{(s)}; i = 1, \dots, n; s = 1, 2, 3\}$ , the full likelihood function of  $\Omega = \{\boldsymbol{\pi}, \mathbf{q}, \boldsymbol{\beta}, \gamma, \lambda_0(\cdot)\}$  is:

$$L(\boldsymbol{\pi}, \mathbf{q}, \boldsymbol{\beta}, \gamma, \lambda_0(t); \mathbf{D}_{obs}) = \prod_{i=1}^n \left( \sum_{j=1}^3 \left[ \pi_j \left\{ \prod_{s=1}^3 \prod_{k=1}^3 (q_{jk}^{(s)})^{I(v_i^{(s)}=k)} \right\} f(X_i, \Delta_i | W_i = j, Z_i) \right] \right), \quad (3.2)$$

where

$$f(X_i, \Delta_i | W_i = j, Z_i) = [\lambda_0(X_i) \exp(\boldsymbol{\beta}\tilde{W}_i + \gamma Z_i)]^{\Delta_i} \exp[-\exp(\boldsymbol{\beta}\tilde{W}_i + \gamma Z_i) \int_0^{X_i} \lambda_0(u) du] \quad (3.3)$$

## 3.2 AN EM ALGORITHM FOR PARAMETER ESTIMATION

Regarding the underlying truth as missing data, the EM algorithm can be implemented for finding the MLE of  $\Omega$  [8, 35]. In the expectation step (E-step), the conditional expectation of the complete data log-likelihood function given the observed data under current parameter estimates is calculated. In the maximization step (M-step), the current parameter estimates are updated by maximizing the conditionally expected log-likelihood function obtained in the E-step. The steps will be repeated until the parameter estimates reach convergence.

### 3.2.1 E-step

The complete-data log-likelihood function for the proposed joint model is:

$$\begin{aligned} & l_{com}(\boldsymbol{\pi}, \mathbf{q}, \boldsymbol{\beta}, \gamma, \lambda_0(t)) \\ &= \sum_{i=1}^n \left[ \sum_{j=1}^3 I(W_i = j) \left\{ \log(\pi_j) + \sum_{s=1}^3 \sum_{k=1}^3 I(V_i^{(s)} = k) \log(q_{jk}^{(s)}) \right\} \right. \\ &+ \left. \Delta_i \left\{ \log(\lambda_0(X_i)) + \gamma Z_i + \boldsymbol{\beta}\tilde{W}_i \right\} - \exp(\boldsymbol{\beta}\tilde{W}_i + \gamma Z_i) \int_0^{X_i} \lambda_0(u) du \right] \quad (3.4) \end{aligned}$$

In the E-step, with current estimate  $\Omega^{(t)}$ , we would need to calculate the conditional expectation of the complete-data log-likelihood,  $E[l_{com}(\boldsymbol{\pi}, \mathbf{q}, \boldsymbol{\beta}, \gamma, \lambda_0(t)) | \Omega^{(t)}]$ .

For any measurable function  $h(\cdot)$ , the conditional expectation of the form  $E[h(\tilde{W}_i) | D_{obs}; \Omega^{(t)}]$  is defined as

$$E[h(\tilde{W}_i) | D_{obs}; \Omega^{(t)}] = \sum_{j=1}^3 Pr(W_i = j | D_{obs}; \Omega^{(t)}) h(\tilde{W}_i) |_{W_i=j}.$$

Then the conditional expectation of the complete-data log-likelihood for the proposed joint model can be calculated by

$$\begin{aligned} E[I(W_i = j) | D_{obs}; \Omega^{(t)}] &= Pr[W_i = j | D_{obs}, \Omega^{(t)}] \\ &= \frac{\pi_j^{(t)} [\prod_{s=1}^3 \prod_{k=1}^3 (q_{jk}^{(s,t)})^{I(v_i^s=k)}] f(X_i, \Delta_i | W_i = j, Z_i; \Omega^{(t)})}{\sum_{j=1}^3 \pi_j^{(t)} [\prod_{s=1}^3 \prod_{k=1}^3 (q_{jk}^{(s,t)})^{I(v_i^s=k)}] f(X_i, \Delta_i | W_i = j, Z_i; \Omega^{(t)})}, \end{aligned}$$

where the calculation of  $f(X_i, \Delta_i | W_i = j, Z_i; \Omega^{(t)})$  is presented in the equation (3.3).

### 3.2.2 M-step

At the M-step, we find the  $\Omega = \Omega^{(t+1)}$  that maximizes  $E[l_{com}(\boldsymbol{\pi}, \mathbf{q}, \boldsymbol{\beta}, \gamma, \lambda_0(t) | \mathbf{D}_{obs}); \Omega^{(t)}]$  and update it as the new estimate of  $\Omega$ . The parameters for the latent class models are updated as:

$$\pi_j^{(t+1)} = \frac{\sum_{i=1}^n E[I(W_i = j) | D_{obs}; \Omega^{(t)}]}{n} \quad (3.5)$$

$$q_{jk}^{(s,t+1)} = \frac{\sum_{i=1}^n I(V_i^{(s)} = k) E[I(W_i = j) | D_{obs}; \Omega^{(t)}]}{\sum_{i=1}^n E[I(W_i = j) | D_{obs}; \Omega^{(t)}]}. \quad (3.6)$$

Given the updated estimates of the coefficients in the Cox model,  $(\gamma^{(t+1)}, \beta^{(t+1)})$ , the baseline hazard function  $\lambda_0(\cdot)$  can be updated with the following formula:

$$\lambda_0^{(t+1)}(u) = \frac{\sum_{i=1}^n \Delta_i I(X_i = u)}{\sum_{i: X_i \geq \mu} E[\exp(\beta^{(t+1)} \tilde{W}_i + \gamma^{(t+1)} Z_i) | D_{obs}; \Omega^{(t)}]}, \quad (3.7)$$

where

$$E[\exp(\beta^{(t+1)} \tilde{W}_i + \gamma^{(t+1)} Z_i) | D_{obs}; \Omega^{(t)}] = \sum_{j=1}^3 Pr[W_i = j | D_{obs}; \Omega^{(t)}] \exp(\beta^{(t+1)} \tilde{W}_i + \gamma^{(t+1)} Z_i) |_{W_i=j}.$$

However, there are no closed-form solutions for  $(\gamma^{(t+1)}, \beta^{(t+1)})$  and in general an iterative optimization algorithm is necessary. In the M-step, it can be shown that the update of Cox model regression parameters is equivalent to fitting a survey-weighted Cox proportional hazards model that was introduced by Binder [3]. A one-step Newton-Raphson approach in Wulfsohn and Tsiatis [3] can be considered as an alternative.

### A. Using a survey-weighted Cox proportional hazards model

By creating some pseudo observations, it can be shown that the update of the Cox regression parameters  $\theta = (\beta, \gamma)$  in the M-step is equivalent to fitting a survey-weighted Cox model. In Binder [3], the estimating equation for the regression parameter  $\theta$  is

$$\sum_{i=1}^N \omega_i \delta_i \left[ z'_i(x_i) - \frac{S^{(1)}(x_i, \boldsymbol{\theta})}{S^{(0)}(x_i, \boldsymbol{\theta})} \right] = 0, \quad (3.8)$$

where

$$\begin{aligned} S^{(0)}(x_i, \boldsymbol{\theta}) &= \frac{1}{N} \sum_{j \in R(x_i)} \omega_j \exp\{z'_j(x_i) \boldsymbol{\theta}\} \\ S^{(1)}(x_i, \boldsymbol{\theta}) &= \frac{1}{N} \sum_{j \in R(x_i)} \omega_j z_j(x_i) \exp\{z'_j(x_i) \boldsymbol{\theta}\}. \end{aligned}$$

In the above equations,  $\omega_i$  is the sampling weight associated with subject  $i$  and is scaled so that  $\sum_i \omega_i = 1$ . Functions `svycoxph` and `svydesign` in the R package `survey` are available to fit such models. In order to show that our update of  $\boldsymbol{\theta}$  in the M-step is equivalent to fitting a survey-weighted Cox model, we first create a new data set with three pseudo observations  $(i0, i1, i2)$  for each subject  $i$ ,  $i = 1, 2, \dots, n$  such that:

- (1) All these pseudo observations have the same values on all observed variables such as other predictors than the true tumor grade, observed event time and censoring code as the corresponding subject  $i$ :  $(X_{ij}, \delta_{ij}, z_{ij}) = (X_i, \delta_i, z_i)$ ,  $j = 0, 1, 2$
- (2)  $W_{i0}^* = 1$ ,  $W_{i1}^* = 2$  and  $W_{i2}^* = 3$ .

Therefore, these pseudo observations represent the cases when the true tumor grade is well, moderate, and poor, respectively. Denote the corresponding weights as  $\omega_{ij} = E[I(W_i = j + 1)|D_{obs}; \boldsymbol{\theta}^{(t)}]$ ,  $j = 0, 1, 2$ . Then we can re-write the score equations for estimating the cox model regression parameters at the M-step as:

$$\begin{aligned}
S_\gamma &= \sum_{i=1}^n \Delta_i \left\{ Z_i - \frac{\sum_{k: X_k \geq X_i} E[Z_k \exp(\gamma' Z_k + \beta' \tilde{W}_k) | D_{obs}; \boldsymbol{\theta}^{(t)}]}{\sum_{k: X_k \geq X_i} E[\exp(\gamma' Z_k + \beta' \tilde{W}_k) | \boldsymbol{\theta}^{(t)}]} \right\} \\
&= \sum_{i=1}^n \left\{ \sum_{j=0}^2 \omega_{ij} \Delta_i Z_i - \Delta_i \sum_{j=0}^2 \omega_{ij} \frac{\sum_{k: X_k \geq X_i} Z_k E[\exp(\gamma' Z_k + \beta' \tilde{W}_k) | D_{obs}; \boldsymbol{\theta}^{(t)}]}{\sum_{k: X_k \geq X_i} E[\exp(\gamma' Z_k + \beta' \tilde{W}_k) | \boldsymbol{\theta}^{(t)}]} \right\} \\
&= \sum_{i=1}^n \sum_{j=0}^2 \omega_{ij} \Delta_{ij} \left\{ Z_{ij} - \frac{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} Z_{kj} \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)}{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)} \right\} \\
S_{\beta_1} &= \sum_{i=1}^n \Delta_i \left\{ E[W_i = 2 | D_{obs}; \boldsymbol{\theta}^{(t)}] - \frac{\sum_{k: X_k \geq X_i} E[W_k = 2 | D_{obs}; \boldsymbol{\theta}^{(t)}] \exp(\gamma' Z_k + \beta' \tilde{W}_k)}{\sum_{k: X_k \geq X_i} E[\exp(\gamma' Z_k + \beta' \tilde{W}_k) | \boldsymbol{\theta}^{(t)}]} \right\} \\
&= \sum_{i=1}^n \sum_{j=0}^2 \omega_{ij} \Delta_{ij} \left\{ I(W_{ij}^* = 2) - \frac{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} I(\omega_{kj}^* = 2) \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)}{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)} \right\} \\
S_{\beta_2} &= \sum_{i=1}^n \Delta_i \left\{ E[W_i = 3 | D_{obs}; \boldsymbol{\theta}^{(t)}] - \frac{\sum_{k: X_k \geq X_i} E[W_k = 3 | D_{obs}; \boldsymbol{\theta}^{(t)}] \exp(\gamma' Z_k + \beta' \tilde{W}_k)}{\sum_{k: X_k \geq X_i} E[\exp(\gamma' Z_k + \beta' \tilde{W}_k) | \boldsymbol{\theta}^{(t)}]} \right\} \\
&= \sum_{i=1}^n \sum_{j=0}^2 \omega_{ij} \Delta_{ij} \left\{ I(W_{ij}^* = 3) - \frac{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} I(\omega_{kj} = 3) \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)}{\sum_{k: X_k \geq X_i} \sum_{j=0}^2 \omega_{kj} \exp(\gamma' Z_{kj} + \beta' \tilde{W}_{kj}^*)} \right\}
\end{aligned}$$

Therefore, the update of  $(\gamma, \beta)$  in the M-step is equivalent to fitting a survey-weighted Cox proportional hazards model on the pseudo observations  $\{X_{ij}, \delta_{ij}, z_{ij}, W_{ij}^*; i = 1, 2, \dots, n; j = 0, 1, 2\}$  with weights  $\{\omega_{ij}, i = 1, 2, \dots, n; j = 0, 1, 2\}$ . The normalization of the weights with their summation being one will not affect the estimating equations.

