UNIVERSITÀ DEGLI STUDI DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
CORSO DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIX

STATISTICAL EVALUATION OF DIAGNOSTIC TESTS UNDER VERIFICATION BIAS

Coordinatore del Corso: Prof. Monica Chiogna

Supervisore: Prof. Monica Chiogna

Co-supervisore: Prof. Gianfranco Adimari

Dottorando: Khanh To Duc

Padova, 31-01, 2017

To my pretty family.

Acknowledgments

Foremost, I would like to express my sincere gratitude to my advisors, Prof. Monica Chiogna and Prof. Gianfranco Adimari, for the continuous support during my PhD studies and related research, for their patience, motivation, and immense knowledge. Their guidance helped me at all times of research and during the writing of this thesis. I could not have imagined having better advisors and mentors for my PhD studies.

Besides my advisors, I would like to thank the rest of my thesis evaluators: Prof. Aldo Solari and Prof. Christos Nakas, for their encouragement, insightful comments, and hard questions.

I am grateful to University of Padova for offering me the financial and living support through the Fondazione Cassa di Risparmio di Padova e Rovigo (CARIPARO) and for my stay in Padova during three years. Beside that, I would like to express sincere thanks to my friends from the 29-th PhD cycle, Claudia, Davide, Elisa, Lucia, Mirko and Paolo. You are my best Italian friends, and I feel very happy to have shared all positive and negative emotions during three years with such an amazing group of people.

Last but not the least, I would like to thank my sweet family and my little lover for their understanding, support and love throughout my graduate study.

Abstract

The use of diagnostic tests to discriminate between disease classes is becoming more and more popular in medicine, which leads to the urgent need for assessing accuracy of diagnostic tests before their implementation. To do that, a common tool is receiver operating characteristic (ROC) analysis. More precisely, the ROC curve and the area under the ROC curve (AUC) are commonly employed when two disease classes (typically, non-diseased and diseased) are considered, whereas the ROC surface and the volume under the ROC surface (VUS) are frequently used when the disease status has three categories (e.g., non-diseased, intermediate and diseased). In estimating such parameters, we assume that the true disease status of each patient can be determined by means of a gold standard test. In practice, unfortunately, the true disease status could be unavailable for all study subjects, due to the expensiveness or invasiveness of the gold standard test. Thus, often only a subset of patients undergoes disease verification. Statistical evaluations of diagnostic accuracy of a test based only on data from subjects with verified disease status are typically biased. This bias is known as verification bias. Various methods have been developed to adjust for verification bias in estimation of the ROC curve and its area for tests with binary or ordinal or continuous results. For the ROC surface and its volume, verification bias correction methods exist for tests with ordinal responses, but not for continuous tests. In this thesis, we propose several bias–corrected methods for estimating the ROC surface and the VUS of continuous diagnostic tests in presence of verification bias. In particular, these methods are constructed based on imputation and re–weighting techniques, and work well when the missingness mechanism of the true disease status is missing at random or missing not at random. The asymptotic behaviors of the estimators are also studied. To illustrate how to use the methods in real applications, two datasets dealing with epithelial ovarian cancer are considered. To support researchers in carrying out the ROC surface analysis in presence of verification bias, an R package and the corresponding Shiny web application have been created.

Sommario

L'uso corrente di test diagnostici per discriminare tra diverse malattie o classi di malattia pone l'accento sulla necessità di una valutazione attenta e fondata della loro accuratezza. Gli strumenti più comunemente impiegati a tal scopo sono basati sulla cosidetta receiver operating characteristic (ROC) analysis. Si utilizzano, in particolare, la curva ROC e l'area sotto la curva ROC (AUC) quando la diagnosi prevede due possibili esiti (tipicamente, non malato e malato), e la superficie ROC e il volume sotteso (VUS) quando la diagnosi si articola su tre classi (ad esempio, sano, stadio iniziale di malattia, stadio avanzato di malattia). Tali strumenti assumono che la vera diagnosi possa essere stabilita per ciascun paziente con certezza utilizzando un test gold standard. Nella pratica, purtroppo, la vera diagnosi potrebbe non essere acquisibile tramite un gold standard per tutti i soggetti coinvolti in uno studio, a causa per esempio del costo o della invasività del gold standard. Così, spesso, la verifica della diagnosi tramite gold standard viene condotta solo per un sottogruppo di pazienti. La valutazione statistica dell'accuratezza diagnostica di un test costruita solo utilizzando i dati dei soggetti con stato di malattia verificato è in genere distorta. Tale effetto è noto come distorsione di verifica. Esistono vari metodi per correggere tale distorsione nella stima della curva ROC e della area sottesa, sia per test diagnostici binari, che ordinali, che continui. Per quanto riguarda la superficie ROC ed il volume sotteso, esistono metodi di correzione della distorsione solo per test diagnostici ordinali. In questa tesi, si propongono diversi metodi per la correzione della distorsione di verfica per la stima della superficie ROC e del VUS per test diagnostici continui. Tali metodi sono costruiti su strategie di imputazione e riponderazione, e sono sviluppati per meccanismi di mancanza del vero stato di malattia sia casuali che non ignorabili. Viene fornito il comportamento asintotico degli stimatori. A titolo illustrativo, l'applicazione dei metodi è mostrata su due insiemi di dati relativi al cancro ovarico epiteliale. Per garantire applicabilità dei metodi, viene fornito un pacchetto R e l'applicazione web Shiny corrispondente.

Table of Contents

List of Figures

List of Tables

List of Symbols and Notation

TCF	True Class Fraction
TPR	True Positive Rate
FPR	False Positive Rate
$\mathcal{D} = (D_1, D_2, D_3)$	true disease status
V	the verification status
T	the continuous diagnostic test
A	the auxiliary covariates
(c_1, c_2), $c_1 < c_2$	the cut point
θ_k, $k = 1, 2, 3$	$\Pr(D_k = 1)$
β_{jk}	$\Pr(T \geq c_j \mid D_k = 1)$
ρ_k	$\Pr(D_k = 1 \mid T, A)$
π	$\Pr(V = 1 \mid T, A)$
$\tau = (\tau_\rho^\top, \tau_\pi^\top)^\top$	the vector of parameters of the models used to estimate $\rho = (\rho_1, \rho_2)^\top$, or π, or both
μ	the volume under ROC surface (VUS)
α	$(\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}, \beta_{23}, \tau^\top)^\top$
$\lambda = (\lambda_1, \lambda_2)^\top$	the nonignorable parameter
$\xi = (\lambda^\top, \tau_\pi^\top, \tau_\rho^\top)^\top$	the vector of nuisance parameters in nonignorable approach
$g_i^{\theta_s}(\alpha)$, $s = 1, 2$	the estimating function component corresponding to θ_s
$g_i^{\beta_{jk}}(\alpha)$	the estimating function component corresponding to β_{jk}
$g_i^{\tau}(\tau)$	the estimating function component corresponding to τ
$G^{\theta_s}(\alpha)$, $s = 1, 2$	$\sum_{i=1}^{n} g_i^{\theta_s}(\alpha)$, the estimating function of θ_s
$G^{\beta_{jk}}(\alpha)$	$\sum_{i=1}^{n} g_i^{\beta_{jk}}(\alpha)$, the estimating function of β_{jk}
$G^{\tau}(\tau)$	$\sum_{i=1}^{n} g_i^{\tau}(\tau)$, the estimating function of τ
$G_{i\ell r}(\mu, \tau)$	the estimating function component of the VUS
$\mathcal{S}$	the score function
$\mathcal{I}$	the Fisher information matrix
$\mathrm{I}(\cdot)$	the indicator function
i, ℓ, r	the index from 1 to n
n	the sample size

Chapter 1

Introduction

1.1 Overview

Nowadays, diagnostic tests are commonly used to detect medical conditions, and hence, they play an important role in medical care. The diagnostic tests are often inexpensive, non-invasive and applied on a large population. A good diagnostic test not only contains medical informations about patients, but also affects the health care provider's plan for managing the patient. However, different diagnostic tests differently distinguish between healthy individuals and diseased subjects. Therefore, assessing accuracy of diagnostic tests before they enter in the clinical practice is crucial and required.

In order to assess the accuracy of a diagnostic test, the comparison between the true disease status and the test results could be performed. To ascertain the true disease status, a perfect test is employed, which is called gold standard (GS) test. When the disease has two classes (typically, diseased and healthy), the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are commonly used for assessing accuracy of diagnostic tests (Pepe, 2003; Zhou et al., 2009). The ROC surface and the volume under the ROC surface (VUS), a generalization version of the ROC curve and of the AUC, have been useful in three–class diagnostic problems (Scurfield, 1996; Mossman, 1999; Dreiseitl et al., 2000). There are some methods for estimating ROC curves and AUCs, for example empirical, parametric, semi-parametric and nonparametric approaches, see Pepe (2003) and Zhou et al. (2009) as general references. These methods have been generalized to ROC surfaces and VUSs (Nakas and Yiannoutsos, 2004; Nakas, 2014; Kang and Tian, 2013; Li and Zhou, 2009; Li et al., 2012; Xiong et al., 2006).

In some situations, however, some subjects do not have the true disease status verified even when they have the test results. This basically comes from the fact that some gold standard tests are too expensive or invasive to be applied to all subjects. Thus, often only a subset of study subjects undergoes a gold standard evaluation and the decision to send a subject to verification typically depends on the test results and other subject's characteristics. Statistical evaluation of diagnostic accuracy based only on data from subjects with verified disease status usually gives biased results. This is known as verification bias (Begg and Greenes, 1983) or work–up bias (Ransohoff and Feinstein, 1978).

Correction for verification bias in two–class diagnostic testing is already discussed in the statistical literature. The available methods are developed on the basis of the type of diagnostic test result (e.g. binary, ordinal and continuous) and the missingness mechanism. Among the others, we cite the papers by Adimari and Chiogna (2015, 2016); Alonzo et al. (2003); Alonzo and Pepe

(2005); Fluss et al. (2009, 2012); He and McDermott (2012); He et al. (2009); Kosinski and Barnhart (2003); Liu and Zhou (2010); Rotnitzky et al. (2006); Zhou and Castelluccio (2003, 2004); Zhou and Rodenberg (1998). However, the issue of correcting the verification bias in the estimation of ROC surfaces and VUSs is very scarcely considered in the ROC analysis context. For example, Chi and Zhou (2008) deal with an ordinal diagnostic test, but there are no methods developed to correct for verification bias in assessing accuracy of a continuous diagnostic test when there are three disease status. This thesis aims to fill in this gap. We develop several methods for estimating ROC surfaces and VUSs of continuous diagnostic tests in presence of verification bias.

The outline of the thesis is as follows. In Chapter 2, we give a quick review on ROC curves and AUCs, and on ROC surfaces and VUSs. Beside that, the definition and impact of verification bias in estimation of ROC curves and their area underneath are also mentioned. Some existing bias–corrected methods for ROC curve analysis of a continuous diagnostic test are also reviewed. We develop some bias–corrected estimators for the ROC surface in Chapter 3. These methods are constructed based on imputation and re–weighting techniques and are suitable in cases where the missingness mechanism of disease status is missing at random. We establish large–sample properties of the proposed estimators based on estimating functions and multivariate delta method. Simulation studies are organized to evaluate the performance of the proposed estimators in both small and large sample size situations. To illustrate the applications of the methods, two distinct data sets, both dealing with epithelial ovarian cancer (EOC), are used. In Chapter 4, by using a numerical technique and repeating the ideas used in Chapter 3, we develop some methods for correcting for verification bias in estimation of the VUS. Consistency and asymptotic distribution of the estimators are also studied. Simulation studies are conducted to evaluate the performance of the bias–corrected VUS estimators. In analogy with Chapter 3, the two data sets on EOC study are used to illustrate the application of the bias–corrected VUS estimators. In Chapter 5, we consider the case in which the true disease status is missing not at random. Based on a likelihood-based approach, we propose the correction methods for VUS in presence of verification bias. We also demonstrate that the proposed estimators are consistent and asymptotically normally distributed. Simulation studies are designed to assess the accuracy of the estimators corresponding to the distinct values of VUS (from low to high value). In Chapter 6, we present an R package and the corresponding Shiny web application, which implement all bias–corrected methods for estimation of ROC surfaces and VUSs under missing at random assumption. Finally, Chapter 7 contains the main conclusions drawn from this project up to date and possible directions for future research.

1.2 Main contributions of the Thesis

Main contributions of the thesis can be summarized as follows:

1. Development of bias–corrected methods for estimation of the ROC surface and the VUS based on the existing methods for the ROC curve analysis, applicable in situations when the true disease status is missing at random.

2. Application to the prediction of patient's response to chemotherapy in advanced–stage epithelial ovarian cancer (EOC).

3. Definition of the approaches to correct for verification bias under missing not at random when disease status has three categories.

4. Creation of an R package and web interface to support researchers in carrying out the ROC surface analysis in presence of verification bias, i.e., when the true disease status is missing at random. This package is freely available for downloading from CRAN-The Comprehensive R Archive Network, and easy to use.

Chapter 2

Assessing the accuracy of a diagnostic test

In this chapter, definitions of some popular tools used to evaluate of a continuous diagnostic test will be given. The impact of verification bias when estimating the accuracy of a diagnostic test is discussed through a hypothetical example. A quick review of the existing bias–corrected methods for assessing the accuracy of a continuous diagnostic test is also given.

2.1 Measure of accuracy for diagnostic tests

Most medical studies consider a dataset, which often contains a medical diagnostic test and the true binary disease status (diseased and non-diseased) determined by the GS test. The diagnostic test can yield a binary, ordinal or continuous result, for which its accuracy could be determined by various ways and depends on the type of test result.

In case of binary tests (positive or negative), the accuracy of the test under study is usually evaluated by the true positive rate (TPR) and false positive rate (FPR), see Zhou et al. (2009) and Pepe (2003). The TPR is the probability that a diseased person is classified as diseased, whereas the FPR is the probability that a normal person is classified as diseased. In the study, a perfect diagnostic test has TPR = 1 and FPR = 0. On the other hand, a useless test has TPR = FPR, i.e., there are no connection between disease and the test outcome.

If the diagnostic tests are measured on ordinal or continuous scales, then they can be dichotomized in practice by using a cut point (also called threshold value) and thus the TPR and FPR could still be applied for measuring the accuracy of the tests. However, in such situations, each different choice of the cut point possibly yields different values of TPR and FPR, and, hence, the evaluation of the diagnostic test changes. Usually, the cut point is varied in the range of the test results, and the entire set of possible values of TPR and FPR is called the receiver operating characteristic (ROC) curve of the test. More specifically, the ROC curve is a plot, on the unit square, of the FPRs on x–axis versus the TPRs on y–axis, for each cut point. Usually, the ROC curve is monotone and lies in the upper triangle of the unit square, which consists of three vertices $(0,0), (0,1)$ and $(1,1)$. The shape of ROC curve allows to evaluate the ability of the test. For example, a ROC curve overlapping with a straight line joining points $(0,0)$ and $(1,1)$ represents a diagnostic test which is a random guess. A perfect test has a ROC curve that is along the left and upper borders of the positive unit quadrant. A commonly used summary measure that aggregates performance information of the test is the area under the ROC curve (AUC). According to the

property of the ROC curve, reasonable values of AUC range from 0.5, suggesting that the test is no better than chance alone, to 1.0, which indicates a perfect test. However, AUC can take a value less than 0.5, which indicates an accuracy worse than random guessing of the diagnosis.

In summary, the TPR, FPR, ROC curve and AUC are, currently, the best–developed statistical tools for measuring the performance of the diagnostic tests. The estimation of these quantities are presented in the context of Pepe (2003) and Zhou et al. (2009).

There are some medical studies, where the disease status involves three categories. For example, the clinical assessment of the presence of HIV–related cognitive dysfunction (AIDS Dementia Complex–ADC, Nakas and Yiannoutsos, 2004); the study of pancreatic cancer (Leichtle et al., 2013); the cohort study for the detection of Glycan biomarkers for liver cancer (Ressom et al., 2008); the study of Alzheimer's Disease (Xiong et al., 2006; Chi and Zhou, 2008). In such situations, the ROC surface is a popular tool for describing the ability of the medical tests having ordinal or continuous results. The ROC surface is defined by plotting three true class fractions (TCF's) by varying the cut point (c_1, c_2) in the unit cube, with $c_1 < c_2$. From a theoretical point of view, the ROC surface is a generalization of the ROC curve in three dimensions. The ROC surface will be the triangular plane with vertices $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$ if all of three TCF's are equal for every pair (c_1, c_2). In this case, we say that the diagnostic test is the random guess. In practice, one can imagine that the graph of ROC surface lies in the unit cube and above the plane of the triangle with three vertices $(0, 0, 1), (0, 1, 0)$, and $(1, 0, 0)$. A summary of the overall diagnostic accuracy of the test under consideration is the volume under the ROC surface (VUS) which can be seen as a generalization of the AUC. The VUS varies from $1/6$ to 1.0, ranging from bad to perfect diagnostic tests. More details related to the ROC surface analysis will be discussed in Section 2.4.

2.2 Verification bias

If all study subjects for which the new diagnostic test is available ultimately have their true disease status verified via the GS test, then ROC curves and also AUCs could be easily estimated. In practice, unfortunately, there are many drawbacks to the use of the GS test, which can be too expensive, or too invasive, or both for regular use. Usually, only a subset of subjects undergoes disease verification and the decision to send a patient to verify the disease status is often based on the test result and other patient characteristics. Then, only data from patients with verified disease status are used to estimate the ROC curve and AUC. This typically leads to a distorted evaluation of the ability of the diagnostic tests. This bias is known as verification bias (Begg and Greenes, 1983) or work-up bias (Ransohoff and Feinstein, 1978). In this paragraph, we present an example to see the impact of verification bias when estimating the accuracy of diagnostic test.

Let us consider a hypothetical medical study that concentrates on a new diagnostic test to detect an illness having probability of disease equal to 0.2. A sample having 1000 subjects randomly selected from population is considered. On the basis of the GS test, the number of diseased and non–diseased subjects are ascertained as 200 and 800, respectively. Of 200 diseased subjects, 160 have a positive test result and the remaining 40 subjects have a negative test result. In the 800 non–diseased people, 80 individuals give a positive and 720 individuals negative test result, respectively. The complete data are presented in Table 2.1.

For the full data, the TPR and FPR are easily obtained as $\frac{160}{200} = 0.8$ and $\frac{80}{800} = 0.1$, respectively. Now, suppose that the GS test in this study is inherently dangerous and also its cost is

Table 2.1: Full data for hypothetical example.

	Test result		
	Positive	Negative	Total
Diseased	160	40	200
Non–Diseased	80	720	800

expensive. Thus, only 75% of patients with a positive test and 10% with a negative test undergo the verification process. The verified sample is showed in Table 2.2.

Table 2.2: Data from subjects that selected to undergo the GS test.

	Test result		
	Positive	Negative	Total
Diseased	120	4	124
Non–Diseased	60	72	132

The estimation based on the verified data give rise to the estimate of TPR is $\frac{120}{124} = 0.968$ and of FPR is $\frac{60}{132} = 0.454$. This shows that the estimation on the verified subjects yields overestimation for the TPR (0.968 vs. 0.8) and FPR (0.454 vs. 0.1). Generally speaking, if the subjects having positive test results are more likely to be verified for disease than those having negative test results, then the bias in the estimation based on verified sample usually increases TPR and FPR (Pepe, 2003; Zhou et al., 2009).

2.3 Existing methods for correcting the ROC curve and AUC of a continuous test

Consider a study with n subjects, for whom the result of a continuous test T is available. The patient's true condition (or disease status), D, is defined by a GS test. D is a binary variable, that is 0 if the subject is non–disease and 1 in case of disease. Further, let V be a binary verification status of a patient, such that $V = 1$ if he/she is underwent the GS test, and $V = 0$ otherwise. In practice, some information, other than the test results, can be obtained for each patient. Let A be a covariate vector for a patient, that may be associated with both D and V.

In order to deal with verification bias, the existing methods usually make an assumption about the mechanism for the missingness of disease verification. In particular, there are three assumptions that are commonly used. The first is missing completely at random (MCAR), which occurs when the verification status V does not depend on any observed measurements (i.e., test result T, the covariates A) and unobserved measurements (i.e., the true disease status D). In other words, the mechanism is MCAR if missing values are randomly distributed across all observations. In medical studies, the selection for disease verification is not controlled by design (i.e., the MCAR assumption usually does not hold); instead, it is often decided by the physician on the basis of test results and other observed covariates. In such situations, we can assume that the verification status V and the response D are mutually independent given the test result T and covariates A,

i.e., $\Pr(V=1|T,A) = \Pr(V=1|D,T,A)$ or, equivalently, $\Pr(D=1|T,A) = \Pr(D=1|V,T,A)$. This assumption is known as the missing at random (MAR) assumption. However, in some cases, the MAR assumption is not realistic, because the decision to send a patient to verification may also depend on some hidden information related to disease status, which can not be determined by the observed measurements. Thus, the missing mechanism is neither MCAR nor MAR, and is known as missing not at random (MNAR).

Under a technical point of view, MCAR and MAR are usually qualified as *ignorable* missing data mechanisms, whereas MNAR is referred as *nonignorable* (NI) mechanism. In fact, for dealing with the missing values, we do not need to take into account the missingness mechanism under MCAR or MAR assumption. But, in order to deal with MNAR, a joint model for the data and the missingness mechanism is necessary to conduct valid inferences. The bias–corrected estimation methods for the ROC curve and AUC under MAR assumption are shortly reviewed in Section 2.3.2, whereas the methods for adjusting for NI verification bias is quickly presented in Section 2.3.3.

2.3.1 Full data estimation

When all subjects are verified by the GS test, we have a full (or complete) data set. For a given cut point c, TPR and FPR are

$$\begin{aligned} \mathrm{TPR}(c) &= \Pr(T \geq c|D=1) = \frac{\Pr(T \geq c, D=1)}{\Pr(D=1)} = \frac{\beta_1}{\theta}, \\ \mathrm{FPR}(c) &= \Pr(T \geq c|D=0) = \frac{\Pr(T \geq c, D=0)}{\Pr(D=0)} = \frac{\beta_0}{1-\theta}. \end{aligned} \tag{2.1}$$

Then, one can employ empirical estimators $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\theta}$ to obtain the nonparametric estimators of TPR and FPR

$$\widehat{\mathrm{TPR}}(c) = \frac{\hat{\beta}_1}{\hat{\theta}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c) D_i}{\sum_{i=1}^{n} D_i}, \qquad \widehat{\mathrm{FPR}}(c) = \frac{\hat{\beta}_0}{1-\hat{\theta}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)(1-D_i)}{\sum_{i=1}^{n} (1-D_i)}, \tag{2.2}$$

where $\mathrm{I}(\cdot)$ is the indicator function.

The AUC is expressed in the following formula (Bamber, 1975; Hanley and McNeil, 1982)

$$\mathrm{AUC} = \Pr(T_i < T_j | D_i = 0, D_j = 1) + \frac{1}{2}\Pr(T_i = T_j | D_i = 0, D_j = 1),$$

for $i \neq j$ from 1 to n. The empirical estimator of AUC is the Mann-Whitney U-statistic, i.e.,

$$\widehat{\mathrm{AUC}} = \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} \mathrm{I}_{ij}(1-D_i)D_j}{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} (1-D_i)D_j}, \tag{2.3}$$

where $\mathrm{I}_{ij} = \mathrm{I}(T_i < T_j) + \frac{1}{2}\mathrm{I}(T_i = T_j)$.