## B. Using a one-step Newton-Raphson algorithm

First, we can update it via a one-step Newton-Raphson method suggested by Wulfsohn and Tsiatis [35] as following:

$$\hat{\boldsymbol{\theta}}^{(k+1)} = \hat{\boldsymbol{\theta}}^{(k)} + I(\hat{\boldsymbol{\theta}}^{(k)}; D_{obs})^{-1} S_{\hat{\boldsymbol{\theta}}^{(k)} | D_{obs}},$$

where  $S_{\hat{\boldsymbol{\theta}}^{(k)}|D_{obs}}$  is the score function and  $I(\hat{\boldsymbol{\theta}}^{(k)}; D_{obs})$  is an information matrix for  $\hat{\boldsymbol{\theta}}^{(k)} = (\gamma^{(k)}, \beta^{(k)})$  at  $k$ th iteration. The elements of the score function are:

$$\begin{aligned} S_{\gamma} &= \sum_{i=1}^n \Delta_i \left[ Z_i - \frac{\sum_{j: X_j \geq X_i} Z_j E[\exp(\gamma' Z_i + \beta' \tilde{W}_i) | D_{obs}; \hat{\boldsymbol{\theta}}^{(k)}]}{\sum_{j: X_j \geq X_i} E[\exp(\gamma' Z_i + \beta' \tilde{W}_i) | \hat{\boldsymbol{\theta}}^{(k)}]} \right] \\ S_{\beta_1} &= \sum_{i=1}^n \Delta_i \left[ E[I(W_i = 2) | D_{obs}; \hat{\boldsymbol{\theta}}^{(k)}] - \frac{\sum_{j: X_j \geq X_i} E[I(W_i = 2) \exp(\gamma' Z_i + \beta' \tilde{W}_i) | D_{obs}; \hat{\boldsymbol{\theta}}^{(k)}]}{\sum_{j: X_j \geq X_i} E[\exp(\gamma' Z_i + \beta' \tilde{W}_i) | \hat{\boldsymbol{\theta}}^{(k)}]} \right] \\ S_{\beta_2} &= \sum_{i=1}^n \Delta_i \left[ E[I(W_i = 3) | D_{obs}; \hat{\boldsymbol{\theta}}^{(k)}] - \frac{\sum_{j: X_j \geq X_i} E[I(W_i = 3) \exp(\gamma' Z_i + \beta' \tilde{W}_i) | D_{obs}; \hat{\boldsymbol{\theta}}^{(k)}]}{\sum_{j: X_j \geq X_i} E[\exp(\gamma' Z_i + \beta' \tilde{W}_i) | \hat{\boldsymbol{\theta}}^{(k)}]} \right]. \end{aligned}$$

The elements of the information matrix  $I(\hat{\boldsymbol{\theta}}^{(k)}; D_{obs})$  are derived by a numerical approximation on the score function  $S_{\hat{\boldsymbol{\theta}}^{(k)}|D_{obs}}$ .

With the local identifiability discussed in Section 2.1.3, the EM algorithm with the Newton-Raphson method in the M-steps will converge to a local maximum of the likelihood function (Equation 3.4). To handle the local identifiability with the ‘‘label-switching’’ phenomenon, the estimates for the latent class model are re-arranged so that  $0 < \hat{\beta}_1 < \hat{\beta}_2$ . This re-arrangement enables us to achieve the global identification on the estimates for the latent class model under the order restrictions on  $\boldsymbol{\beta}$ . For example, if  $0 > \hat{\beta}_1 > \hat{\beta}_2$ , the labels on  $\phi = \{\pi_1, \pi_2, \pi_3, q_{1k}^{(s)}, q_{2k}^{(s)}, q_{3k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$  are switched to  $\phi' = \{\pi_3, \pi_2, \pi_1, q_{3k}^{(s)}, q_{2k}^{(s)}, q_{1k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$ , where  $\sum_{j=1}^3 \pi_j = \sum_{k=1}^3 q_{jk}^{(s)} = 1$ . If  $0 < \hat{\beta}_2 < \hat{\beta}_1$ , the labels on  $\phi$  are switched to  $\phi' = \{\pi_1, \pi_3, \pi_2, q_{1k}^{(s)}, q_{3k}^{(s)}, q_{2k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$ , where  $\sum_{j=1}^3 \pi_j = \sum_{k=1}^3 q_{jk}^{(s)} = 1$ . The order restriction is the critical assumption for the proposed joint model to guarantee the global identification. Hence, testing for the order restriction is necessary. The testing procedure will be discussed in Chapter 4.3.

### 3.3 VARIANCE ESTIMATION FROM THE PROFILE LIKELIHOOD

A likelihood function can be used to estimate variances for parameters of interest with its observed information matrix. However, the use of full likelihood function is not feasible when the dimension of parameter space is too high or the model has a semi-parametric nature.

Due to its semi-parametric nature, the variance of the parameter estimates for the Cox model in the proposed joint model cannot be readily derived from the usual likelihood approach with finite number of parameters. Instead of using the full likelihood function, we apply the profile likelihood method for inference on those estimates [19]. Denote  $\Omega = (\theta, \phi)$  with  $\theta = (\beta_1, \beta_2, \gamma)$  as the Cox regression parameters in the proposed joint model and  $\phi$  the other nuisance parameters, including parameters for the latent class model and baseline hazard for the Cox model. Then, the profile likelihood for  $\theta = (\beta_1, \beta_2, \gamma)$  is

$$L_{pl}(\theta) = L(\theta, \hat{\phi}(\theta); D_{obs}) = \max_{\phi} L(\theta, \phi; D_{obs}). \quad (3.9)$$

With the maximum likelihood estimators  $\hat{\Omega}$ , the asymptotic variances for  $\hat{\theta}$  are estimated by inverting the negative of the second derivatives of the logarithm of the profile likelihood :

$$\widehat{var}(\hat{\theta}) = \left[ -\frac{\partial^2 L_{pl}(\theta)}{\partial \theta \partial \theta^T} \right]^{-1} \Big|_{\theta = \hat{\theta}_{MLE}}.$$

Usually, we do not have the analytical forms for the second derivatives of the log-profile likelihood function. We adapt a numerical approximation to derive the second derivatives of the log-profile likelihood function [19]. With  $h = 0.001$ ,  $\theta = (\beta_1, \beta_2, \gamma)$ , and  $l_{pl}(\theta) = \log L_{pl}(\theta, \phi; D_{obs})$ , we approximate the second derivatives of the log-profile likelihood scores as below:

$$\begin{aligned} \frac{\partial^2 l_{pl}}{\partial \gamma^2} &= \frac{l_{pl}(\gamma + h, \beta_1, \beta_2) - 2l_{pl}(\gamma, \beta_1, \beta_2) + l_{pl}(\gamma - h, \beta_1, \beta_2)}{h^2} + o(h) \\ \frac{\partial^2 l_{pl}}{\partial \beta_1^2} &= \frac{l_{pl}(\gamma, \beta_1 + h, \beta_2) - 2l_{pl}(\gamma, \beta_1, \beta_2) + l_{pl}(\gamma, \beta_1 - h, \beta_2)}{h^2} + o(h) \\ \frac{\partial^2 l_{pl}}{\partial \beta_2^2} &= \frac{l_{pl}(\gamma, \beta_1, \beta_2 + h) - 2l_{pl}(\gamma, \beta_1, \beta_2) + l_{pl}(\gamma, \beta_1, \beta_2 - h)}{h^2} + o(h) \\ \frac{\partial^2 l_{pl}}{\partial \gamma \partial \beta_1} &= \frac{1}{4h^2} \{l_{pl}(\gamma + h, \beta_1 + h, \beta_2) + l_{pl}(\gamma - h, \beta_1 - h, \beta_2) \\ &\quad - l_{pl}(\gamma - h, \beta_1 + h, \beta_2) - l_{pl}(\gamma + h, \beta_1 - h, \beta_2)\} + o(h) \\ \frac{\partial^2 l_{pl}}{\partial \gamma \partial \beta_2} &= \frac{1}{4h^2} \{l_{pl}(\gamma + h, \beta_1, \beta_2 + h) + l_{pl}(\gamma - h, \beta_1, \beta_2 - h) \\ &\quad - l_{pl}(\gamma - h, \beta_1, \beta_2 + h) - l_{pl}(\gamma + h, \beta_1, \beta_2 - h)\} + o(h) \\ \frac{\partial^2 l_{pl}}{\partial \beta_1 \partial \beta_2} &= \frac{1}{4h^2} \{l_{pl}(\gamma, \beta_1 + h, \beta_2 + h) + l_{pl}(\gamma, \beta_1 - h, \beta_2 - h) \\ &\quad - l_{pl}(\gamma, \beta_1 - h, \beta_2 + h) - l_{pl}(\gamma, \beta_1 + h, \beta_2 - h)\} + o(h), \end{aligned}$$

and  $\frac{\partial^2 l_{pl}}{\partial \theta_i \partial \theta_j} = \frac{\partial^2 l_{pl}}{\partial \theta_j \partial \theta_i}$ , where  $i, j = 1, 2, 3$ .

## 4.0 STATISTICAL TESTING WITH ORDER RESTRICTED HYPOTHESIS

Hypothesis testing for statistical model parameter  $\boldsymbol{\theta}$  is mainly focused on the null hypothesis  $H_0 : \boldsymbol{\theta} = 0$  against  $H_1 : \boldsymbol{\theta} \neq 0$ . In this conventional statistical inference, likelihood-base approaches, such as likelihood ratio, score, and Wald tests have been commonly used for hypothesis testing [18]. Under a large sample, the likelihood-based approaches gain the optimal statistical inferences even the regularity conditions are somewhat violated [26]. However, under a constrained parameter space, the optimal statistical inferences may not be achieved regardless of a sample size [26]. Molenberghs and Verbeke [18] provided the frameworks of the likelihood-based approaches for one-sided testing in constrained parameter spaces with some illustrative examples. Under constrained one-sided alternatives, Self and Liang [24] provided the theoretical works on the likelihood ratio test and its null distribution. Silvapulle and Silvapulle [27] proposed Wald-type tests under restricted one-sided alternatives. Silvapulle [28] also provided the score test under a constrained alternatives.

Even the regularity conditions on the likelihood-based approaches are met, the explicit form of maximum likelihood estimators and likelihood function are sometimes hard to be derived due to the complexity of statistical model [26]. As an alternative of likelihood-based approaches, the Union-Intersection (UI) principle introduced by Roy [23] has been considered for constrained statistical inferences. With preserving the statistical properties setting in likelihood-based approaches, the UI principle provides a flexibility of an alternative hypothesis by allowing complex parametric or beyond parametric parameters [26]. Sen [25] introduced the application of the UI principle for the Cox proportional hazards model with various order restricted alternatives. With formulating a partial likelihood of the parameters of interest, Sen [25] presented more efficient statistical inferences on the order constrained alternatives.

In the following sections, we first review the statistical hypothesis testing under standard conditions and the Union-Intersection principle for the Cox proportional hazards model introduced by Sen [25]. Then, the use of the Union-Intersection principle for the proposed joint model is discussed in Section 4.3.

#### 4.1 LIKELIHOOD-BASE APPROACHES

Let  $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\psi}^T)^T$  denote a  $1 \times p$  vector of parameters. In the parameter vector  $\boldsymbol{\theta}$ ,  $\boldsymbol{\beta}$  is the vector of parameters of interest and  $\boldsymbol{\psi}$  is a vector of nuisance parameters. Under the null hypothesis  $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ , the log-likelihood and score functions are denoted by  $l(\boldsymbol{\theta})$  and  $S(\boldsymbol{\theta}) = \partial l(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ , respectively. Denote  $J(\boldsymbol{\theta})$  as the matrix of second derivatives of the log-likelihood function. Then,  $S(\boldsymbol{\theta})$  can be decomposed into  $S(\boldsymbol{\theta}) = (S_{\boldsymbol{\beta}}(\boldsymbol{\theta}), S_{\boldsymbol{\psi}}(\boldsymbol{\theta}))$ .  $J(\boldsymbol{\theta})$  can be partitioned as

$$\begin{pmatrix} J_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) & J_{\boldsymbol{\beta}\boldsymbol{\psi}}(\boldsymbol{\theta}) \\ J_{\boldsymbol{\psi}\boldsymbol{\beta}}(\boldsymbol{\theta}) & J_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}) \end{pmatrix}$$

and its inverse  $J^{-1}(\boldsymbol{\theta})$  can also be partitioned as

$$\begin{pmatrix} J^{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\theta}) & J^{\boldsymbol{\beta}\boldsymbol{\psi}}(\boldsymbol{\theta}) \\ J^{\boldsymbol{\psi}\boldsymbol{\beta}}(\boldsymbol{\theta}) & J^{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\theta}) \end{pmatrix}.$$

##### 4.1.1 General two-sided alternative: $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ vs. $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$

The Wald test statistic is defined as

$$T_w = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T [J^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

The score test statistic is defined as

$$T_s = [S_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))]^T [J^{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))] [S_{\boldsymbol{\beta}}(\boldsymbol{\beta}_0, \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))],$$

where  $\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)$  is a restricted maximum partial likelihood estimator at  $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ .

The likelihood ratio test statistic is defined as

$$T_{lr} = 2[l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) - l(\boldsymbol{\beta}_0, \hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0))],$$

where  $\hat{\boldsymbol{\psi}}(\boldsymbol{\beta}_0)$  is a restricted maximum partial likelihood estimator at  $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ .

The asymptotic null distributions of the three tests are equivalent to a chi-square with the degree of freedom  $p$ , where  $p$  is the number of restrictions imposed on  $\boldsymbol{\beta}$  by  $H_0$ .

#### 4.1.2 Constrained one-sided alternative: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ vs. $H_1 : \boldsymbol{\beta} \in C$

In Molenberghs and Verbeke [18], the authors defined the differences between constrained and unconstrained one-sided test by whether negative estimates for  $\boldsymbol{\beta}$  are allowed in the null hypothesis. For the same alternative hypothesis  $H_1 : \boldsymbol{\beta} > \mathbf{0}$ , the null hypothesis for the unconstrained one-sided hypothesis testing is defined as  $H_0 : \boldsymbol{\beta} \leq \mathbf{0}$  and that for the constrained one-sided hypothesis testing sets  $H_0 : \boldsymbol{\beta} = \mathbf{0}$ . From the null hypothesis, any negative values of  $\boldsymbol{\beta}$  will be replaced by  $\boldsymbol{\beta} = 0$  under the constrained one-sided hypothesis [18].

For general form of alternative hypotheses, Molenberghs and Verbeke [18] denoted  $C$  as a closed and convex cone in the Euclidean space, with vertex at the origin. With  $\boldsymbol{\beta}_0 = \mathbf{0}$  and  $\mathbf{Z} = N^{-1/2}S_{\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}}', \mathbf{0}')$ , a one-sided score statistic is defined as [18, 28]

$$T_s = \mathbf{Z}' J^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}}', \mathbf{0}') \mathbf{Z} - \inf \left\{ (\mathbf{Z} - \mathbf{b})' J^{\boldsymbol{\beta}\boldsymbol{\beta}}(\hat{\boldsymbol{\psi}}', \mathbf{0}') (\mathbf{Z} - \mathbf{b}) \mid \mathbf{b} \in C \right\}.$$

A one-sided Wald statistic is defined as [18, 27]

$$T_w = \hat{\boldsymbol{\beta}}' V_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \hat{\boldsymbol{\beta}} - \inf \left\{ (\hat{\boldsymbol{\beta}} - \mathbf{b})' V_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} (\hat{\boldsymbol{\beta}} - \mathbf{b}) \mid \mathbf{b} \in C \right\},$$

where  $V$  is the asymptotic variance-covariance matrix of  $\boldsymbol{\theta}$  and  $V_{\boldsymbol{\beta}\boldsymbol{\beta}}$  is the corresponding sub-matrix.

A one-sided likelihood ratio test is defined as

$$T_{lr} = 2[\sup\{l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \mid \boldsymbol{\beta} \in C\} - \sup\{l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\psi}}) \mid \boldsymbol{\beta} = \mathbf{0}\}].$$

The asymptotic p-values of the three tests can be calculated from the mixture of chi-bar square distribution. With an observed test statistic  $t_{obs}$  and  $p$ , which is the number of restrictions imposed on  $\beta$  by  $H_0$ , the asymptotic p-value can be calculated by

$$\lim_{n \rightarrow \infty} pr(t_{obs} \geq c | H_0) = \sum_{i=0}^p \omega_i(p, J^{\beta\beta}(\theta)) pr(\chi_i^2 \geq t_{obs}).$$

When we test the null hypothesis with  $p$  restrictions on  $\beta$ , the mixture of chi-bar square distribution can be calculated by

$$\sum_{i=0}^p 2^{-p} \binom{i}{p} \chi_i^2.$$

For example, if the null hypothesis is  $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3$ , which can be rewrote as  $H_0 : \lambda_1 = \lambda_2 = \lambda_3 = 0$ , where  $\lambda_p = \beta_p - \beta_{p-1}$ ,  $p = 1, 2, 3$ , the null distribution under the constrained one-sided alternative can be defined as

$$\sum_{i=0}^p 2^{-p} \binom{i}{p} \chi_i^2 = \frac{1}{8} \chi_0^2 + \frac{3}{8} \chi_1^2 + \frac{3}{8} \chi_2^2 + \frac{1}{8} \chi_3^2.$$

## 4.2 THE UNION-INTERSECTION PRINCIPLE FOR COX PROPORTIONAL HAZARDS MODELS

In a multi-parameter setting, testing for the null hypothesis  $H_0 : \beta = \mathbf{0}$  against the alternative hypothesis with order restrictions on the parameters, such as  $H_1 : 0 \leq \beta_1 \leq \beta_2 \leq \dots \leq \beta_p$ , may be interested. In such a case, the statistical testing with the likelihood-based approaches may not be optimal and hard to be conducted [26]. Roy [23] proposed the Union-Intersection (UI) principle that can preserve the statistical properties of the likelihood-based approaches and adapt more various settings of alternative hypotheses with good robustness. Sen [25] showed the use of the UI principle for testing order restricted alternatives on the Cox model parameters with partial likelihood scores. We review Sen's approaches for the Cox model in this section and discuss the application of the UI principle to the proposed joint model in Section 4.3.

For simplicity, Sen [25] considered the general Cox proportional hazards model with categorical covariates  $c_i$  and a continuous covariate  $z_i$ . With the observable dataset  $(X_i, \delta_i, \mathbf{c}_i, z_i)$ , where  $X_i$  and  $\delta_i$  are the observed failure time and the failure indicator for each subject  $i$ , the hazard rate is defined as

$$h_i(t; \mathbf{c}_i, z_i) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{c}_i + \gamma' z_i),$$

where  $i = 1, \dots, n$ ;  $h_0(t)$  is an unknown, arbitrary nonnegative function. Here, the  $p$  parameters of interest are denoted as  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$  and  $\mathbf{c}_i$  is the set of vectors  $\{(0, \dots, 0)', (1, \dots, 0)', \dots, (0, \dots, 1)'\}$  for a subject  $i$ .

Based on the observable dataset  $(X_i, \delta_i, \mathbf{c}_i, z_i), i = 1, \dots, n$ , the test for  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  against  $H_1^>, H_1^*$  or  $H_1^{*>}$  may be interested, where

$$H_1^> : \quad \beta_j \geq 0, j = 1, \dots, p$$

$$H_1^* : \quad \beta_1 \leq \dots \leq \beta_p$$

$$H_1^{*>} : \quad 0 \leq \beta_1 \leq \dots \leq \beta_p.$$

Even the global test for the Cox model parameters based on the partial likelihood is still valid, it may not be optimal approaches for these alternative hypotheses regarding efficiency and/or power properties [26]. To implement the UI principle for testing the constrained hypotheses under the Cox model, we first build the partial likelihood function with the observable dataset  $(X_i, \delta_i, \mathbf{c}_i, z_i), i = 1, \dots, n$ . The partial likelihood function is given by [25]

$$L_N^P(\boldsymbol{\beta}, \gamma) = \prod_{i=1}^N \left\{ \frac{\exp(\boldsymbol{\beta}' \mathbf{c}_i + \gamma' z_i)}{\sum_{j: X_j \geq X_i} \exp(\boldsymbol{\beta}' \mathbf{c}_j + \gamma' z_j)} \right\}^{\delta_i}.$$

Based on the maximum partial likelihood estimator of  $\gamma, \hat{\gamma}_N^0$ , the partial likelihood scores for  $\boldsymbol{\beta}$  are defined by:

$$\hat{U}_N = N^{-1/2} \left( \frac{\partial}{\partial \boldsymbol{\beta}} \right) \log L_N^P(\boldsymbol{\beta}, \gamma) \Big|_{\boldsymbol{\beta}=\mathbf{0}, \gamma=\hat{\gamma}_N^0} \quad (4.1)$$

$$= N^{-1/2} \sum_{i=1}^N \delta_i \left\{ \mathbf{c}_i - \frac{\sum_{j: X_j \geq X_i} \mathbf{c}_j \exp(z_j' \hat{\gamma}_N^0)}{\sum_{j: X_j \geq X_i} \exp(z_j' \hat{\gamma}_N^0)} \right\}. \quad (4.2)$$

Here,  $\hat{\gamma}_N^0$  can be obtained by solving the below equation under the null hypothesis  $H_0 : \boldsymbol{\beta} = \mathbf{0}$ .

$$\sum_{i=1}^N \delta_i \left\{ z_i - \frac{\sum_{j: X_j \geq X_i} z_j \exp(\gamma' z_j)}{\sum_{j: X_j \geq X_i} \exp(\gamma' z_j)} \right\} = 0.$$