If not all patients have their disease status verified, the nonparametric estimators (2.2) and (2.3) can be applied after removal of all missing values. The resulting estimators are called Naïve estimators. However, they only perform well when the missing mechanism is MCAR. Under MAR or MNAR, if one tries to use the Naïve estimators, i.e., the expressions (2.2) and (2.3) based only on verified subjects, then one gets biased and inconsistent estimates.

2.3.2 Bias–corrected estimators under MAR assumption

In presence of verification bias, Alonzo and Pepe (2005) proposed four partially parametric estimators to assess the continuous diagnostic (or screening) tests under the MAR assumption. In particular, the authors constructed the estimates based on imputation and re-weighting techniques (Roberts et al., 1987; Reilly and Pepe, 1995; Horvitz and Thompson, 1952; Robins et al., 1995), i.e., full imputation (FI), mean score imputation (MSI), inverse probability weighting (IPW) and semiparameter efficient (SPE) estimators. The FI estimators of $\mathrm{TPR}(c)$ and $\mathrm{FPR}(c)$ are

$$\widehat{\mathrm{TPR}}_{\mathrm{FI}}(c) = \frac{\hat{\beta}_{1,\mathrm{FI}}}{\hat{\theta}_{\mathrm{FI}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\hat{\rho}_i}{\sum_{i=1}^{n} \hat{\rho}_i}, \quad \widehat{\mathrm{FPR}}_{\mathrm{FI}}(c) = \frac{\hat{\beta}_{0,\mathrm{FI}}}{1-\hat{\theta}_{\mathrm{FI}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)(1-\hat{\rho}_i)}{\sum_{i=1}^{n} (1-\hat{\rho}_i)}.$$

Here, the estimates $\hat{\rho}_i$ of $\rho_i = \Pr(D_i = 1|T_i, A_i)$ are obtained by using some suitable parametric model (e.g., logistic regression model) estimated on verified subjects. The MSI estimators only impute the disease status for subjects who did not undergo the GS test, resulting to be

$$\begin{aligned} \widehat{\mathrm{TPR}}_{\mathrm{MSI}}(c) &= \frac{\hat{\beta}_{1,\mathrm{MSI}}}{\hat{\theta}_{\mathrm{MSI}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i D_i + (1-V_i)\hat{\rho}_i\right\}}{\sum_{i=1}^{n} \left\{V_i D_i + (1-V_i)\hat{\rho}_i\right\}}, \\ \widehat{\mathrm{FPR}}_{\mathrm{MSI}}(c) &= \frac{\hat{\beta}_{0,\mathrm{MSI}}}{1-\hat{\theta}_{\mathrm{MSI}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_i)\right\}}{\sum_{i=1}^{n} \left\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_i)\right\}}. \end{aligned} \tag{2.4}$$

The IPW method weights each verified subject by the inverse of the conditional verification probability $\pi_i = \Pr(V_i = 1|T_i, A_i)$ (i.e. the probability that the subject is selected for verification). Therefore, the estimators are

$$\begin{aligned} \widehat{\mathrm{TPR}}_{\mathrm{IPW}}(c) &= \frac{\hat{\beta}_{1,\mathrm{IPW}}}{\hat{\theta}_{\mathrm{IPW}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c) V_i D_i \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} V_i D_i \hat{\pi}_i^{-1}}, \\ \widehat{\mathrm{FPR}}_{\mathrm{IPW}}(c) &= \frac{\hat{\beta}_{0,\mathrm{IPW}}}{1-\hat{\theta}_{\mathrm{IPW}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c) V_i (1-D_i) \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} V_i (1-D_i) \hat{\pi}_i^{-1}}. \end{aligned} \tag{2.5}$$

The estimates $\hat{\pi}_i$ need to be obtained by using parametric regression models such as logistic or probit models. Finally, the SPE estimators are defined as follow

$$\begin{aligned} \widehat{\mathrm{TPR}}_{\mathrm{SPE}}(c) &= \frac{\hat{\beta}_{1,\mathrm{SPE}}}{\hat{\theta}_{\mathrm{SPE}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i D_i \hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)\hat{\rho}_i \hat{\pi}_i^{-1}\right\}}{\sum_{i=1}^{n} \left\{V_i D_i \hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)\hat{\rho}_i \hat{\pi}_i^{-1}\right\}}, \\ \widehat{\mathrm{FPR}}_{\mathrm{SPE}}(c) &= \frac{\hat{\beta}_{0,\mathrm{SPE}}}{1-\hat{\theta}_{\mathrm{SPE}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i (1-D_i) \hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)(1-\hat{\rho}_i) \hat{\pi}_i^{-1}\right\}}{\sum_{i=1}^{n} \left\{V_i (1-D_i) \hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)(1-\hat{\rho}_i) \hat{\pi}_i^{-1}\right\}}. \end{aligned} \tag{2.6}$$

Alonzo and Pepe (2005) showed that SPE estimators are doubly robust because they are consistent if either the π_i's or the ρ_i's are consistently estimated. However, it is worth noting that SPE esti-

mates may not be range-respecting, i.e., they could fall outside the interval $(0, 1)$. This happens because the quantities $\left\{V_i(1-D_i)\hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)(1-\hat{\rho}_i)\hat{\pi}_i^{-1}\right\}$ or $\left\{V_i D_i \hat{\pi}_i^{-1} - (V_i - \hat{\pi}_i)\hat{\rho}_i\hat{\pi}_i^{-1}\right\}$ can be negative.

The four bias–corrected methods mentioned above perform well if and only if the parametric models are correctly specified (the disease and/or verification model). A wrong specification of such parametric models can negatively affect the behavior of the estimators, that are no longer consistent. To reduce the impact of misspecified models, He and McDermott (2012) proposed to use propensity score stratification; however, this method only applies to a binary diagnostic test. For a continuous test, a fully nonparametric framework is suggested by Adimari and Chiogna (2015). Specifically, the authors employed K nearest-neighbor (KNN) imputation (Ning and Cheng, 2012; Cheng, 1994) to obtain the bias-corrected estimators for the ROC curves. In fact, the KNN estimators are

$$\begin{aligned}
\widehat{\mathrm{TPR}}_{\mathrm{KNN}}(c) &= \frac{\hat{\beta}_{1,\mathrm{KNN}}}{\hat{\theta}_{\mathrm{KNN}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i D_i + (1-V_i)\hat{\rho}_{i,K}\right\}}{\sum_{i=1}^{n}\left\{V_i D_i + (1-V_i)\hat{\rho}_{i,K}\right\}}, \\
\widehat{\mathrm{FPR}}_{\mathrm{KNN}}(c) &= \frac{\hat{\beta}_{0,\mathrm{KNN}}}{1-\hat{\theta}_{\mathrm{KNN}}} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c)\left\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{i,K})\right\}}{\sum_{i=1}^{n}\left\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{i,K})\right\}}.
\end{aligned} \tag{2.7}$$

Here, $\hat{\rho}_{i,K}$ is the estimate of ρ_i obtained by using KNN imputation.

For each of the above methods, an estimated bias-corrected ROC curve can be obtained by plotting $\widehat{\mathrm{TPR}}_*(c)$ versus $\widehat{\mathrm{FPR}}_*(c)$ for all cut points c, where the star $*$ indicates FI, MSI, IPW, SPE and KNN. Note that, the SPE estimate of the ROC curve could be non monotone, because of its behavior. Therefore, the authors suggest to use isotonic regression (Robertson et al., 1988), to force the SPE ROC curve to be monotone.

Based on the bias–corrected ROC curves described above, one can employ the trapezoidal rule (Bamber, 1975) to get empirical estimators of the AUC. This solution is supported by Alonzo and Pepe (2005), who suggest to use the bootstrap resampling method to obtain the asymptotic variance. On the other hand, based on U-statistics and IPW, He and colleagues derived the closed-form expressions for directly estimating the AUC (He et al., 2009). In fact, they assume that the verification process $\pi_i = \Pr(V_i = 1 | T_i, A_i)$ were known and define

$$\widehat{\mathrm{AUC}}_{\mathrm{DIR}} = \frac{\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} \mathrm{I}(T_i < T_j)\mathrm{I}(D_i < D_j) V_i V_j \pi_i^{-1}\pi_j^{-1}}{\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} \mathrm{I}(D_i < D_j) V_i V_j \pi_i^{-1}\pi_j^{-1}}. \tag{2.8}$$

However, in practice, the verification probabilities are unknown, hence, in such situations, the authors suggest to use consistent estimates $\hat{\pi}_i$. Therefore, the direct estimation (2.8) can be referred to as parametric framework, in the sense that it requires a parametric model (e.g., logistic regression) to estimate π. Hence, it also could suffer from the effect of misspecification. The nonparametric estimators for the AUC in the setting of verification bias is proposed by Adimari and Chiogna (2016). In fact, the authors used KNN imputation, again, like in the method for the

ROC curves. The KNN estimator for the AUC is

$$\widehat{\text{AUC}}_{\text{KNN}} = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n} \text{I}(T_i < T_j)(1-\hat{D}_i)\hat{D}_j}{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}(1-\hat{D}_i)\hat{D}_j}, \tag{2.9}$$

where $\hat{D}_i = V_i D_i + (1-V_i)\hat{\rho}_{i,K}$.

2.3.3 NI verification bias

As we already mentioned in the beginning of this section, the MNAR assumption may be suitable in some situations. Many publications deal with NI verification bias, for instance, Baker (1995); Zhou and Castelluccio (2003, 2004); Zhou and Rodenberg (1998); Kosinski and Barnhart (2003) focus on the binary and ordinal diagnostic test. For the continuous test, Rotnitzky et al. (2006); Fluss et al. (2009); Liu and Zhou (2010) developed various methods for adjusting for the verification bias.

In principle, we need to define a joint model of the data and missingness mechanism. According this idea, Rotnitzky et al. (2006) used the untestable selection model

$$\log\left\{\frac{\Pr(V=0|T,A,D)}{\Pr(V=1|T,A,D)}\right\} = h(T,A) + q(T,A)D,$$

where $q(T,A)$ is an arbitrary specified function and $h(T,A)$ is an arbitrary unknown function. Under this model, the missing mechanism is MAR if $q(T,A)=0$ for all T and A; otherwise, it is NI. Using this model, doubly robust estimators of the AUC and the ROC curves are derived by Rotnitzky et al. (2006) and Fluss et al. (2009), respectively. However, a drawback of this model is the fixedness of $q(T,V)$. To cope with this problem, the authors recommend using a sensitivity analysis by repeating the estimation of AUC (and also TPR and FPR) under a variety of reasonable choices of $q(T,A)$. This may work well, but in the general cases, we might face troubles related to the range of $q(T,A)$, which can be large, and to the computation, which can be heavy.

On the other hand, Liu and Zhou (2010) suggested to use the verification model

$$\Pr(V=1|T,A,D) = \frac{\exp(h(T,A;\beta)+\lambda D)}{1+\exp(h(T,A;\beta)+\lambda D)}.$$

Here, $h(T,A;\beta)$ is an linear predictor; and λ is an unknown parameter called the nonignorable parameter. To estimate the parameters, Liu and Zhou used a likelihood-based approach, together with a disease model for the whole sample

$$\Pr(D=1|T,A) = \frac{\exp(m(T,A;\gamma))}{1+\exp(m(T,A;\gamma))},$$

where $m(T,A;\gamma)$ is an arbitrary linear predictor. After that, the authors employed the imputation and re–weighting techniques to correct the ROC curve and the area underneath, i.e., FI, MSI, IPW and pseudo doubly robust (PDR) estimators. Unfortunately, this strategy requires both the disease model and the verification model exactly specified and a large sample size (may be several thousands). To construct verification and disease models, the authors recommend selecting covariates based on scientific knowledge in the literature. A small sample size may occur in some medical studies, because of the cost or other difficulties. In such situations, Liu and Zhou suggest to use the MAR assumption, instead. But, we have to be careful with this option, because a distorted result may be generated.

2.4 The ROC surface analysis

In a three–class diagnostic problem, the ROC surface analysis is frequently used for the evaluation of diagnostic markers. The theoretical construction of the ROC surface and VUS was introduced for the first time by Scurfield (1996). However, this article did not provide any application on real data. After that, in an independent study, Mossman (1999) proposed a similar construction and gave two applications on window-rating data and psychiatric data. Based on the Mossman's construction, Dreiseitl et al. (2000) proposed the three-way ROC surface and derived a nonparametric estimate of variance for the VUS. The ROC surface construction for a continuous diagnostic test is provided by Nakas and Yiannoutsos (2004), as a direct generalization of the ROC curve to three–class diagnostic problems.

We model the disease status by a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$, where D_k is a binary variable that takes 1 if the subject belongs to class k, $k = 1, 2, 3$. Without loss of generality, we assume that the subjects from class 3 tend to have higher test results than subjects in class 2 and the latter tend to have higher test results than subjects in class 1, i.e., $T|D_1 < T|D_2 < T|D_3$. This implies that the disease classes are ordered with respect to the test result, a condition often referred to as monotone ordering. For given a pair of cut points (c_1, c_2), with $c_1 < c_2$, subjects are classified into class 1 if $T < c_1$; class 2 if $c_1 \leq T < c_2$; and class 3 otherwise. The true class fractions of the test T at (c_1, c_2) are defined as

$$\begin{aligned}
\mathrm{TCF}_1(c_1) &= \Pr(T < c_1|\text{class } 1) = 1 - \Pr(T \geq c_1|D_1 = 1), \\
\mathrm{TCF}_2(c_1, c_2) &= \Pr(c_1 < T < c_2|\text{class } 2) \\
&= \Pr(T \geq c_1|D_2 = 1) - \Pr(T \geq c_2|D_2 = 1), \\
\mathrm{TCF}_3(c_2) &= \Pr(T > c_2|\text{class } 3) = \Pr(T \geq c_2|D_3 = 1).
\end{aligned}$$

The plot of $(\mathrm{TCF}_1, \mathrm{TCF}_2, \mathrm{TCF}_3)$ by varying the pair (c_1, c_2) produces the ROC surface of T in the unit cube. Figure 2.1 shows the ROC surface of a given diagnostic test and the triangular plane of the random guess. According to this figure, the projection of the ROC surface to the plane defined by TCF_2 versus TCF_1 yields the ROC curve between classes 1 and 2. Similarly, on projecting the ROC surface to the plane defined by the axes TCF_2 and TCF_3, the ROC curve between classes 2 and 3 is produced (see also Nakas (2014)).

The general formula of the VUS of the diagnostic test T, say μ, is defined as (Nakas and Yiannoutsos, 2004)

$$\begin{aligned}
\mu &= \Pr\left(T_i < T_\ell < T_r|D_{1i} = 1, D_{2\ell} = 1, D_{3r} = 1\right) \\
&\quad + \frac{1}{2}\Pr\left(T_i < T_\ell = T_r|D_{1i} = 1, D_{2\ell} = 1, D_{3r} = 1\right) \\
&\quad + \frac{1}{2}\Pr\left(T_i = T_\ell < T_r|D_{1i} = 1, D_{2\ell} = 1, D_{3r} = 1\right) \\
&\quad + \frac{1}{6}\Pr\left(T_i = T_\ell = T_r|D_{1i} = 1, D_{2\ell} = 1, D_{3r} = 1\right)
\end{aligned}$$

or, equivalently,

$$\mu = \frac{\mathbb{E}\left(D_{1i}D_{2\ell}D_{3r}\mathrm{I}_{i\ell r}\right)}{\mathbb{E}\left(D_{1i}D_{2\ell}D_{3r}\right)}, \qquad i \neq \ell \neq r, \tag{2.10}$$

where $\mathrm{I}_{i\ell r} = \mathrm{I}(T_i < T_\ell < T_r) + \frac{1}{2}\mathrm{I}(T_i < T_\ell = T_r) + \frac{1}{2}\mathrm{I}(T_i = T_\ell < T_r) + \frac{1}{6}\mathrm{I}(T_i = T_\ell = T_r)$ and $\mathrm{I}(\cdot)$ is the indicator function. Under correct ordering, a suitable value of μ lies between 1/6 to 1. More specifically, $\mu = 1/6$ indicates the classification rule is uninformative, while the value 1

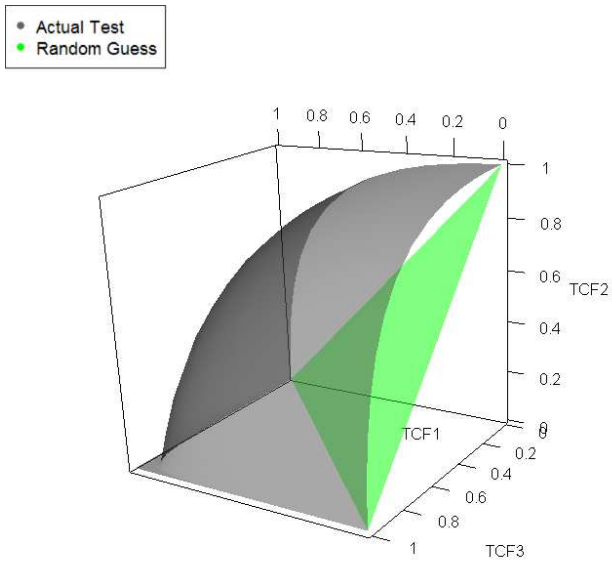


Figure 2.1: An example of the ROC surface

indicates the perfect diagnostic test. Moreover, it is worth noting that the VUS is invariant under monotonically increasing data transformations (Nakas and Yiannoutsos, 2004).

When all subjects are verified, the nonparametric estimators of the true class fractions and VUS are given by

$$\begin{aligned}
\widehat{\mathrm{TCF}}_1(c_1) &= 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) D_{1i}}{\sum_{i=1}^{n} D_{1i}}, \\
\widehat{\mathrm{TCF}}_2(c_1, c_2) &= \frac{\sum_{i=1}^{n} \{\mathrm{I}(T_i \geq c_1) - \mathrm{I}(T_i \geq c_2)\} D_{2i}}{\sum_{i=1}^{n} D_{2i}}, \\
\widehat{\mathrm{TCF}}_3(c_2) &= \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) D_{3i}}{\sum_{i=1}^{n} D_{3i}},
\end{aligned} \tag{2.11}$$

and

$$\hat{\mu}_{\mathrm{NP}} = \frac{\sum_{i=1}^{n} \sum_{\ell=1, \ell \neq i}^{n} \sum_{\substack{r=1 \\ r \neq \ell, r \neq i}}^{n} \mathrm{I}_{i\ell r} D_{1i} D_{2\ell} D_{3r}}{\sum_{i=1}^{n} \sum_{\ell=1, \ell \neq i}^{n} \sum_{\substack{r=1 \\ r \neq \ell, r \neq i}}^{n} D_{1i} D_{2\ell} D_{3r}}. \tag{2.12}$$

In recent years, many methods have been developed for estimating the ROC surface and VUS in absence of verification bias. Nakas and Yiannoutsos (2004) and Nakas (2014) gave some interesting results about the ROC surface analysis. In their context, the ROC surface is formulated by a functional form and a nonparametric approach for the VUS estimation is also provided. Conversely, parametric estimation of VUS is given in Xiong et al. (2006), where the assumption of normality distribution was used, whereas Li and Zhou (2009) tackled the nonparametric and semi-parametric estimation of the ROC surface. Li et al. (2012) proposed a regression approach to ROC surface, and

in Kang and Tian (2013) a kernel smoothing based approach for estimation of VUS is employed.

Like the case of two–class diagnostic problem, verification bias occurs when we try to use the estimators (2.11) for TCFs and (2.12) for VUS when the missing mechanism is MAR or MNAR. The issue of correcting for the verification bias in ROC surface analysis is very scarcely considered in the statistical literature. Until now, only Chi and Zhou (2008) discussed about the issue. The authors proposed maximum likelihood estimates for ROC surface and VUS under the MAR assumption. However, these results only concern ordinal diagnostic tests.

Chapter 3

Bias–corrected methods for estimating the ROC surface

ROC surfaces are the 3D plots of true class fractions $(\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1, c_2), \mathrm{TCF}_3(c_2))$ by varying the cut points (c_1, c_2). Thus, correcting for verification bias a ROC surface boils down to bias-corrected estimation for TCFs.

Recall that the disease status is presented as the trinomial vector $\mathcal{D} = (D_1, D_2, D_3)$ such that $D_k = 1$ if the subject is belong to class k with $k = 1, 2, 3$. Thus, D_k is a Bernoulli random variable having mean $\theta_k = \Pr(D_k = 1)$, $k = 1, 2, 3$, with $\theta_1 + \theta_2 + \theta_3 = 1$. Let $\beta_{jk} = \Pr(T \geq c_j, D_k = 1)$ with $j = 1, 2$ and $k = 1, 2, 3$. In this notation,

$$\begin{aligned} \mathrm{TCF}_1(c_1) &= 1 - \frac{\Pr(T \geq c_1, D_1 = 1)}{\Pr(D_1 = 1)} = 1 - \frac{\beta_{11}}{\theta_1}, \\ \mathrm{TCF}_2(c_1, c_2) &= \frac{\Pr(T \geq c_1, D_2 = 1) - \Pr(T \geq c_2, D_2 = 1)}{\Pr(D_2 = 1)} = \frac{\beta_{12} - \beta_{22}}{\theta_2}, \\ \mathrm{TCF}_3(c_2) &= \frac{\Pr(T \geq c_2, D_3 = 1)}{\Pr(D_3 = 1)} = \frac{\beta_{23}}{\theta_3}. \end{aligned} \tag{3.1}$$

Thus, our goal is find bias–corrected estimators of the quantities $\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}$ and β_{23}.

In this chapter, we propose five bias–corrected approaches, which work under MAR assumption and can be seen as an extension of estimators reviewed in Subsection 2.3. In expressions (2.1) and (3.1), we note that parameters θ and θ_k, so as β_1 and β_{jk}, play, in essence, a similar role. Therefore, estimators of θ_k and β_{jk} can be obtained by mimicking what was done in the two-class problem.