Let  $\omega_{ij} = I(X_j \geq X_i) \exp(z_j' \hat{\gamma}_N^0)$ , for  $i, j = 1, \dots, N$ . Then, we define  $\bar{V}_{11}, \bar{V}_{12}, \bar{V}_{22}, \bar{V}_{11.2}$ , which are second derivatives of the partial likelihood scores with respect to  $\boldsymbol{\beta}, \gamma$ , by

$$\begin{aligned} \bar{V}_{11} &= N^{-1} \sum_{i=1}^N \delta_i \left\{ \left( \sum_{i=1}^N \omega_{ij} \right)^{-1} \sum_{i=1}^N \omega_{ij} \mathbf{c}_j \mathbf{c}_j' - \left( \sum_{i=1}^N \omega_{ij} \right)^{-2} \left( \sum_{i=1}^N \omega_{ij} \mathbf{c}_j \right) \left( \sum_{i=1}^N \omega_{ij} \mathbf{c}_j' \right) \right\} \\ \bar{V}_{12} &= N^{-1} \sum_{i=1}^N \delta_i \left\{ \left( \sum_{i=1}^N \omega_{ij} \right)^{-1} \sum_{i=1}^N \omega_{ij} \mathbf{c}_j z_j' - \left( \sum_{i=1}^N \omega_{ij} \right)^{-2} \left( \sum_{i=1}^N \omega_{ij} \mathbf{c}_j \right) \left( \sum_{i=1}^N \omega_{ij} z_j' \right) \right\} \\ \bar{V}_{22} &= N^{-1} \sum_{i=1}^N \delta_i \left\{ \left( \sum_{i=1}^N \omega_{ij} \right)^{-1} \sum_{i=1}^N \omega_{ij} z_j z_j' - \left( \sum_{i=1}^N \omega_{ij} \right)^{-2} \left( \sum_{i=1}^N \omega_{ij} z_j \right) \left( \sum_{i=1}^N \omega_{ij} z_j' \right) \right\} \\ \bar{V}_{11.2} &= \bar{V}_{11} - \bar{V}_{12} \bar{V}_{22}^{-1} \bar{V}_{21}. \end{aligned}$$

For given  $\mathbf{b} \geq \mathbf{0}$ , we can write  $H_1^> : \beta_j \geq 0$  as  $H_1^> : \cup_B H_{\mathbf{b}}$ , where  $B = \{\mathbf{b} : \mathbf{b} \geq \mathbf{0}\}$ . Under  $H_0$ ,  $(\mathbf{b}' \hat{U}_N) / (\mathbf{b}' \bar{V}_{11.2} \mathbf{b})$  closely follows  $N(0, 1)$  for a given  $\mathbf{b}$ . Hence, the Union-Intersection test statistic for testing  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  vs.  $H_1^> : \boldsymbol{\beta} \geq \mathbf{0}$  can be defined as

$$T_N^{(1)} = \sup \left\{ (\mathbf{b}' \hat{U}_N) / (\mathbf{b}' \bar{V}_{11.2} \mathbf{b})^{1/2} : \mathbf{b} \in B \right\}.$$

The supremum of  $T_N^{(1)}$  can be obtained by maximizing  $\mathbf{b}' \hat{U}_N$ , where  $\mathbf{b} > \mathbf{0}$  and  $\mathbf{b}' \bar{V}_{11.2} \mathbf{b}$  is a constant. Based on the Kuhn-Tucker-Lagrange (KTL-) point formula theorem, we can find the solution for maximizing  $\mathbf{b}' \hat{U}_N$ , that is  $\mathbf{b}^* \hat{U}_N$  [25]. Let  $h(\mathbf{b}) = -\mathbf{b}' \hat{U}_N$ ,  $h_1(\mathbf{b}) = -\mathbf{b}$ , and  $h_2(\mathbf{b}) = \mathbf{b}' \bar{V}_{11.2} \mathbf{b} - 1$ , then the Lagrangian function is

$$L(\mathbf{b}, t_1, t_2) = h(\mathbf{b}) + t_1' h_1(\mathbf{b}) + t_2' h_2(\mathbf{b})$$

and  $(\mathbf{b}^*, t_1^*, t_2^*)$  is a KTL-point that satisfies the system of conditions

$$\begin{aligned} t_1^* &\geq \mathbf{0}, & \mathbf{b}^* &\geq \mathbf{0}, & \mathbf{b}^* \bar{V}_{11.2} \mathbf{b}^* &= 1 \\ t_1^* \mathbf{b}^* &= \mathbf{0}, & -\hat{U}_N - \mathbf{t}_1^* + 2\mathbf{t}_2^* \bar{V}_{11.2} \mathbf{b}^* &= \mathbf{0}. \end{aligned}$$

Let  $\mathbf{a}$  be any subset of  $P = \{1, \dots, p\}$ , where  $p$  is the number of components in  $\hat{U}_N$  and  $\bar{\mathbf{a}}$  be the complementary subset ( $\emptyset \subseteq \mathbf{a} \subseteq P$ ). For each subset  $\mathbf{a}$ , the partial likelihood score  $\hat{U}'_N$  and the variance-covariance matrix  $\bar{V}_{11.2}$  are partitioned by :

$$\hat{U}'_N = (\hat{U}'_{N(\mathbf{a})}, \hat{U}'_{N(\bar{\mathbf{a}})}), \quad \bar{V}_{11.2} = \begin{pmatrix} \bar{V}_{11.2(\mathbf{a}\mathbf{a})} & \bar{V}_{11.2(\mathbf{a}\bar{\mathbf{a}})} \\ \bar{V}_{11.2(\bar{\mathbf{a}}\mathbf{a})} & \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})} \end{pmatrix}.$$

With the sub-vectors of  $\hat{U}'_N$  and the sub-blocks of  $\bar{V}_{11.2}$ ,  $\hat{U}^*_{N(\mathbf{a})}$  and  $\bar{V}^*_{11.2(\mathbf{a})}$  for each subset  $\mathbf{a}$  are calculated as below:

$$\hat{U}^*_{N(\mathbf{a})} = \hat{U}_{N(\mathbf{a})} - \bar{V}_{11.2(\mathbf{a}\bar{\mathbf{a}})} \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \quad (4.3)$$

$$\bar{V}^*_{11.2(\mathbf{a})} = \bar{V}_{11.2(\mathbf{a}\mathbf{a})} - \bar{V}_{11.2(\mathbf{a}\bar{\mathbf{a}})} \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \bar{V}_{11.2(\bar{\mathbf{a}}\mathbf{a})}. \quad (4.4)$$

For some set  $\mathbf{a} \subseteq P$ , a KTL-point is  $(a^*, t_1^*, t_2^*)$ , where

$$\begin{aligned} t_1^* &= \begin{pmatrix} \mathbf{0} \\ -\bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \end{pmatrix} \geq \mathbf{0} \\ 2t_2^* a^* &= a^* \hat{U}_N = \{\hat{U}'_N \bar{V}_{11.2}^{-1} \hat{U}_N - \hat{U}'_{N(\bar{\mathbf{a}})} \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}'_{N(\bar{\mathbf{a}})}\}^{\frac{1}{2}} \\ 2t_2^* &= \bar{V}_{11.2}^{-1} (\hat{U}_N + t_1^*). \end{aligned}$$

From the KTL-point, the solution for  $b^* \hat{U}_N$  is given by [25]

$$b^* \hat{U}_N = (\hat{U}_{N(\mathbf{a})}^* \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2},$$

where  $\hat{U}_{N(\mathbf{a})}^* > 0$  and  $\bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0$ .

Hence, the UI-test statistic is defined by [25]

$$\begin{aligned} T_N^{(1)} &= \sup \left\{ (\mathbf{b}' \hat{U}_N) / (\mathbf{b}' \bar{V}_{11.2} \mathbf{b})^{1/2} : \mathbf{b} \in B \right\} \\ &= \sum_{\emptyset \subseteq \mathbf{a} \subseteq P} \left\{ (\hat{U}_{N(\mathbf{a})}^* \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2} I(\hat{U}_{N(\mathbf{a})}^* > 0) I(\bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0) \right\}, \quad (4.5) \end{aligned}$$

where  $\mathbf{a}$  is a subset of  $P$ , and  $\hat{U}_{N(\mathbf{a})}^*$  and  $\bar{V}_{11.2(\mathbf{a})}^{*-1}$  are calculated by the equation (4.3) and (4.4).

Under the null hypothesis with large  $n$ , for each  $\mathbf{a}$ , ( $\emptyset \subseteq \mathbf{a} \subseteq P$ ),

- $\hat{U}_N$  is asymptotically normal with null mean vector and variance-covariance matrix  $v_{11.2}$ .
- $\bar{V}_{11.2}$  converges in probability to  $v_{11.2}$ .
- $\hat{U}_{N(\mathbf{a})}^*$  is asymptotically normal with null mean vector and variance-covariance matrix  $v_{11.2(\mathbf{a})}$ .
- $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})}$  is asymptotically normal with null mean vector and variance-covariance matrix  $v_{11.2(\bar{\mathbf{a}})}^{-1}$ .
- $(\hat{U}_{N(\mathbf{a})}^{*\prime} \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2}$  has asymptotically the central chi distribution with  $k_{\mathbf{a}}$  degrees of freedom, which is the cardinality of the set  $\mathbf{a}$ .
- For every  $x \geq 0$ , the three events  $I((\hat{U}_{N(\mathbf{a})}^{*\prime} \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2} \leq x)$ ,  $I(\hat{U}_{N(\mathbf{a})}^* > 0)$ ,  $I(\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0)$  are asymptotically mutually independent.

Therefore, the p-value can be obtained by

$$\begin{aligned}
P \left\{ T_N^{(1)} \leq x | H_0 \right\} &= P \left\{ \sum_{\emptyset \subseteq \mathbf{a} \subseteq P} \left\{ (\hat{U}_{N(\mathbf{a})}^{*\prime} \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2} I(\hat{U}_{N(\mathbf{a})}^* > 0) I(\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0) \right\} \leq x \right\} \\
&= \sum_{\emptyset \subseteq \mathbf{a} \subseteq P} P \left\{ \hat{U}_{N(\mathbf{a})}^* > 0 \right\} P \left\{ \bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0 \right\} P \left\{ \chi_k \leq x \right\} \\
&= \sum_{k=0}^p \omega_k P \left\{ \chi_k \leq x \right\}, \forall x \geq 0, \sum_{k=1}^p \omega_k = 1.
\end{aligned} \tag{4.6}$$

The null distribution of the UI-test statistic is a chi-bar distribution with the weight  $\omega_k$  for each  $k$ , where  $k$  is the number of elements in some subset  $\mathbf{a}$ . The weight,  $\omega_k$  can be computed by multiplying  $P(\hat{U}_{N(\mathbf{a})}^* > 0)$  by  $P(\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0)$ , where  $\hat{U}_{N(\mathbf{a})}^* \sim N(0, v_{11.2(\mathbf{a})})$  and  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \sim N(0, \bar{V}_{11.2(\bar{\mathbf{a}})}^{-1})$ . The weights  $\omega_k$  can be calculated by the exact formulas or a numerical integration.

For testing  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  vs.  $H_1^{* >} : 0 \leq \beta_1 \leq \dots \leq \beta_p$ , the reparameterization may be necessary to use Equation (4.5) and (4.6) for the test statistic and its p-value, respectively. Let  $\psi_j = \beta_j - \beta_{j-1}$ ,  $j = 1, \dots, p$ , where  $\beta_0 = \mathbf{0}$ , and  $\mathbf{d}'_i = (d_{i1}, \dots, d_{ip})$ ,  $i = 1, \dots, N$ , where  $d_{ij} = \sum_{s=j}^p c_{is}$ , for  $j = 1, \dots, p$ ,  $i = 1, \dots, N$ . Then, the Cox model can be written as

$$h_i(t; \mathbf{c}_i, z_i) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{c}_i + \gamma' z_i) = h_0(t) \exp(\boldsymbol{\psi}' \mathbf{d}_i + \gamma' z_i).$$

Then, the null and alternative hypotheses are reduced to  $H_0 : \boldsymbol{\psi} = \mathbf{0}$  and  $H_1^> : \boldsymbol{\psi} \geq \mathbf{0}$ . The UI-test for testing  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  vs.  $H^{*>} : 0 \leq \beta_1 \leq \dots \leq \beta_p$  is the same as that for testing  $H_1^> : \beta_j \geq 0, j = 1, \dots, p$ .