3.1 Parametric methods

3.1.1 Full imputation

For each $j = 1, 2$ and $k = 1, 2, 3$, the FI estimators of θ_k and β_{jk} are obtained as

$$\hat{\theta}_{k,\mathrm{FI}} = \widehat{\Pr}(D_k = 1) = \frac{1}{n} \sum_{i=1}^{n} \hat{\rho}_{ki}, \tag{3.2}$$

$$\hat{\beta}_{jk,\mathrm{FI}} = \widehat{\Pr}(T \geq c_j, D_k = 1) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) \hat{\rho}_{ki}, \tag{3.3}$$

where $\hat{\rho}_{ki}$ is an estimate of $\rho_{ki} = \Pr(D_{ki} = 1 | T_i, A_i)$ given by some suitable model, such as the multinomial logistic or probit regression model, applied to the verified sample units. Note that in

what follows we will perform estimation within the framework of maximum likelihood. Therefore, the full imputation estimators $\widehat{\mathrm{TCF}}_{1,\mathrm{FI}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{FI}}(c_1,c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{FI}}(c_2)$ are

$$\begin{aligned}
\widehat{\mathrm{TCF}}_{1,\mathrm{FI}}(c_1) &= 1-\frac{\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_1)\hat{\rho}_{1i}}{\sum_{i=1}^{n}\hat{\rho}_{1i}},\\
\widehat{\mathrm{TCF}}_{2,\mathrm{FI}}(c_1,c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(c_1\leq T_i<c_2)\hat{\rho}_{2i}}{\sum_{i=1}^{n}\hat{\rho}_{2i}},\\
\widehat{\mathrm{TCF}}_{3,\mathrm{FI}}(c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_2)\hat{\rho}_{3i}}{\sum_{i=1}^{n}\hat{\rho}_{3i}}.
\end{aligned}$$

It is worth noting that estimators $\hat{\theta}_{k,\mathrm{FI}}$ and $\hat{\beta}_{jk,\mathrm{FI}}$ in (3.2) and (3.3) are the solutions of the estimating equations

$$\sum_{i=1}^{n}(\hat{\rho}_{ki}-\theta_k) = 0, \tag{3.4}$$

$$\sum_{i=1}^{n}\{\mathrm{I}(T_i\geq c_j)\hat{\rho}_{ki}-\beta_{jk}\} = 0. \tag{3.5}$$

3.1.2 Mean score imputation

By inspection of (2.4), we get the MSI estimators of θ_k, $k=1,2,3$, as follows

$$\hat{\theta}_{k,\mathrm{MSI}}=\widehat{\Pr}(D_k=1)=\frac{1}{n}\sum_{i=1}^{n}\left[V_iD_{ki}+(1-V_i)\hat{\rho}_{ki}\right].$$

The estimators of β_{jk} are given by

$$\hat{\beta}_{jk,\mathrm{MSI}}=\widehat{\Pr}(T\geq c_j,D_k=1)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)\left[V_iD_{ki}+(1-V_i)\hat{\rho}_{ki}\right].$$

Then, the MSI estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1,c_2)$ and $\mathrm{TCF}_3(c_2)$ are :

$$\begin{aligned}
\widehat{\mathrm{TCF}}_{1,\mathrm{MSI}}(c_1) &= 1-\frac{\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_1)\left[V_iD_{1i}+(1-V_i)\hat{\rho}_{1i}\right]}{\sum_{i=1}^{n}\left[V_iD_{1i}+(1-V_i)\hat{\rho}_{1i}\right]},\\
\widehat{\mathrm{TCF}}_{2,\mathrm{MSI}}(c_1,c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(c_1\leq T_i<c_2)\left[V_iD_{2i}+(1-V_i)\hat{\rho}_{2i}\right]}{\sum_{i=1}^{n}\left[V_iD_{2i}+(1-V_i)\hat{\rho}_{2i}\right]},\\
\widehat{\mathrm{TCF}}_{3,\mathrm{MSI}}(c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_2)\left[V_iD_{3i}+(1-V_i)\hat{\rho}_{3i}\right]}{\sum_{i=1}^{n}\left[V_iD_{3i}+(1-V_i)\hat{\rho}_{3i}\right]}.
\end{aligned}$$

Again, we can obtain $\hat{\theta}_k$ and $\hat{\beta}_{jk}$ as solution of the estimating equations

$$\sum_{i=1}^{n} \{V_i(D_{ki} - \theta_k) + (1 - V_i)(\hat{\rho}_{ki} - \theta_k)\} = 0, \tag{3.6}$$

$$\sum_{i=1}^{n} \{V_i(\mathrm{I}(T_i \geq c_j)D_{ki} - \beta_{jk}) + (1 - V_i)(\mathrm{I}(T_i \geq c_j)\hat{\rho}_{ki} - \beta_{jk})\} = 0. \tag{3.7}$$

3.1.3 Inverse probability weighting

From the IPW estimators of β_1 and θ in (2.5), we derive, by analogy,

$$\hat{\theta}_{k,\mathrm{IPW}} = \widehat{\Pr}(D_k = 1) = \frac{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{ki}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1}},$$

$$\hat{\beta}_{jk,\mathrm{IPW}} = \widehat{\Pr}(T \geq c_j, D_k = 1) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) V_i \hat{\pi}_i^{-1} D_{ki}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1}}.$$

The estimates $\hat{\pi}_i$ are obtained in the same way as in the two-class case, i.e, by using parametric regression models such as logistic or probit models. Again, in what follows we will employ maximum likelihood estimation. Then, the IPW estimators $\widehat{\mathrm{TCF}}_{1,\mathrm{IPW}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{IPW}}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{IPW}}(c_2)$ are

$$\widehat{\mathrm{TCF}}_{1,\mathrm{IPW}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) V_i \hat{\pi}_i^{-1} D_{1i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{1i}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{IPW}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) V_i \hat{\pi}_i^{-1} D_{2i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{2i}},$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{IPW}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) V_i \hat{\pi}_i^{-1} D_{3i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{3i}},$$

and the estimating equations corresponding to $\hat{\theta}_{k,\mathrm{IPW}}$ and $\hat{\beta}_{jk,\mathrm{IPW}}$ are

$$\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} (D_{ki} - \theta_k) = 0, \tag{3.8}$$

$$\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} (\mathrm{I}(T_i \geq c_j) D_{ki} - \beta_{jk}) = 0. \tag{3.9}$$

Note that the IPW estimators only use verified subjects.

3.1.4 Semiparametric efficient

Similarly to three previous cases, the SPE estimators of β_{jk} and θ_k are derived in analogy to $\hat{\beta}_{1,\mathrm{SPE}}$ and $\hat{\theta}_{\mathrm{SPE}}$ in (2.6), i.e,

$$\hat{\theta}_{k,\mathrm{SPE}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{V_i D_{ki}}{\hat{\pi}_i} - \frac{\hat{\rho}_{ki}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}, \tag{3.10}$$

$$\hat{\beta}_{jk,\mathrm{SPE}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i \geq c_j)\left\{\frac{V_i D_{ki}}{\hat{\pi}_i} - \frac{\hat{\rho}_{ki}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}. \tag{3.11}$$

Therefore, we obtain

$$\widehat{\mathrm{TCF}}_{1,\mathrm{SPE}}(c_1) = 1 - \frac{\sum_{i=1}^{n}\mathrm{I}(T_i \geq c_1)\left\{\frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}}{\sum_{i=1}^{n}\left\{\frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{SPE}}(c_1,c_2) = \frac{\sum_{i=1}^{n}\mathrm{I}(c_1 \leq T_i < c_2)\left\{\frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}}{\sum_{i=1}^{n}\left\{\frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}},$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{SPE}}(c_2) = \frac{\sum_{i=1}^{n}\mathrm{I}(T_i \geq c_2)\left\{\frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}}{\sum_{i=1}^{n}\left\{\frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i-\hat{\pi}_i)}{\hat{\pi}_i}\right\}}.$$

The estimators $\hat{\theta}_{k,\mathrm{SPE}}$ and $\hat{\beta}_{jk,\mathrm{SPE}}$ solve the estimating equations

$$\sum_{i=1}^{n}\left\{\frac{V_i}{\hat{\pi}_i}\left[\mathrm{I}(T_i \geq c_j)D_{ki} - \beta_{jk}\right] - \frac{V_i-\hat{\pi}_i}{\hat{\pi}_i}\left[\mathrm{I}(T_i \geq c_j)\hat{\rho}_{ki} - \beta_{jk}\right]\right\} = 0, \tag{3.12}$$

$$\sum_{i=1}^{n}\left\{\frac{V_i}{\hat{\pi}_i}\left(D_{ki} - \theta_k\right) - \frac{V_i-\hat{\pi}_i}{\hat{\pi}_i}\left(\hat{\rho}_{ki} - \theta_k\right)\right\} = 0. \tag{3.13}$$

The SPE estimators are also known to be doubly robust estimators, in the sense that they are consistent if either the ρ_{ki}'s or the π_i's are estimated consistently. However, SPE estimates could fall outside the interval $(0,1)$. This happens because the quantities $V_i D_{ki}\hat{\pi}_i^{-1} - \hat{\rho}_{ki}(V_i-\hat{\pi}_i)\hat{\pi}_i^{-1}$ can be negative.

3.1.5 Asymptotic distribution theory

The parameters of interest $\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1,c_2)$ and $\mathrm{TCF}_3(c_2)$ are functions of θ_1, θ_2, β_{11}, β_{12}, β_{22}, β_{23} and $\tau = (\tau_\rho^\top, \tau_\pi^\top)^\top$, where τ is the vector of parameters of the models used to estimate $\rho = (\rho_1^\top, \rho_2^\top)^\top$, or π, or both. Let us denote $\alpha = (\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}, \beta_{23}, \tau^\top)^\top$. The estimators (FI, MSI, IPW, SPE) of α are obtained by solving suitable estimating equations. Hence, we use results in Alonzo et al. (2003) and Alonzo and Pepe (2005) to give consistency and asymptotic normality of the proposed bias–corrected estimators.

According to equations (3.4), (3.5), (3.6), (3.7), (3.8), (3.9), (3.12) and (3.13), let $G_*^{\theta_s}(\alpha) = \sum_{i=1}^{n} g_{i,*}^{\theta_s}(\alpha)$ and $G_*^{\beta_{jk}}(\alpha) = \sum_{i=1}^{n} g_{i,*}^{\beta_{jk}}(\alpha)$ be the estimating functions for θ_s and β_{jk}, with $k = 1,2,3$, s and $j = 1,2$, for one of the four previously introduced approaches (the star indicates FI, MSI, IPW, SPE). We assume that $\hat{\tau}$ is the solution to a classic set of estimating equations of the form $G^\tau(\alpha) = \sum_{i=1}^{n} g_i^\tau(\alpha) = 0$. Specifically, we will employ classic score equations derived from: a multinomial

logistic regression model for estimation of the disease process; and from a logistic regression model for estimation of the verification process. The estimate $\hat{\alpha}_*$ of α is then obtained by solving $G_*(\alpha) = \sum_{i=1}^{n} g_{i,*}(\alpha) = 0$, where $g_{i,*}(\alpha) = \left(g_{i,*}^{\theta_1}(\alpha), g_{i,*}^{\theta_2}(\alpha), g_{i,*}^{\beta_{11}}(\alpha), g_{i,*}^{\beta_{12}}(\alpha), g_{i,*}^{\beta_{22}}(\alpha), g_{i,*}^{\beta_{23}}(\alpha), g_i^{\tau}(\alpha)^\top\right)^\top$.

Let $\alpha_0 = \left(\theta_{10}, \theta_{10}, \beta_{110}, \beta_{120}, \beta_{220}, \beta_{230}, \tau_0^\top\right)^\top$ be the true value of α. We assume that

(A1) $\mathcal{D}$ is missing at random (MAR);

(A2) the data $(\mathcal{D}_i, T_i, A_i^\top, V_i)^\top$ are i.i.d;

(A3) $(T, A^\top)^\top$ is a bounded random vector;

(A4) $\mathbb{E}\left[\frac{\partial}{\partial \alpha^\top} g_{i,*}(\alpha_0)\right]$ is negative definite;

(A5) ρ_{ki} and π_i are bounded away from 0.

We consider also the following standard regularity conditions.

(C1) $g_{i,*}(\alpha_0)$ are i.i.d and $\mathbb{E}\{g_{i,*}(\alpha_0)\} = 0$.

(C2) Elements of $G_*(\alpha)$, $\frac{\partial}{\partial \alpha^\top} G_*(\alpha)$, and $\frac{\partial^2}{\partial \alpha \partial \alpha^\top} G_*(\alpha)$ exist in a bounded δ-neighborhood of α_0, $N_\delta(\alpha_0)$.