### 4.3 THE UNION-INTERSECTION PRINCIPLE FOR THE PROPOSED JOINT MODEL

In the proposed joint model in Chapter 3, we assume that the more aggressive and progressed tumor has the higher risk of recurrence to obtain the full identification of the parameter estimates for tumor grades. Under this assumption, the order restriction is imposed on the Cox regression parameters:  $0 \leq \beta_1 \leq \beta_2$ .

$$\lambda(t_i|Z_i, \tilde{W}_i) = \lambda_0(t_i) \exp(\boldsymbol{\beta} \tilde{W}_i + \gamma Z_i),$$

where  $\tilde{W}_i = (W_{i1}, W_{i2}) = (I(W_i = 2), I(W_i = 3))$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  in the case of tumor grade and  $\lambda_0(t_i)$  represents the baseline hazard function for a subject  $i$ . To test the order restriction on the Cox regression parameters, the Union-Intersection (UI) principle can be considered. Under the null hypothesis  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  and the alternative hypothesis  $H_1 : 0 \leq \beta_1 \leq \beta_2$ , we adapt the UI principle for the Cox proportional hazards model with  $H_1 : \boldsymbol{\beta} \geq \mathbf{0}$  proposed by Sen [25].

Instead of building partial likelihood scores as in Sen [25], we use the profile likelihood scores with treating the parameters  $\phi = (\lambda_0(t_i), \boldsymbol{\pi}, \mathbf{q})$  as nuisance parameters, where  $(\boldsymbol{\pi}, \mathbf{q})$  are the prevalences and classification rates in the sub-model for multiple ratings and  $\lambda_0(t_i)$  are the baseline hazards in the sub-model for time-to-event data. If we consider the profile likelihood such as  $\log L_{PL}(\boldsymbol{\beta})$ , the additional iterative procedures along with the survey-weighted cox model in the M-steps are required to update  $\gamma$ . However, with the profile likelihood  $\log L_{PL}(\boldsymbol{\beta}, \gamma)$ , the survey-weighted Cox model in the M-steps is enough to update the nuisance parameters  $\phi = (\lambda_0(t_i), \boldsymbol{\pi}, \mathbf{q})$  while  $(\boldsymbol{\beta}, \gamma)$  are fixed. Hence, we can expect that the use of profile likelihood  $\log L_{PL}(\boldsymbol{\beta}, \gamma)$  can reduce the programming demands for the proposed joint model.

Denote  $\hat{\gamma}_N^0$ , which is the parameter estimate of  $\gamma$  under the null hypothesis and it can be derived by fitting a cox model with only including an auxiliary variable, such as tumor sizes. The profile likelihood score for  $\boldsymbol{\beta}$  under  $H_0$  and the variance-covariance matrix can be defined by

$$\hat{U}_N = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_{PL}(\boldsymbol{\beta}, \gamma) |_{\boldsymbol{\beta}=\mathbf{0}, \gamma=\hat{\gamma}_N^0} \quad (4.7)$$

$$V_{11.2} = - \left[ \begin{array}{cc} \frac{\partial^2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{\partial \boldsymbol{\beta} \partial \gamma} \\ \frac{\partial^2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{\partial \gamma^2} & \end{array} \right]_{\boldsymbol{\beta}=\mathbf{0}, \gamma=\hat{\gamma}_0} = \left[ \begin{array}{cc} \bar{V}_{11} & \bar{V}_{12} \\ \bar{V}_{21} & \bar{V}_{22} \end{array} \right]. \quad (4.8)$$

Then  $\bar{V}_{11.2}$  can be calculated by the sub-blocks of  $V_{11.2}$  with the following equation:

$$\bar{V}_{11.2} = \bar{V}_{11} - \bar{V}_{12} \bar{V}_{22}^{-1} \bar{V}_{21}. \quad (4.9)$$

The first and second derivatives in the equations (4.7) and (4.8) are approximated by a numerical differentiation. With  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  and  $h = 0.001$ ,  $\mathbf{h}_1 = h * (1, 0)$ ,  $\mathbf{h}_2 = h * (0, 1)$ ,  $\mathbf{h}_3 = h * (1, 1)$ ,  $\mathbf{h}_4 = h * (-1, 1)$ ,  $\mathbf{h}_5 = h * (1, -1)$ , we use the following numerical differentiation to calculate the first and second order partial derivatives, where  $\log L_{PL} = \log L_{PL}(\boldsymbol{\beta}, \gamma)$ .

$$\begin{aligned} \frac{\partial \log L_{PL}}{\partial \beta_1} &= \frac{\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_1, \gamma) - \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_1, \gamma)}{2h} + o(h) \\ \frac{\partial \log L_{PL}}{\partial \beta_2} &= \frac{\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_2, \gamma) - \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_2, \gamma)}{2h} + o(h) \\ \frac{\partial \log L_{PL}}{\partial \gamma} &= \frac{\log L_{PL}(\boldsymbol{\beta}, \gamma + h) - \log L_{PL}(\boldsymbol{\beta}, \gamma - h)}{2h} + o(h) \\ \frac{\partial^2 \log L_{PL}}{\partial \beta_1^2} &= \frac{\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_1, \gamma) + \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_1, \gamma) - 2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{h^2} + o(h) \\ \frac{\partial^2 \log L_{PL}}{\partial \beta_2^2} &= \frac{\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_2, \gamma) + \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_2, \gamma) - 2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{h^2} + o(h) \\ \frac{\partial^2 \log L_{PL}}{\partial \gamma^2} &= \frac{\log L_{PL}(\boldsymbol{\beta}, \gamma + h) + \log L_{PL}(\boldsymbol{\beta}, \gamma - h) - 2 \log L_{PL}(\boldsymbol{\beta}, \gamma)}{h^2} + o(h) \end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 \log L_{PL}}{\partial \beta_1 \partial \gamma} &= \frac{1}{4h^2} [\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_1, \gamma + h) + \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_1, \gamma - h) \\
&\quad - \log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_1, \gamma - h) - \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_1, \gamma + h)] + o(h) \\
\frac{\partial^2 \log L_{PL}}{\partial \beta_1 \partial \beta_2} &= \frac{1}{4h^2} [\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_3, \gamma) + \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_3, \gamma) \\
&\quad - \log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_4, \gamma) - \log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_5, \gamma)] + o(h) \\
\frac{\partial^2 \log L_{PL}}{\partial \beta_2 \partial \gamma} &= \frac{1}{4h^2} [\log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_2, \gamma + h) + \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_2, \gamma - h) \\
&\quad - \log L_{PL}(\boldsymbol{\beta} + \mathbf{h}_2, \gamma - h) - \log L_{PL}(\boldsymbol{\beta} - \mathbf{h}_2, \gamma + h)] + o(h).
\end{aligned}$$

With  $\hat{U}_N$  and  $\bar{V}_{11.2}$ ,  $\hat{U}_{N(\mathbf{a})}^*$  and  $\bar{V}_{11.2(\mathbf{a})}^*$  are computed for any subset  $\mathbf{a}$  of  $P = \{1, \dots, p\}$ , where  $p$  is the dimension of  $\hat{U}_N$ .

$$\hat{U}_{N(\mathbf{a})}^* = \hat{U}_{N(\mathbf{a})} - \bar{V}_{11.2(\mathbf{a}\bar{\mathbf{a}})} \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \quad (4.10)$$

$$\bar{V}_{11.2(\mathbf{a})}^* = \bar{V}_{11.2(\mathbf{a}\mathbf{a})} - \bar{V}_{11.2(\mathbf{a}\bar{\mathbf{a}})} \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \bar{V}_{11.2(\bar{\mathbf{a}}\mathbf{a})}. \quad (4.11)$$

Next, the below conditions for each subset  $\mathbf{a}$  and the complementary subset  $\bar{\mathbf{a}}$  are evaluated.

$$\hat{U}_{N(\mathbf{a})}^* > 0 \quad \bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0$$

Without loss of generality, we assume that  $p = 2$ , where  $\boldsymbol{\beta} = (\beta_1, \beta_2)$  in the proposed joint model. Then, any subset of  $p$  is  $P = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ . The first condition is always met for a null set and the second condition is always met for  $P = \{1, 2\}$ . Among the subsets, only one element (for the proposed joint model,  $\{1\}$  or  $\{2\}$ ) satisfies both conditions. Once the element is found, the Union-Intersection statistic can be computed, where  $I(\hat{U}_{N(\mathbf{a})}^* > 0)$  and  $I(\bar{V}_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0)$

$$T_N^{(1)} = (\hat{U}_{N(\mathbf{a})}^* \bar{V}_{11.2(\mathbf{a})}^{*-1} \hat{U}_{N(\mathbf{a})}^*)^{1/2} = t. \quad (4.12)$$

Under the null hypothesis,  $\hat{U}_{N(\mathbf{a})}^* \sim N(\mathbf{0}, v_{11.2(\mathbf{a})}^*)$ ,  $\bar{V}_{(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \sim N(\mathbf{0}, v_{11.2(\bar{\mathbf{a}}\bar{\mathbf{a}})}^{-1})$ . Hence, the distribution of the UI statistic  $T_N^{(1)}$  is

Table 5: Weights for the mixture distribution for the UI-test statistic

$\mathbf{a}$	$k$	$Pr\{\hat{U}_{N(\mathbf{a})}^* > 0\}$	$Pr\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1}\hat{U}_{N(\bar{\mathbf{a}})} \leq 0\}$	$\omega$
$\emptyset$	0	1	$p_1$	$\omega_0$
$\{1\}$	1	1/2	1/2	1/4
$\{2\}$	1	1/2	1/2	1/4
$\{1, 2\}$	2	$p_2$	1	$\omega_2$

$$\begin{aligned}
 P\left\{T_N^{(1)} \leq t | H_0\right\} &= \sum_{\emptyset \subseteq \mathbf{a} \subseteq P} P\left\{\hat{U}_{N(\mathbf{a})}^* > 0\right\} P\left\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1}\hat{U}_{N(\bar{\mathbf{a}})} \leq 0\right\} P\left\{\chi_k \leq t\right\} \\
 &= \sum_{k=0}^p \omega_k P\left\{\chi_k \leq t\right\}, \forall x \geq 0, \sum_{k=1}^p \omega_k = 1,
 \end{aligned} \tag{4.13}$$

Here,  $\chi_k^2$  is a chi-square random variable, with  $k$  is the number of elements in  $\mathbf{a}$ .  $\chi_0^2$  is equal to zero with probability 1. The weight  $\omega_k$  can be computed by the product of two normal orthant probabilities and be written as

$$\omega_k = \sum_{\dim(\mathbf{a})=k} Pr\{\hat{U}_{N(\mathbf{a})}^* > 0 | H_0\} Pr\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1}\hat{U}_{N(\bar{\mathbf{a}})} \leq 0 | H_0\}. \tag{4.14}$$

With  $P = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ , the weights for each  $k$  are calculated as Table 5. In Table 5,  $p_1 = Pr\{\hat{U}_{N(\mathbf{a})}^* > 0\}$  with  $\mathbf{a} = \{1, 2\}$  and  $p_2 = Pr\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1}\hat{U}_{N(\bar{\mathbf{a}})} \leq 0\}$  with  $\bar{\mathbf{a}} = \{1, 2\}$  are evaluated by using a numerical integration, where  $\hat{U}_{N(\mathbf{a})}^* \sim N(0, v_{11.2(\mathbf{a})})$  and  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1}\hat{U}_{N(\bar{\mathbf{a}})} \sim N(0, \bar{V}_{11.2(\bar{\mathbf{a}})}^{-1})$ . R function `pmvnorm` in R package `mvtnorm` can be used to calculate the multi-dimensional normal orthant probabilities. For example, the two dimensional orthant probabilities can be computed by the below statement.

```

p1 <- pmvnorm(lower = c(l1,l2) ,upper = c(u1,u2) ,
+ mean=rep(0, 2), sigma = V11.2, algorithm = Miwa())

```

- lower: Specify the lower limit of the integral region.
- upper: Specify the upper limit of the integral region.
- mean: Specify the mean vector.
- sigma: Specify the variance-covariance matrix.
- algorithm: Select the algorithm for a numerical integration.

If the limit of integral region is infinity, we specify small or large value to the lower/upper limit (e.g. lower = c(0,0), upper = c(10000, 10000)). The details can be found in the document for R package mvtnorm '<http://cran.r-project.org/web/packages/mvtnorm/mvtnorm.pdf>'.

Then, the P-value is calculated as

$$\text{P-value} = 1 - \{P(\chi_0^2 \leq t) * \omega_0 + P(\chi_1^2 \leq t) * \omega_1 + P(\chi_1^2 \leq t) * \omega_1 + P(\chi_2^2 \leq t) * \omega_2\},$$

where  $t$  is the value of test statistic in the equation (4.12).

For testing the null hypothesis  $H_0 : \boldsymbol{\beta} = \mathbf{0}$  vs.  $H_1 : 0 \leq \beta_1 \leq \beta_2$ , the reparameterization on  $\boldsymbol{\beta}$  is needed. Let  $\lambda_j = \beta_j - \beta_{j-1}$  with  $\beta_0 = 0$ , then the alternative hypothesis will be reduced to  $H_1 : \boldsymbol{\lambda} \geq \mathbf{0}$ , where  $\lambda_1 = \beta_1$  and  $\lambda_2 = \beta_2 - \beta_1$ . Now, the profile likelihood score  $\hat{U}_N$  (Eq.4.7) and the variance-covariance matrix  $\bar{V}_{11.2}$  (Eq. 4.8) can be updated with respect to  $\boldsymbol{\lambda}$  by using a chain rule. Once the profile likelihood score and the variance-covariance matrix are updated, the equations (4.10) and (4.11) are calculated with the values from the equations (4.15) and (4.16). The test statistics (Eq.4.12) and p-value can also be calculated in the same way. The profile likelihood score  $\hat{U}_N$  for  $\boldsymbol{\lambda}$  can be calculated by

$$\hat{U}_N^* = \frac{\partial}{\partial \boldsymbol{\lambda}} \log L_{PL}(\boldsymbol{\lambda}, \gamma)|_{\boldsymbol{\lambda}=\mathbf{0}, \gamma=\hat{\gamma}_N^0} = \frac{\partial}{\partial \boldsymbol{\beta}} \log L_{PL}(\boldsymbol{\beta}, \gamma) \frac{\partial}{\partial \boldsymbol{\lambda}} \boldsymbol{\beta}(\boldsymbol{\lambda}) \quad (4.15)$$

The second derivatives for the variance-covariance matrix  $\bar{V}_{11.2}^*$  can be calculated by

$$\frac{\partial^2}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}^T} \log L_{PL}(\boldsymbol{\lambda}, \gamma)|_{\boldsymbol{\lambda}=\mathbf{0}, \gamma=\hat{\gamma}_N^0} = \frac{\partial}{\partial \boldsymbol{\lambda}} \boldsymbol{\beta}(\boldsymbol{\lambda}) \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \log L_{PL}(\boldsymbol{\beta}, \gamma) \frac{\partial}{\partial \boldsymbol{\lambda}} \boldsymbol{\beta}(\boldsymbol{\lambda})^T, \quad (4.16)$$

where

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \boldsymbol{\beta}(\boldsymbol{\lambda}) = \begin{bmatrix} \frac{\partial \beta_1}{\partial \lambda_1} & \frac{\partial \beta_1}{\partial \lambda_2} \\ \frac{\partial \beta_2}{\partial \lambda_1} & \frac{\partial \beta_2}{\partial \lambda_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

## 5.0 SIMULATION STUDIES

We conducted simulation studies to examine the statistical properties for the proposed joint model in Chapter 3. To confirm the practical utility of our methods, we considered simulation setups that are similar to the National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14 sub-study. Based on the structure of the NSABP B-14 dataset, we simulated  $n$  subjects with a total seven variables  $D_i = \{W_i, V_i^{(1)}, V_i^{(2)}, V_i^{(3)}, X_i, \Delta_i, Z_i\}$  for each subject  $i$ , where

- $W_i$ : True tumor grade, where  $W_i = j, j = 1, 2, 3$
- $(V_i^{(1)}, V_i^{(2)}, V_i^{(3)})$ : Tumor grade ratings from three pathologists, where  $V_i^{(s)} = k, k = 1, 2, 3; s = 1, 2, 3$
- $X_i$ : Observed failure time
- $\Delta_i$ : Observed censoring indicator, 1 for failure event, 0 for censoring event
- $Z_i$ : Tumor size

Tumor grade ratings  $(V_i^{(1)}, V_i^{(2)}, V_i^{(3)})$  from three pathologists, true tumor grade  $W_i$  and tumor size  $Z_i$  for each subject  $i$  were simulated such that:

- (1) Simulate true tumor grades ranged from 1 to 3 for  $n$  subjects based on a multinomial distribution with the prevalence  $(\pi_1, \pi_2, \pi_3) = (0.3, 0.5, 0.2)$ . With the R function `rMultinom(rbind(0.3, 0.5, 0.2), n)`, generate  $W_i$ . The value of  $W_i$  is 1, 2, or 3.
- (2) Generate tumor grade ratings from three pathologists given a simulated true tumor grade for each subject  $i$ . Tumor grade ratings for each pathologist are independently generated given the level of true tumor grade generated at (1). A multinomial distribution with the classification rates  $\hat{q}_{jk}^{(s)}$ ;  $j, k, s = 1, 2, 3$  in Table 6 are used to generate tumor grade ratings. For example, if  $W_i = 2$ , we would generate  $V_i^{(s)}$  from a multinomial distribution

Table 6: Simulation setup: Classification rates for simulation studies

True grade	Rater#1			Rater#2			Rater#3		
	W	M	P	W	M	P	W	M	P
W	0.6	0.3	0.1	0.5	0.35	0.15	0.6	0.15	0.25
M	0.07	0.87	0.06	0.2	0.7	0.1	0.04	0.91	0.05
P	0.05	0.15	0.8	0.03	0.23	0.74	0.1	0.3	0.6

with the classification rates  $(q_{21}^{(s)}, q_{22}^{(s)}, q_{23}^{(s)})$ . Let  $n_2 = \sum I(W_i = 2)$ , then  $V_i^{(2)}$  given  $W_i = 2$  are generated with the R function  $rMultinom(rbind(0.2, 0.7, 0.1), n_2)$

- (3) Given a true tumor grade  $W_i$  for each subject  $i$ , generate tumor sizes from three log-normal distributions with the mean and variance estimates from the B-14 dataset:  $\log(Z_i) \sim N(\mu_i, \sigma^2)$ , where  $\mu_i = (0.54, 0.65, 0.81)$  for  $W_i = 1, 2$ , and 3, respectively, and common variance  $\sigma^2 = 0.29$  for all grades. For example, if  $W_i = 2$ , generate  $Z_i^* = \log(Z_i)$  with the R function  $rnorm(n_2, 0.65, 0.29)$  and take an exponential of  $Z_i^*$  so that  $Z_i = \exp(Z_i^*)$ .

With the true regression coefficient  $\{\gamma, \beta_1, \beta_2\} = \{0.3, 0.5, 2.0\}$ , the survival time  $T_i$  was generated from a Cox proportional hazards model with constant baseline hazard  $\lambda_0(t) \equiv \lambda_0$ , such that

$$\lambda(t|Z_i, W_i) = \lambda_0 \exp\{0.3Z_i + 0.5I(W_i = 2) + 2.0I(W_i = 3)\},$$

where  $I(W_i = j), j = 2, 3$ .

The censoring time  $C_i$  was generated from an exponential distribution with rate  $\lambda_C$  with assuming that the censoring is non-informative. We considered a simulation setup for the sample size 668 and 1300 with 20% or 80% censoring rates, by setting  $(\lambda_0, \lambda_c)$  to  $(0.045, 0.035)$  or  $(0.01, 0.15)$ , respectively. The observed failure time  $X_i$  were simulated such that:

- (1) For  $i = 1, 2, \dots, n$ , generate  $u_{i1}, u_{i2}$  from a uniform distribution  $U(0, 1)$ , independently.

- (2) With the simulated true tumor grade and the tumor size given the true tumor grade, generate a failure time  $T_i$  for  $n$  subjects by

$$T_i = \frac{-\log(u_{i1})}{\lambda_0 \exp\{0.3Z_i + 0.5I(W_i = 2) + 2.0I(W_i = 3)\}}.$$

- (3) Generate censoring time  $c_i$  for  $n$  subjects by

$$c_i = \frac{-\log(u_{i2})}{\lambda_c}.$$

- (4) Generate time-to-event data and censoring indicator. The observed time-to-event data are  $X_i = \min\{C_i, T_i\}$  and the censoring indicator is  $\Delta_i = I(T_i \leq C_i)$ .

Considering each of four scenarios with different sample sizes  $n = (668, 1300)$  and censoring proportions  $\Delta\% = (20, 80)$ , we conducted 500 replications with different seed numbers. For the  $s$ th replication,  $s = 1, 2, \dots, S$ , we set the initial values for the parameter estimates  $(\boldsymbol{\pi}, \mathbf{q}, \gamma, \beta_1, \beta_2)$  for running the EM algorithm in Section 3.2. To get the initial values, we fitted two sub-models separately. The parameter estimates for  $(\boldsymbol{\pi}, \mathbf{q})$  from the equation (2.1) were considered as the initial values for  $(\boldsymbol{\pi}, \mathbf{q})$  in the EM algorithm in Section 3.2. The Cox regression parameters fitted with including true tumor grade and tumor size were considered as the initial values for  $(\gamma, \beta_1, \beta_2)$  in the EM algorithm. With the initial values, the parameter estimates and the standard errors of the Cox model parameters were computed as described in Chapter 3. Due to the intensive computation, we run 20 replications per a seed number. Hence, total 25 seed numbers for 500 replications were generated by the R function *.Random.seed* for the simulation studies.

The results with 500 replications for the sample size 668 and 1300 are shown in Table 7. The empirical bias (Bias), the empirical standard deviations of the estimates over all replications (Emp.SD), the average of estimated standard errors (ASE), and the coverage probabilities of 95% confidence interval are reported to examine the statistical properties for the proposed joint model. Let  $\hat{\Omega}^{(t)} = (\hat{\gamma}^{(t)}, \hat{\beta}_1^{(t)}, \hat{\beta}_2^{(t)})$ ,  $t = 1, 2, \dots, 500$  be the estimates for  $t$ th simulated data, and  $\Omega = \{\gamma, \beta_1, \beta_2\} = \{0.3, 0.5, 2.0\}$  be the true regression parameters. For example, the simulation parameters for  $\gamma$  are computed as below:

## Empirical Bias

$$\text{Emp.bias} = \frac{1}{500} \sum_{t=1}^{500} \hat{\gamma}^{(t)} - \gamma$$

## Empirical Standard Deviation

$$\text{Emp.SD} = \frac{1}{499} \sum_{t=1}^{500} (\hat{\gamma}^{(t)} - \bar{\gamma})^2,$$

where  $\bar{\gamma} = \frac{1}{500} \sum_{t=1}^{500} \hat{\gamma}^{(t)}$ .

## Average of estimated standard errors

$$\text{ASE} = \frac{1}{500} \sum_{t=1}^{500} \widehat{S}(\hat{\gamma}^{(t)}),$$

where  $\widehat{S}(\hat{\gamma}^{(t)})$  is the estimated standard error for  $t$ th simulated dataset.

## Coverage probability of 95% confidence interval

$$\text{CP} = \frac{1}{500} \sum_{t=1}^{500} I(\hat{\gamma}^{(t)} \in [\gamma - 1.96 * S_{\hat{\gamma}}^{(t)}, \gamma + 1.96 * S_{\hat{\gamma}}^{(t)}]).$$

Instead of imposing the order restrictions on the parameter estimates in the M-steps, we incorporated the order restrictions when we interpreted the parameter estimates at the convergences of the EM algorithm. For each case, we switched the labels on  $\beta$  by re-parameterizing  $(\beta_1, \beta_2)$  so that the order restrictions are preserved.