(C3) $g_{i,*}(\alpha)$, $\frac{\partial}{\partial \alpha^\top} g_{i,*}(\alpha)$, and $\frac{\partial^2}{\partial \alpha \partial \alpha^\top} g_{i,*}(\alpha)$ are uniformly bounded in $N_\delta(\alpha_0)$.

Under the assumptions (A1)–(A5) and conditions (C1)–(C3), we obtain the asymptotic results summarized in the following theorem.

Theorem 3.1.1. *Let* $\mathrm{TCF}_{10}(c_1), \mathrm{TCF}_{20}(c_1, c_2), \mathrm{TCF}_{30}(c_2)$ *be the true parameter values. The FI, MSI, IPW or SPE bias-corrected estimators* $\widehat{\mathrm{TCF}}_{1,*}(c_1), \widehat{\mathrm{TCF}}_{2,*}(c_1, c_2)$ *and* $\widehat{\mathrm{TCF}}_{3,*}(c_2)$ *are consistent. Furthermore,*

$$\sqrt{n}\left[\begin{pmatrix} \widehat{\mathrm{TCF}}_{1,*}(c_1) \\ \widehat{\mathrm{TCF}}_{2,*}(c_1, c_2) \\ \widehat{\mathrm{TCF}}_{3,*}(c_2) \end{pmatrix} - \begin{pmatrix} \mathrm{TCF}_{10}(c_1) \\ \mathrm{TCF}_{20}(c_1, c_2) \\ \mathrm{TCF}_{30}(c_2) \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}_3\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{\partial h(\alpha_0)}{\partial \alpha^\top} \Sigma \frac{\partial h^\top(\alpha_0)}{\partial \alpha^\top}\right), \quad (3.14)$$

where $h(\alpha) = \left(1 - \frac{\beta_{11}}{\theta_1}, \frac{\beta_{12} - \beta_{22}}{\theta_2}, \frac{\beta_{23}}{1 - (\theta_1 + \theta_2)}\right)^\top$ *and*

$$\Sigma = \left[\mathbb{E}\left\{\frac{\partial}{\partial \alpha^\top} g_{i,*}(\alpha_0)\right\}\right]^{-1} \mathbb{C}\mathrm{ov}\{g_{i,*}(\alpha_0)\} \left[\mathbb{E}\left\{\frac{\partial}{\partial \alpha^\top} g_{i,*}^\top(\alpha_0)\right\}\right]^{-1}.$$

Proof. We apply Theorem 1 and Theorem 2 of Alonzo et al. (2003). Under assumptions (A1)–(A5) and conditions (C1)–(C3), $\hat{\alpha}_*$ is consistent and $\sqrt{n}\,(\hat{\alpha}_* - \alpha_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. Thus, $\widehat{\mathrm{TCF}}_{1,*}(c_1) = 1 - \hat{\beta}_{11}/\hat{\theta}_1$, $\widehat{\mathrm{TCF}}_{2,*}(c_1, c_2) = (\hat{\beta}_{12} - \hat{\beta}_{22})/\hat{\theta}_2$ and $\widehat{\mathrm{TCF}}_{3,*}(c_2) = \hat{\beta}_{23}/(1 - (\hat{\theta}_1 + \hat{\theta}_2))$ are consistent for the true $\mathrm{TCF}_{10}(c_1)$, $\mathrm{TCF}_{20}(c_1, c_2)$ and $\mathrm{TCF}_{30}(c_2)$ and, by and application of the multivariate delta method, result (3.14) follows. In next parts, we check conditions (C1)–(C3) for each estimator, i.e., FI, MSI, IPW and SPE, under assumptions (A1)–(A5). This is done when a multinomial logistic regression model is used for the estimation of the disease process and a logistic regression model or a probit model is used for the estimation of the verification process. □

The above theorem gives a general result for all estimates, i.e., FI, MSI, IPW and SPE. In the last part, the explicit form of the asymptotic variance–covariance matrix is obtained. In practice, the variance–covariance matrix Σ is replaced by a consistent estimate $\hat{\Sigma}$

$$\hat{\Sigma} = n\left[\sum_{i=1}^{n} \frac{\partial}{\partial \alpha^\top} g_{i,*}(\hat{\alpha})\right]^{-1} \left[\sum_{i=1}^{n} g_{i,*}(\hat{\alpha}) g_{i,*}(\hat{\alpha})^\top\right] \left[\sum_{i=1}^{n} \frac{\partial}{\partial \alpha^\top} g_{i,*}^\top(\hat{\alpha})\right]^{-1}.$$

It is worth noting that SPE estimators of θ_k and β_{jk} in (3.10) and (3.11), will inherit the double robustness property of $\hat{\theta}_{\text{SPE}}$ and $\hat{\beta}_{1,\text{SPE}}$ in (2.6). That is, $\hat{\theta}_{k,\text{SPE}}$ and $\hat{\beta}_{jk,\text{SPE}}$ remain consistent if only one of the disease model $\Pr(D_k = 1|T, A)$ or the verification model $\Pr(V = 1|T, A)$ is correctly specified in the estimation process; they are inconsistent if both models are misspecified. Clearly, this property also holds for the estimators $\widehat{\text{TCF}}_{1,\text{SPE}}(c_1)$, $\widehat{\text{TCF}}_{2,\text{SPE}}(c_1, c_2)$ and $\widehat{\text{TCF}}_{3,\text{SPE}}(c_2)$.

Next, we discuss validity of conditions (C1), (C2) and (C3) for the proposed estimators. The discussion covers first the elements of the estimating functions corresponding to the parameter τ. Then, we pass on to the elements of the estimating functions corresponding to the parameters $\theta_1, \theta_2, \theta_{11}, \beta_{12}, \beta_{22}, \beta_{23}$, specializing the discussion to the various methods. Finally, we give the explicit form of the variance-covariance matrix in Theorem 3.1.1. Recall that α_0 denotes the true value of α.

Parameter τ. Note that estimators FI, MSI and SPE require a multinomial logistic or probit regression model to estimate the disease probabilities $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$ with $k = 1, 2, 3$. In the following, we adopt the multinomial logistic model, but arguments similar to those given below also hold for the multinomial probit model, despite the rather more complex algebra (see Daganzo 1979, Chap. 3, as a general reference).

The estimating function for the nuisance parameter $\tau \equiv \tau_\rho = \left(\tau_{\rho_1}^\top, \tau_{\rho_2}^\top\right)^\top$,

$$G^{\tau_\rho}(\alpha) = \left(G^{\tau_{\rho_1}}(\alpha)^\top, G^{\tau_{\rho_2}}(\alpha)^\top\right)^\top \equiv \left(\left(\sum_{i=1}^n g_i^{\tau_{\rho_1}}(\alpha)\right)^\top, \left(\sum_{i=1}^n g_i^{\tau_{\rho_2}}(\alpha)\right)^\top\right)^\top,$$

is obtained as the first derivative of the log likelihood function. With the multinomial logistic model, we get

$$G^{\tau_\rho}(\alpha) = \left(\left(\sum_{i=1}^n V_i U_i (D_{1i} - \rho_{1i})\right)^\top, \left(\sum_{i=1}^n V_i U_i (D_{2i} - \rho_{2i})\right)^\top\right)^\top,$$

where U_i is an arbitrary regressor; for simplicity, we take $U_i = (1, T_i, A_i^\top)^\top$. Under assumption (A2), condition (C1) trivially holds. Moreover, we get

$$\begin{aligned} \tfrac{\partial}{\partial \tau_{\rho_1}^\top} g_i^{\tau_{\rho_1}}(\alpha) &= -V_i U_i U_i^\top \rho_{1i}(1 - \rho_{1i}), & \tfrac{\partial}{\partial \tau_{\rho_2}^\top} g_i^{\tau_{\rho_1}}(\alpha) &= V_i U_i U_i^\top \rho_{1i}\rho_{2i}, \\ \tfrac{\partial}{\partial \tau_{\rho_2}^\top} g_i^{\tau_{\rho_2}}(\alpha) &= -V_i U_i U_i^\top \rho_{2i}(1 - \rho_{2i}), & \tfrac{\partial}{\partial \tau_{\rho_1}^\top} g_i^{\tau_{\rho_2}}(\alpha) &= V_i U_i U_i^\top \rho_{1i}\rho_{2i}, \end{aligned} \tag{3.15}$$

and $\frac{\partial}{\partial \theta_s} g_i^{\tau_\rho}(\alpha) = 0$, $\frac{\partial}{\partial \beta_{jk}} g_i^{\tau_\rho}(\alpha) = 0$ for each s, j, k. The second–order partial derivatives can be easily derived. Hence, for $G^{\tau_\rho}(\alpha)$, condition (C2) holds and, by assumption (A3)–(A5) condition (C3) also holds.

The IPW and SPE estimators require estimates of $\pi_i = \Pr(V_i = 1|T_i, A_i)$. With T and A as covariates, we can use the logistic or probit models to this end. In these cases, conditions (C1)–(C3) are satisfied by the score functions

$$G^{\tau_\pi}(\alpha) = \sum_{i=1}^n g_i^{\tau_\pi}(\alpha) = \sum_{i=1}^n U_i (V_i - \pi_i)$$

or

$$G^{\tau_\pi}(\alpha) = \sum_{i=1}^n g_i^{\tau_\pi}(\alpha) = \sum_{i=1}^n \left[\frac{V_i U_i \phi(U_i^\top \tau_\pi)}{\Phi(U_i^\top \tau_\pi)} - (1 - V_i)\frac{U_i \phi(U_i^\top \tau_\pi)}{1 - \Phi(U_i^\top \tau_\pi)}\right],$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal random variable, respectively. Recall that τ_π is the component of the nuisance parameter τ corresponding the model for estimating π. The first-order derivatives are

$$\frac{\partial}{\partial \tau_\pi^\top} g_i^{\tau_\pi}(\alpha) = -U_i U_i^\top \pi_i(1-\pi_i), \tag{3.16}$$

or

$$\begin{aligned}\frac{\partial}{\partial \tau_\pi^\top} g_i^{\tau_\pi}(\alpha) &= -\frac{V_i U_i U_i^\top \phi(U_i^\top \tau_\pi)\left[-U_i^\top \tau_\pi \Phi(U_i^\top \tau_\pi) - \phi(U_i^\top \tau_\pi)\right]}{\Phi^2(U_i^\top \tau_\pi)} \\ &\quad - (1-V_i)\frac{U_i U_i^\top \phi(U_i^\top \tau_\pi)\left[U_i^\top \tau_\pi(\Phi(U_i^\top \tau_\pi)-1) + \phi(U_i^\top \tau_\pi)\right]}{\left[1-\Phi(U_i^\top \tau_\pi)\right]^2}.\end{aligned} \tag{3.17}$$