- (1) If  $\beta_1, \beta_2 > 0$  but  $\beta_2 < \beta_1$ , then we set  $\beta'_1 = \beta_2$  and  $\beta'_2 = \beta_1$ .
- (2) If  $\beta_1 < 0$  and  $\beta_2 > 0$ , we set  $\beta'_1 = -\beta_1$  and  $\beta'_2 = -\beta_1 + \beta_2$ .

In the same time, the labels on the  $(\pi, \mathbf{q})$  were also re-arranged so that  $0 < \widehat{\beta}_1 < \widehat{\beta}_2$ . For example,

- (1) If  $0 > \beta_1 > \beta_2$ , the labels on  $\phi = \{\pi_1, \pi_2, \pi_3, q_{1k}^{(s)}, q_{2k}^{(s)}, q_{3k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$  are switched to  $\phi' = \{\pi_3, \pi_2, \pi_1, q_{3k}^{(s)}, q_{2k}^{(s)}, q_{1k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$ , where  $\sum_{j=1}^3 \pi_j = \sum_{k=1}^3 q_{jk}^{(s)} = 1$ .

Table 7: Simulation results with  $(\gamma, \beta_1, \beta_2) = (0.3, 0.5, 2.0)$ .

% censoring	n	parms	$N = 500$			
			Bias	Emp.SD	ASE	CP
20%	668	$\gamma$	0.026	0.038	0.021	0.779
		$\beta_1$	-0.007	0.149	0.095	0.923
		$\beta_2$	-0.004	0.196	0.121	0.913
	1300	$\gamma$	0.023	0.028	0.02	0.700
		$\beta_1$	-0.005	0.103	0.095	0.924
		$\beta_2$	0.004	0.130	0.122	0.923
80%	668	$\gamma$	0.009	0.058	0.06	0.943
		$\beta_1$	0.069	0.361	0.272	0.896
		$\beta_2$	0.036	0.355	0.266	0.876
	1300	$\gamma$	0.013	0.043	0.044	0.928
		$\beta_1$	-0.011	0.236	0.199	0.916
		$\beta_2$	-0.018	0.234	0.193	0.904

(2) If  $0 < \beta_2 < \beta_1$ , the labels on  $\phi$  are switched to  $\phi' = \{\pi_1, \pi_3, \pi_2, q_{1k}^{(s)}, q_{3k}^{(s)}, q_{2k}^{(s)}; k = 1, 2, \dots, K; s = 1, 2, \dots, S\}$ , where  $\sum_{j=1}^3 \pi_j = \sum_{k=1}^3 q_{jk}^{(s)} = 1$ .

In Table 7, the simulation results for 20% censoring show that the proposed joint model provides the parameter estimates for  $\beta$  with negligible biases. Also, no significant difference between the empirical standard deviations and the average of estimated standard errors for  $\beta$  is detected for 20% censoring. The coverage probabilities of 95% confidence interval for the  $\beta$  are close to the nominal level of 0.95. However, the simulation results for 20% censoring show that the parameter estimate for  $\gamma$  is biased and the coverage probability is not close to the nominal level.

When the sets of weights include extreme values, such as 0, 1, the current R package `survey` may not converge well. Although the computation efforts can be reduced by using the R package `survey` to update the parameters having no closed-form of solutions, the systematic biases on the parameters may be expected through the estimation procedures. For further checking procedure, we will compare the current simulation results with the results from the EM algorithm with a one-step Newton-Raphson method. As the censoring proportion is increased from 20% to 80%, the biases for  $(\beta_1, \beta_2)$  are also increased and the coverage probability is not close to the nominal level. Based on our knowledge, if the association between the unknown true status and the time of disease recurrence is not strong enough, the estimates in a finite sample can be overestimated due to the "label swapping" on the model parameters. For example, when we generate a finite sample  $(X_1, X_2)$ , where  $X_1 \sim N(\mu_1, 1), X_2 \sim N(\mu_2, 1)$  with imposing the order restriction on  $\mu, \mu_1 > \mu_2$ , the sample means of  $X_2, \bar{X}_2$  can be larger than  $\bar{X}_1$  due to the "label switching" phenomenon if the magnitude of the difference between  $\mu_1$  and  $\mu_2$  is not large enough.

We expect the bias, relative to the empirical standard deviation, will diminish as sample size increases and the gap between the  $\beta_1$  and  $\beta_2$  increases. As censoring proportion decreases, we expect the bias will also decrease. Because of the numerical approximation in variance estimation in Section 3.3, the average of standard error estimates could be slightly different from the empirical standard deviations of the estimates.

## 6.0 APPLICATION: ANALYSIS OF DATA FROM THE NSABP B-14 DATA

The National Surgical Adjuvant Breast and Bowel Project (NSABP) B-14 trial is a phase 3 randomized controlled trial to investigate the prognosis of patients with node-negative, estrogen-receptor-positive breast cancer treated with Tamoxifen and/or chemotherapy. We applied the proposed joint model to a sub-study of the NSABP B-14 trial with disease-free survival as the endpoint or outcome of interest. With the tissue samples and clinical information of patients enrolled in B-14 trial, the validation study of a 21-gene panel assay with comparing classical clinical factors such as age, tumor size, and tumor grade was carried out [20]. With a patients paraffin-embedded tumor tissue, the tumor grades for a patient were independently measured by three different pathologists from the NSABP, Stanford University Medical Center, and the University of California at San Francisco, School of Medicine, respectively. All three experienced pathologists were blinded from any clinical information and independently evaluated the tumor grade following the modified Bloom-Richardson grading criteria [20]. According to the modified Bloom-Richardson grading system, the pathologists scored three features: the tubule formation, the count of cell mitosis, and uniformity in cell size, shape and staining character of the nuclei. Each of these features is scored from 1 to 3. Once all three features in Table 1 are scored, the total sum of three scores are computed and categorized into one of three grade (Table 2). A tumor is considered as a well, moderately, poorly differentiated tumor if the total sum of score is 3 to 5, 6 to 7, or 8 to 9, respectively [9].

The data analysis of this exemplary dataset was programmed in R. The survey-weighted Cox model in the M-step was applied by using the R functions **svydesign** and **svycoxph** in the R package **survey**. The weights were proportional to  $E[W_i = j | D_{obs}; \theta^{(t)}]$  for a subject  $i$  and a level of true tumor grade  $j$ , which was calculated in the E-step. The details are shown in Section 3.2.2. The standard errors of cox parameter estimates were derived by

Table 8: Estimated prevalences and classification rates from the proposed joint model

True grade	$\pi$	Rater#1			Rater#2			Rater#3		
		W	M	P	W	M	P	W	M	P
W	0.28	0.44	0.55	0.01	0.87	0.13	0	0.62	0.37	0
M	0.49	0.07	0.84	0.08	0.17	0.74	0.09	0.05	0.90	0.05
P	0.23	0	0.22	0.78	0.02	0.20	0.78	0.01	0.46	0.55

using the profile likelihood approach proposed by Murphy and van der Vaart [19]. With the derived standard errors from the profile likelihood, the p-values for the parameters of interest were calculated by the Wald test and the Union-Intersection Principle in Section 4.3. In the analysis, we used tumor size for each subject as an auxiliary information and included tumor grade readings from three pathologists with considering a well-differentiated tumor as the reference category in the sub-model for time-to event data (Equation 3.3). Accordingly, tumor size was used as one of covariates in the sub-model for time-to-event data and tumor grade readings were used in the sub-model for multiple ratings with an unknown true tumor grade.

Since three pathologists at three different sites in the U.S. independently assessed tumor grades for 668 subjects in the NSABP B-14 study, it is reasonable to assume that three tumor grade readings are independent each other given a true tumor grade. In addition, any clinical information such as tumor size was not disclosed to all pathologists during tumor grade reading. Hence, we can also assume that the auxiliary information is independent of tumor grade readings given the underlying true tumor grade.

Table 4 in Section 2.1.3 shows the estimates of prevalence and classification rates from the latent class model proposed by Dawid and Skene [7]. The assumption of conditional independence on the tumor grade readings is met with the p-value of the Pearson Chi-square goodness-of-fit test at 0.46 ( $\chi_6^2 = 5.7$ ). Table 8 shows the estimates of prevalence and classification rates from the proposed joint model. Both tables show very similar estimates of prevalence and classification rates.

Table 9: Parameter estimates and standard errors of Cox parameter estimates from the joint model.

	Cox model only						Joint model	
	Pathologist 1		Pathologist 2		Pathologist 3		Est.(SE)	P-value
	Est.(SE)	P-value	Est.(SE)	P-value	Est.(SE)	P-value		
$\hat{\gamma}$	0.18 (0.05)	<.001	0.18 (0.05)	<.001	0.12 (0.05)	0.01	0.16 (0.04)	<.001
$\hat{\beta}_1$	0.30 (0.35)	0.39	0.62 (0.29)	0.03	0.52 (0.34)	0.13	0.55 (0.24)	<.001
$\hat{\beta}_2$	1.21 (0.35)	<.001	1.64 (0.28)	<0.001	1.86 (0.36)	<.001	1.86 (0.17)	<.001

As a naive approach, each of the three Cox models includes tumor grade readings from one of three pathologists and tumor size as the predictors. We compare the degree of association between tumor grade and time to breast cancer recurrence determined from the three Cox models and the proposed joint model. Table 9 summarizes the parameter estimates and standard errors under three Cox models and the joint model. The estimated regression coefficients of tumor size  $\hat{\gamma}$  and the corresponding standard errors are quiet similar across all four models. However, we notice large differences in the estimates of regression coefficients for tumor grades between three naive models and the joint model ( $\hat{\beta}_1$  for grade 2 vs. grade 1 ;  $\hat{\beta}_2$  for grade 3 vs. grade 1). Compared to the estimates from the joint model, the naive Cox models can lead to biased results for the association between tumor grades and the risk of breast cancer recurrence. For example, when we use tumor grade readings from the first pathologist, the degree of association between tumor grades and the risk of breast cancer recurrence can be underestimated ( $\hat{\beta}_1^{(1)} = 0.3$ ,  $\hat{\beta}_2^{(1)} = 1.21$  from the naive Cox model;  $\hat{\beta}_1 = 0.55$ ,  $\hat{\beta}_2 = 1.86$  from the joint model). Comparing to the standard errors from the naive cox models, the joint model provides more efficient estimates for the association between tumor grades and time to breast cancer recurrence. The three naive Cox models only use tumor grade readings from one of three pathologists. On the other hand, the joint model combines all available tumor grade readings from three pathologists, so that it estimates the Cox parameters with larger sample size.

Table 10: The conditions for calculating the U-I test statistic for the NSABP B-14 data.

$\mathbf{a}$	$\bar{\mathbf{a}}$	$\hat{U}_{N(\mathbf{a})}^*$	$\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})}$	$I(\hat{U}_{N(\mathbf{a})}^* > 0)$	$I(\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0)$
$\emptyset$	$\{1, 2\}$	-	$\begin{bmatrix} -2.44 \\ 1.86 \end{bmatrix}$	1	0
$\{1\}$	$\{2\}$	-25.65	0.23	0	0
$\{2\}$	$\{1\}$	28.27	-0.63	1	1
$\{1, 2\}$	$\emptyset$	$\begin{bmatrix} -2.437 \\ 1.861 \end{bmatrix}$	-	0	1

Since the p-values for  $H_0 : \boldsymbol{\beta} = (\beta_1, \beta_2)' = \mathbf{0}$  in Table 9 are less than 0.05, we tested the order restrictions on  $\boldsymbol{\beta}$ , which is  $0 \leq \beta_1 \leq \beta_2$ . The Union-Intersection principle for the proposed joint model in Section 4.3 was conducted with the re-parameterization on  $\boldsymbol{\beta}$ . With  $\lambda_j = \beta_j - \beta_{j-1}$ , where  $\beta_0 = 0$ , the alternative hypothesis was reduced to  $H_1 : \boldsymbol{\lambda} \geq \mathbf{0}$ , where  $\lambda_1 = \beta_1$  and  $\lambda_2 = \beta_2 - \beta_1$ .

For each subset  $\mathbf{a}$  of  $\mathbf{P} = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ , we first evaluated the two conditions:  $\hat{U}_{N(\mathbf{a})}^* > 0$  and  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0$ , where  $\hat{U}_{N(\mathbf{a})}^* \sim N(0, v_{11.2(\mathbf{a})})$ ,  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \sim N(0, \bar{V}_{11.2(\bar{\mathbf{a}})})$ , which are defined in Eq.4.10 and Eq.4.11 and the chain rule for re-parameterization in Eq.4.15 and Eq.4.16. Table 10 shows the values of two conditions for each partition  $\mathbf{a}$ . The value of  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})}$  is always greater than zero when  $\mathbf{a}$  is the empty set. On the other hand,  $\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})}$  is always equal to or less than less than zero, when  $\bar{\mathbf{a}}$  is the empty set. Among a total  $2^2 = 4$  partitions, only one partition will have both  $I(\hat{U}_{N(\mathbf{a})}^* > 0)$  and  $I(\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0)$  equal to one. The U-I test statistic 7.252 was calculated on the term for which both indicators are one,  $\mathbf{a} = 2, \bar{\mathbf{a}} = 1$ .

$$T_N = (\hat{U}_{N(2)}^* \bar{V}_{11.2(2)}^{*-1} \hat{U}_{N(2)}^*)^{1/2} = 7.252.$$

The weights for a mixture chi-bar distribution for each partition are shown in Table 11. For  $\mathbf{a} = \emptyset$  and  $\mathbf{a} = \{1, 2\}$ , the weights were calculated by the R package `mvtnorm` with the

Table 11: Estimated weights for the mixture distribution for the U-I test statistic

$\mathbf{a}$	$k$	$Pr\{\hat{U}_{N(\mathbf{a})}^* > 0\}$	$Pr\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0\}$	$\omega$
$\emptyset$	0	1	0.399	0.399
$\{1\}$	1	0.5	0.5	0.25
$\{2\}$	1	0.5	0.5	0.25
$\{1, 2\}$	2	0.101	1	0.101

upper limits of integral region  $c = (10000, 10000)$  and the lower limits of integral region  $c = (0, 0)$ . For example, the weights  $\omega_0$  and  $\omega_2$  represent

$$\begin{aligned} \omega_0 &= Pr\{\hat{U}_{N(\mathbf{a})}^* > 0\} + Pr\{\bar{V}_{11.2(\bar{\mathbf{a}})}^{-1} \hat{U}_{N(\bar{\mathbf{a}})} \leq 0\} \\ &= 1 + Pr\{\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2) \leq 0 : \boldsymbol{\epsilon} \sim N(0, v_{11.2(\bar{\mathbf{a}})})\} \\ \omega_2 &= Pr\{\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2) \leq 0 : \boldsymbol{\epsilon} \sim N(0, v_{11.2(\mathbf{a})})\} + 1. \end{aligned}$$

The P-value with  $t = 7.252$  and the weights  $\omega_k, k = 0, 1, 2$  is

$$\begin{aligned} \text{P-value} &= 1 - \{P(\chi_0^2 \leq t) * \omega_0 + P(\chi_1^2 \leq t) * \omega_1 + P(\chi_2^2 \leq t) * \omega_2 + P(\chi_3^2 \leq t) * \omega_3\} \\ &= 0.0062, \end{aligned}$$

where  $P(\chi_0^2 \leq t) = 1$ .

## 7.0 CONCLUSION

The main objective of the proposed method is to achieve global identification of model parameters in the latent class models for modeling discrete diagnostic tests without a gold standard. The global identification becomes possible when there is a known trend between the underlying truth and the risk of an event of interest. In addition, the proposed joint modeling approach enables us to utilize all ratings from multiple independent ratings to provide an accurate assessment of the association between the unknown true status and the time to event of interest. In our motivating example, the true tumor grade for patients was inaccessible due to the subjective nature of the tumor grading system. The critical assumption of conditional independence among multiple ratings is more likely to be true because the three pathologists evaluated patients' tumor grade independently and none of them was aware of any clinical information related to the tissue samples. The result from the Pearson chi-square goodness-of-fit test also concurs with its p-value larger than 0.05.

In some other circumstances, however, the conditional independence assumption may be violated. Under the violation of conditional independence, the estimates from a misspecified model will be biased as discussed in Section 2.1.4 [29, 32]. In addition to achieving global identification, the proposed method produces more efficient estimates compared to the Cox proportional hazards model that only includes the ratings from a single rater. The reason is that the proposed joint model incorporates all accessible ratings from multiple raters on the same subject.

Due to the absence of closed-form of solutions for the Cox model parameters, some joint modeling approaches use a one-step Newton-Raphson algorithm to update estimates in M-steps [15, 16, 30, 31]. For the proposed joint model, we employed the Newton-Raphson algorithm in the M-steps to update the Cox model parameters. We found that the Newton-

Raphson algorithm in the M-steps is equivalent to running a survey-weighted Cox model with some pseudo observations and weights created accordingly. However, the current R package `survey` may not converge well with extreme values in the sets of weights. Biases in some parameter estimates shown in the simulation studies may be because of convergence in using the R package `survey` or potential programming errors. We will work on comparing results with a one-step Newton-Raphson method to these with the survey-weighted Cox model to check whether the degree of biases can be reduced.

For testing the existence of a monotone association between the unknown true status and the risk of an event, we modified the application of the Union-Intersection principle for the Cox proportional hazards model as proposed by Sen [25]. Since the proposed joint model has the high dimensional nuisance parameters, we used the profile likelihood function instead of the partial likelihood function for the Cox proportional hazards model. We construct the test statistic of the Union-Intersection principle based on the corresponding profile likelihood score functions and their asymptotic variance-covariance matrix. Application of the Union-Intersection principle achieves more statistical power than the likelihood-based testing without considering the order restriction. In the future, we will perform additional simulation studies for evaluating statistical properties of the order restricted hypothesis testing for the proposed joint model regarding its statistical power.

In clinical studies, our proposed joint model can be widely applied to handle the issue of local identifiability to evaluate the accuracy of a diagnostic test without a gold standard. Also, the proposed method is useful to improve the precision of clinical decision making when diagnostic test results are predictive of a risk of disease recurrence and so patients can get more appropriate treatment consequently.

## BIBLIOGRAPHY

- [1] AITKIN, M., ANDERSON, D., AND HINDE, J. Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)* (1981), 419–461.
- [2] ALLMAN, E. S., MATIAS, C., AND RHODES, J. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* (2009), 3099–3132.
- [3] BINDER, D. A. Fitting cox’s proportional hazards models from survey data. *Biometrika* 79, 1 (1992), 139–147.
- [4] CHEN, H.-S., LAI, K., AND YING, Z. Goodness-of-fit tests and minimum power divergence estimators for survival data. *Statistica Sinica* 14, 1 (2004), 231–248.
- [5] COX, D. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (1972), 187–220.
- [6] CRESSIE, N., AND READ, T. R. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)* (1984), 440–464.
- [7] DAWID, A., AND SKENE, A. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics* 28 (1979), 20–28.
- [8] DEMPSTER, A., LAIRD, N., AND RUBIN, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39 (1977), 1–38.
- [9] ELSTON, C., AND ELLIS, I. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 19, 5 (1991), 403–410.
- [10] GARRETT, E. S., AND ZEGER, S. L. Latent class model diagnosis. *Biometrics* 56, 4 (2000), 1055–1067.
- [11] GOODMAN, L. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61, 2 (1974), 215–231.

- [12] GREENE, F., PAGE, D., FLEMING, I., BALCH, C., AND FRITZ, A. *AJCC cancer staging handbook: from the AJCC cancer staging manual*. Springer Verlag, 2002.
- [13] HENDERSON, R., DIGGLE, P., AND DOBSON, A. Joint modelling of longitudinal measurements and event time data. *Biostatistics* 1, 4 (2000), 465–480.
- [14] HUI, S., AND ZHOU, X. Evaluation of diagnostic tests without gold standards. *Statistical methods in medical research* 7, 4 (1998), 354–370.
- [15] LARSEN, K. Joint analysis of time-to-event and multiple binary indicators of latent classes. *Biometrics* 60, 1 (2004), 85–92.
- [16] LARSEN, K. The cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics* 61, 4 (2005), 1049–1055.
- [17] MCHUGH, R. Efficient estimation and local identification in latent class analysis. *Psychometrika* 21, 4 (1956), 331–347.
- [18] MOLENBERGHS, G., AND VERBEKE, G. Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician* 61, 1 (2007), 22–27.
- [19] MURPHY, S., AND VAN DER VAART, A. On profile likelihood. *Journal of the American Statistical Association* 95, 450 (2000), 449–465.
- [20] PAIK, S., SHAK, S., TANG, G., KIM, C., BAKER, J., CRONIN, M., BAEHNER, F., WALKER, M., WATSON, D., PARK, T., ET AL. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* 351, 27 (2004), 2817–2826.
- [21] QU, Y., TAN, M., AND KUTNER, M. Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics* 52 (1996), 797–810.
- [22] ROSENBERG, J., CHIA, Y., AND PLEVritis, S. The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the us seer database. *Breast cancer research and treatment* 89, 1 (2005), 47–54.
- [23] ROY, S. N. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* 24 (1953), 220–238.
- [24] SELF, S. G., AND LIANG, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* 82, 398 (1987), 605–610.
- [25] SEN, P. K. Subhypotheses testing against restricted alternatives for the cox regression model. *Journal of Statistical Planning and Inference* 10, 1 (1984), 31–42.
- [26] SEN, P. K. Union–intersection principle and constrained statistical inference. *Journal of statistical planning and inference* 137, 11 (2007), 3741–3752.

- [27] SILVAPULLE, M. J. Robust wald-type tests of one-sided hypotheses in the linear model. *Journal of the American Statistical Association* 87, 417 (1992), 156–161.
- [28] SILVAPULLE, M. J., AND SILVAPULLE, P. A score test against one-sided alternatives. *Journal of the American Statistical Association* 90, 429 (1995), 342–349.
- [29] SUBTIL, A., DE OLIVEIRA, M. R., AND GONÇALVES, L. Conditional dependence diagnostic in the latent class model: A simulation study. *Statistics & Probability Letters* 82, 7 (2012), 1407–1412.
- [30] THEISSIG, F., KUNZE, K., HAROSKE, G., AND MEYER, W. Histological grading of breast cancer: interobserver, reproducibility and prognostic significance. *Pathology-Research and Practice* 186, 6 (1990), 732–736.
- [31] TSIATIS, A., DEGRUTTOLA, V., AND WULFSOHN, M. Modeling the relationship of survival to longitudinal data measured with error. applications to survival and cd4 counts in patients with aids. *Journal of the American Statistical Association* 90, 429 (1995), 27–37.
- [32] VACEK, P. M. The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics* 17 (1985), 959–968.
- [33] VON DAVIER, M. Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a monte carlo study. *Methods of Psychological Research Online* 2, 2 (1997), 29–48.
- [34] WALTER, S., AND IRWIG, L. Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of clinical epidemiology* 41, 9 (1988), 923–937.
- [35] WULFSOHN, M., AND TSIATIS, A. A joint model for survival and longitudinal data measured with error. *Biometrics* 53 (1997), 330–339.
- [36] XU, J., AND ZEGER, S. Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50, 3 (2001), 375–387.