FI and MSI estimators. According to equations (3.4), (3.5), (3.6) and (3.7), the estimating functions $G_*^{\theta_s}(\alpha)$ for FI and MSI estimators can be presented in the form

$$G_{\mathrm{IE}}^{\theta_s}(\alpha) \equiv \sum_{i=1}^n g_{i,\mathrm{IE}}^{\theta_s}(\alpha) = \sum_{i=1}^n \left\{V_i\left[mD_{si} - \theta_s + (1-m)\rho_{si}\right] + (1-V_i)(\rho_{si}-\theta_s)\right\},$$

with $s = 1, 2$. Similarly,

$$\begin{aligned}G_{\mathrm{IE}}^{\beta_{jk}}(\alpha) &\equiv \sum_{i=1}^n g_{i,\mathrm{IE}}^{\beta_{jk}}(\alpha) = \sum_{i=1}^n \Big\{V_i\left[m\mathrm{I}(T_i \geq c_j)D_{ki} - \beta_{jk} + (1-m)\mathrm{I}(T_i \geq c_j)\rho_{ki}\right] \\ &\quad + (1-V_i)\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk}\right]\Big\},\end{aligned}$$

for $j = 1, 2$ and $k = 1, 2, 3$. Here, the notation IE means "imputation estimator". The estimating function corresponds to the FI estimator if $m = 0$, to the MSI estimator if $m = 1$. Using the conditional expectation and the assumption (A1), $\mathbb{E}\left[g_{i,\mathrm{IE}}^{\theta_s}(\alpha_0)\right]$ equals

$$\begin{aligned}&\mathbb{E}_{D_s,T_i,A_i}\left[\mathbb{E}\left[g_{i,\mathrm{IE}}^{\theta_s}(\alpha_0)|T_i,A_i\right]\right] \\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\mathbb{E}\left[\left\{V_i\left[mD_{si} - \theta_{s0} + (1-m)\rho_{si}\right] + (1-V_i)\left[\rho_{si}-\theta_{s0}\right]\right\}|T_i,A_i\right]\right] \\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\pi_i\left[m\mathbb{E}\left[D_{si}|T_i,A_i\right] - \theta_{s0} + (1-m)\rho_{si}\right] + (1-\pi_i)(\rho_{si}-\theta_{s0})\right] \\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\pi_i\left[m\rho_{si} - \theta_{s0} + (1-m)\rho_{si}\right] + (1-\pi_i)(\rho_{si}-\theta_{s0})\right] \\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\pi_i(\rho_{si}-\theta_{s0}) + (1-\pi_i)(\rho_{si}-\theta_{s0})\right] \\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\rho_{si}-\theta_{s0}\right] = 0.\end{aligned}$$

Similarly, we compute the expected value of the estimating function components $g_{i,\mathrm{IE}}^{\beta_{jk}}(\alpha_0)$ as follows

$$\begin{aligned}&\mathbb{E}_{D_k,T_i,A_i}\left[\mathbb{E}\left[g_{i,\mathrm{IE}}^{\beta_{jk}}(\alpha_0)|T_i,A_i\right]\right] \\ &= \mathbb{E}_{D_{ki},T_i,A_i}\Bigg[\mathbb{E}\Bigg[\Big\{V_i\left[m\mathrm{I}(T_i \geq c_j)D_{ki} - \beta_{jk0} + (1-m)\mathrm{I}(T_i \geq c_j)\rho_{ki}\right] \\ &\quad + (1-V_i)\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0}\right]\Big\}\Big|T_i,A_i\Bigg]\Bigg] \\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\pi_i\left[m\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0} + (1-m)\mathrm{I}(T_i \geq c_j)\rho_{ki}\right] + (1-\pi_i)(\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0})\right] \\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\pi_i(\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0}) + (1-\pi_i)(\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0})\right] \\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk0}\right] = 0.\end{aligned}$$

Hence, under assumption (A2), condition (C1) holds for $G_{\mathrm{IE}}^{\theta_s}(\alpha)$ and $G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$.

We now verify conditions (C2) and (C3). The partial derivative of $G_{\mathrm{IE}}^{\theta_s}(\alpha)$ with respect to β_{jk} equals 0 for all j,k. Moreover,

$$\begin{aligned}\frac{\partial}{\partial\theta_{s'}}G_{\mathrm{IE}}^{\theta_s}(\alpha) &= \sum_{i=1}^{n}\frac{\partial}{\partial\theta_{s'}}\left\{V_i\left[mD_{si}-\theta_s+(1-m)\rho_{si}\right]+(1-V_i)\left[\rho_{si}-\theta_s\right]\right\}\\ &= \sum_{i=1}^{n}\mathrm{I}(s'=s)\left\{-V_i-(1-V_i)\right\}=-n\mathrm{I}(s'=s)\end{aligned}$$

and

$$\frac{\partial}{\partial\tau_\rho}G_{\mathrm{IE}}^{\theta_s}(\alpha)=\left(\left(\frac{\partial}{\partial\tau_{\rho_1}}G_{\mathrm{IE}}^{\theta_s}(\alpha)\right)^\top,\left(\frac{\partial}{\partial\tau_{\rho_2}}G_{\mathrm{IE}}^{\theta_s}(\alpha)\right)^\top\right)^\top.$$

For each $l=1,2$ and $s=1,2$, we have

$$\frac{\partial}{\partial\tau_{\rho_l}}G_{\mathrm{IE}}^{\theta_s}(\alpha)=\sum_{i=1}^{n}(1-mV_i)\frac{\partial}{\partial\tau_{\rho_l}}\rho_{si}.$$

Recall that, under the multinomial logistic model,

$$\rho_{si}=\frac{e^{U_i^\top\tau_s}}{1+e^{U_i^\top\tau_{\rho_1}}+e^{U_i^\top\tau_{\rho_2}}},\qquad s=1,2. \tag{3.18}$$

Thus, we obtain

$$\begin{aligned}\tfrac{\partial}{\partial\tau_{\rho_1}}\rho_{1i} &= U_i\rho_{1i}(1-\rho_{1i}),\quad \tfrac{\partial}{\partial\tau_{\rho_2}}\rho_{1i} &= -U_i\rho_{1i}\rho_{2i},\\ \tfrac{\partial}{\partial\tau_{\rho_2}}\rho_{2i} &= U_i\rho_{2i}(1-\rho_{2i}),\quad \tfrac{\partial}{\partial\tau_{\rho_1}}\rho_{2i} &= -U_i\rho_{1i}\rho_{2i}.\end{aligned} \tag{3.19}$$

The derivatives of $G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$ are

$$\frac{\partial}{\partial\theta_s}G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)=0,\qquad \frac{\partial}{\partial\beta_{j'k'}}G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)=-n\mathrm{I}(j'k'=jk)$$

and

$$\frac{\partial}{\partial\tau_{\rho_l}}G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)=\sum_{i=1}^{n}(1-mV_i)\mathrm{I}(T_i\geq c_j)\frac{\partial}{\partial\tau_{\rho_l}}\rho_{ki},$$

where $\frac{\partial}{\partial\tau_{\rho_l}}\rho_{si}$ is in (3.19). Hence, we have the explicit form of the partial derivatives of both $G_{\mathrm{IE}}^{\theta_s}(\alpha)$ and $G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$. The only not null elements of the second–order partial derivative of $G_{\mathrm{IE}}^{\theta_s}(\alpha)$ and $G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$ are those corresponding to the matrices $\frac{\partial^2}{\partial\tau\partial\tau^\top}G_{\mathrm{IE}}^{\theta_s}(\alpha)$ and $\frac{\partial^2}{\partial\tau\partial\tau^\top}G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$. These elements involve the derivatives with respect to τ of quantities in (3.19). It follows that conditions (C2) and (C3) hold for $G_{\mathrm{IE}}^{\theta_s}(\alpha)$ and $G_{\mathrm{IE}}^{\beta_{jk}}(\alpha)$ for each s,j,k.

IPW estimator. Recall that the estimating function for θ_s is

$$G_{\mathrm{IPW}}^{\theta_s}(\alpha)=\sum_{i=1}^{n}g_{i,\mathrm{IPW}}^{\theta_s}(\alpha)=\sum_{i=1}^{n}\frac{V_i}{\pi_i}\left(D_{si}-\theta_s\right)\quad s=1,2,$$

and for the parameter β_{jk} is

$$G_{\mathrm{IPW}}^{\beta_{jk}}(\alpha)=\sum_{i=1}^{n}g_{i,\mathrm{IPW}}^{\beta_{jk}}(\alpha)=\sum_{i=1}^{n}\frac{V_i}{\pi_i}\left(\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk}\right)\quad j=1,2;\ k=1,2,3.$$

We show that these estimating functions are unbiased under assumptions (A1) and (A2). In fact, we get

$$\begin{aligned}\mathbb{E}\left[V_i\pi_i^{-1}\left(D_{si}-\theta_{s0}\right)\right] &= \mathbb{E}_{D_s,T,A}\left[\mathbb{E}\left(V_i\pi_i^{-1}\left(D_{si}-\theta_{s0}\right)|,T_i,A_i\right)\right]\\ &= \mathbb{E}_{D_s,T,A}\left[\pi_i^{-1}\mathbb{E}\left(V_i|T_i,A_i\right)\left(\rho_{si}-\theta_{s0}\right)\right]\\ &= \mathbb{E}_{D_s,T,A}\left[\rho_{si}-\theta_{s0}\right]=0,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{D_k,T_i,A_i}\left[\mathbb{E}\left[g^{\beta_{jk}}_{i,\mathrm{IPW}}(\alpha_0)|T_i,A_i\right]\right] &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\mathbb{E}\left[\left\{\frac{V_i}{\pi_i}\left(\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk0}\right)\right\}\middle|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\mathrm{I}(T_i\geq c_j)\rho_{ik}-\beta_{jk0}\right]=0.\end{aligned}$$

Therefore, condition (C1) holds for $G^{\theta_s}_{\mathrm{IPW}}(\alpha)$ and $G^{\beta_{jk}}_{\mathrm{IPW}}(\alpha)$, all s,j,k.

Next, we obtain the partial derivatives

$$\begin{aligned}\tfrac{\partial}{\partial\theta_{s'}}G^{\theta_s}_{\mathrm{IPW}}(\alpha) &= -\sum_{i=1}^n\frac{V_i}{\pi_i}\mathrm{I}(s'=s), & \tfrac{\partial}{\partial\beta_{jk}}G^{\theta_s}_{\mathrm{IPW}}(\alpha) &= 0,\\ \tfrac{\partial}{\partial\theta_s}G^{\beta_{jk}}_{\mathrm{IPW}}(\alpha) &= 0, & \tfrac{\partial}{\partial\beta_{j'k'}}G^{\beta_{jk}}_{\mathrm{IPW}}(\alpha) &= -\sum_{i=1}^n\frac{V_i}{\pi_i}\mathrm{I}(j'k'=jk),\end{aligned}$$

and, for the logistic model (used to estimate the verification process)

$$\tfrac{\partial}{\partial\tau_\pi}G^{\theta_s}_{\mathrm{IPW}}(\alpha) = -\sum_{i=1}^n\frac{V_i(D_{si}-\theta_s)U_i}{e^{U_i^\top\tau_\pi}},\quad \tfrac{\partial}{\partial\tau_\pi}G^{\beta_{jk}}_{\mathrm{IPW}}(\alpha) = -\sum_{i=1}^n\frac{V_i(\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk})U_i}{e^{U_i^\top\tau_\pi}},$$

or the probit model

$$\begin{aligned}\frac{\partial}{\partial\tau_\pi}G^{\theta_s}_{\mathrm{IPW}}(\alpha) &= -\sum_{i=1}^n\frac{V_i(D_{si}-\theta_s)U_i\phi(U_i^\top\tau_\pi)}{\Phi^2(U_i^\top\tau_\pi)},\\ \frac{\partial}{\partial\tau_\pi}G^{\beta_{jk}}_{\mathrm{IPW}}(\alpha) &= -\sum_{i=1}^n\frac{V_i(\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk})U_i\phi(U_i^\top\tau_\pi)}{\Phi^2(U_i^\top\tau_\pi)}.\end{aligned}$$

The computation of the second-order derivatives is similar and the results imply that the conditions (C2) and (C3) hold.

SPE estimator. Recall that

$$G^{\theta_s}_{\mathrm{SPE}}(\alpha)=\sum_{i=1}^n\left\{\frac{V_i}{\pi_i}\left(D_{si}-\theta_s\right)-\frac{V_i-\pi_i}{\pi_i}\left(\rho_{si}-\theta_s\right)\right\},\quad s=1,2,$$

$$G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha)=\sum_{i=1}^n\left\{\frac{V_i}{\pi_i}\left[\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk}\right]-\frac{V_i-\pi_i}{\pi_i}\left[\mathrm{I}(T_i\geq c_j)\rho_{ki}-\beta_{jk}\right]\right\},\quad j=1,2;\ k=1,2,3.$$

Under assumption (A1), $\mathbb{E}\left[g^{\theta_k}_{i,\mathrm{SPE}}(\alpha_0)\right]$ equals

$$\begin{aligned}&\mathbb{E}_{D_s,T_i,A_i}\left[\mathbb{E}\left[g^{\theta_s}_{i,\mathrm{SPE}}(\alpha_0)|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\mathbb{E}\left[\left\{\frac{V_i}{\pi_i}\left(D_{si}-\theta_{s0}\right)-\frac{V_i-\pi_i}{\pi_i}\left(\rho_{si}-\theta_{s0}\right)\right\}|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\pi_i^{-1}\left[\mathbb{E}\left[D_{si}|T_i,A_i\right]-\theta_{s0}\right]\pi_i\right]-\pi_i^{-1}\mathbb{E}_{D_{si},T_i,A_i}\left[\mathbb{E}\left[(V_i-\pi_i)(\rho_{si}-\theta_{s0})|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\rho_{si}-\theta_{s0}\right]-\pi_i^{-1}\mathbb{E}_{D_{si},T_i,A_i}\left[(\rho_{si}-\theta_{s0})\mathbb{E}\left[(V_i-\pi_i)|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{si},T_i,A_i}\left[\rho_{si}-\theta_{s0}\right]=0.\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}_{D_k,T_i,A_i}\left[\mathbb{E}\left[g^{\beta_{jk}}_{i,\mathrm{SPE}}(\alpha_0)|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\mathbb{E}\left[\left\{\frac{V_i}{\pi_i}\left[\mathrm{I}(T_i\geq c_j)D_{ki}-\beta_{jk0}\right]-\frac{V_i-\pi_i}{\pi_i}\left[\mathrm{I}(T_i\geq c_j)\rho_{ki}-\beta_{jk0}\right]\right\}\middle|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\left[\mathrm{I}(T_i\geq c_j)\mathbb{E}\left[D_{ki}|T_i,A_i\right]-\beta_{jk0}\right]\right]\\ &\quad-\pi_i^{-1}\mathbb{E}_{D_{ki},T_i,A_i}\left[\mathbb{E}\left[(V_i-\pi_i)(\mathrm{I}(T_i\geq c_j)\rho_{ki}-\beta_{jk0})|T_i,A_i\right]\right]\\ &= \mathbb{E}_{D_{ki},T_i,A_i}\left[\mathrm{I}(T_i\geq c_j)\rho_{ki}-\beta_{jk0}\right]=0.\end{aligned}$$

Therefore, condition (C1) holds for $G^{\theta_s}_{\mathrm{SPE}}(\alpha)$ and $G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha)$, all s, j, k.

Next, we obtain the partial derivatives

$$\begin{aligned} \tfrac{\partial}{\partial\theta_{s'}} G^{\theta_s}_{\mathrm{SPE}}(\alpha) &= -n\mathrm{I}(s'=s) & \qquad \tfrac{\partial}{\partial\beta_{jk}} G^{\theta_s}_{\mathrm{SPE}}(\alpha) &= 0 \\ \tfrac{\partial}{\partial\theta_{s}} G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha) &= 0 & \qquad \tfrac{\partial}{\partial\beta_{j'k'}} g^{\beta_{jk}}_{\mathrm{SPE}}(\alpha) &= -n\mathrm{I}(j'k'=jk) \end{aligned}$$

and the partial derivative with respect to $\tau_\rho \equiv (\tau_{\rho_1}^\top, \tau_{\rho_2}^\top)^\top$

$$\frac{\partial}{\partial\tau_{\rho_l}} G^{\theta_s}_{\mathrm{SPE}}(\alpha) = \sum_{i=1}^{n} -\frac{V_i-\pi_i}{\pi_i}\frac{\partial}{\partial\tau_{\rho_l}}\rho_{si}; \qquad \frac{\partial}{\partial\tau_{\rho_l}} G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha) = \sum_{i=1}^{n} -\frac{V_i-\pi_i}{\pi_i}\mathrm{I}(T_i\geq c_j)\frac{\partial}{\partial\tau_{\rho_l}}\rho_{si},$$

where $\frac{\partial}{\partial\tau_{\rho_l}^\top}\rho_{si}$ is given in (3.19). The partial derivative with respect to τ_π, are

$$\frac{\partial}{\partial\tau_{\pi}} G^{\theta_s}_{\mathrm{SPE}}(\alpha) = \sum_{i=1}^{n} \frac{V_iU_i(\rho_{si}-D_{si})}{e^{U_i^\top\tau_\pi}}; \qquad \frac{\partial}{\partial\tau_{\pi}} G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha) = \sum_{i=1}^{n} \frac{V_iU_i\mathrm{I}(T_i\geq c_j)(\rho_{ki}-D_{ki})}{e^{U_i^\top\tau}};$$

when the logistic model is used for the verification process. If the probit model is used, we have

$$\begin{aligned} \frac{\partial}{\partial\tau_{\pi}} G^{\theta_s}_{\mathrm{SPE}}(\alpha) &= \sum_{i=1}^{n} \frac{V_iU_i(D_{si}-\rho_{si})\phi(U_i^\top\tau_\pi)}{\Phi^2(U_i^\top\tau_\pi)}, \\ \frac{\partial}{\partial\tau_{\pi}} G^{\beta_{jk}}_{\mathrm{SPE}}(\alpha) &= \sum_{i=1}^{n} \frac{V_iU_i\mathrm{I}(T_i\geq c_j)(D_{si}-\rho_{ki})\phi(U_i^\top\tau_\pi)}{\Phi^2(U_i^\top\tau_\pi)}. \end{aligned}$$

Also in this case, computation of the second–order partial derivatives develops similarly and the results imply that the conditions (C2) and (C3) hold.

Asymptotic covariance matrix. Recall that the asymptotic covariance matrix of TCF estimators is obtained as

$$\frac{\partial h(\alpha_0)}{\partial\alpha^\top}\Sigma\frac{\partial h^\top(\alpha_0)}{\partial\alpha^\top},$$

where $h(\alpha) = \left(1-\frac{\beta_{11}}{\theta_1}, \frac{\beta_{12}-\beta_{22}}{\theta_2}, \frac{\beta_{23}}{1-(\theta_1+\theta_2)}\right)^\top$ and

$$\Sigma = \left[\mathbb{E}\left\{\frac{\partial}{\partial\alpha^\top}g_{i,*}(\alpha_0)\right\}\right]^{-1}\mathbb{E}\{g_{i,*}(\alpha_0)g_{i,*}(\alpha_0)^\top\}\left[\mathbb{E}\left\{\frac{\partial}{\partial\alpha^\top}g_{i,*}^\top(\alpha_0)\right\}\right]^{-1}.$$

It is easy to derive that

$$\frac{\partial h(\alpha)}{\partial\alpha^\top} = \begin{pmatrix} \frac{\beta_{11}}{\theta_1^2} & 0 & -\frac{1}{\theta_1} & 0 & 0 & 0 & 0 \\ 0 & -\frac{\beta_{12}-\beta_{22}}{\theta_2^2} & 0 & \frac{1}{\theta_2} & -\frac{1}{\theta_2} & 0 & 0 \\ \frac{\beta_{23}}{(1-\theta_1-\theta_2)^2} & \frac{\beta_{23}}{(1-\theta_1-\theta_2)^2} & 0 & 0 & 0 & \frac{1}{1-\theta_1-\theta_2} & 0 \end{pmatrix}.$$

The elements $g_{i,*}(\alpha)$ of the estimating functions $G_*(\alpha)$ are given in the previous paragraphs. Now, we derive the explicit form for $\frac{\partial}{\partial\alpha^\top}g_{i,*}(\alpha)$.

First, we consider the class of imputation estimators. We get

$$\begin{aligned}
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\theta_1}(\alpha) &= \left(-1,0,0,0,0,0,A_{11i}^\top,A_{21i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\theta_2}(\alpha) &= \left(0,-1,0,0,0,0,A_{12i}^\top,A_{22i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\beta_{11}}(\alpha) &= \left(0,0,-1,0,0,0,B_{111i}^\top,B_{121i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\beta_{12}}(\alpha) &= \left(0,0,0,-1,0,0,B_{112i}^\top,B_{122i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\beta_{22}}(\alpha) &= \left(0,0,0,0,-1,0,B_{212i}^\top,B_{222i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IE}}^{\beta_{23}}(\alpha) &= \left(0,0,0,0,0,-1,B_{213i}^\top,B_{223i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha^\top}g_{i,\mathrm{IE}}^{\tau_{\rho_1}}(\alpha) &= \left(0,0,0,0,0,0,C_{11i},C_{21i}\right),\\
\frac{\partial}{\partial\alpha^\top}g_{i,\mathrm{IE}}^{\tau_{\rho_2}}(\alpha) &= \left(0,0,0,0,0,0,C_{12i},C_{22i}\right),
\end{aligned}$$

where

$$A_{lsi}=(1-mV_i)\frac{\partial}{\partial\tau_{\rho_l}}\rho_{si},\quad B_{jlki}=(1-mV_i)\mathrm{I}(T_i\geq c_j)\frac{\partial}{\partial\tau_{\rho_l}}\rho_{ki},\quad C_{lsi}=\frac{\partial}{\partial\tau_{\rho_l}^\top}g_i^{\tau_{\rho_s}}(\alpha),$$

with $j,l,s=1,2$ and $k=1,2,3$ (see (3.15) and (3.19) for the multinomial logistic modeling of the disease process). Thus,

$$\frac{\partial}{\partial\alpha^\top}g_{i,\mathrm{IE}}(\alpha)=\begin{pmatrix}
-1 & 0 & 0 & 0 & 0 & 0 & A_{11i}^\top & A_{21i}^\top\\
0 & -1 & 0 & 0 & 0 & 0 & A_{12i}^\top & A_{22i}^\top\\
0 & 0 & -1 & 0 & 0 & 0 & B_{111i}^\top & B_{121i}^\top\\
0 & 0 & 0 & -1 & 0 & 0 & B_{112i}^\top & B_{122i}^\top\\
0 & 0 & 0 & 0 & -1 & 0 & B_{212i}^\top & B_{222i}^\top\\
0 & 0 & 0 & 0 & 0 & -1 & B_{213i}^\top & B_{223i}^\top\\
0 & 0 & 0 & 0 & 0 & 0 & C_{11i} & C_{21i}\\
0 & 0 & 0 & 0 & 0 & 0 & C_{12i} & C_{22i}
\end{pmatrix}.$$

Then, we consider the IPW estimators. Let

$$A_{ki}=\frac{\partial}{\partial\tau_\pi}g_{i,\mathrm{IPW}}^{\theta_k}(\alpha),\qquad B_{jki}=\frac{\partial}{\partial\tau_\pi}g_{i,\mathrm{IPW}}^{\beta_{jk}}(\alpha),\qquad C_i=\frac{\partial}{\partial\tau_\pi^\top}g_i^{\tau_\pi}(\alpha).$$

Note that these quantities change according to the model, logit or probit, chosen for the verification process. We obtain

$$\begin{aligned}
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\theta_1}(\alpha) &= \left(-V_i\pi^{-1},0,0,0,0,0,A_{1i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\theta_2}(\alpha) &= \left(0,-V_i\pi^{-1},0,0,0,0,A_{2i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\beta_{11}}(\alpha) &= \left(0,0,-V_i\pi^{-1},0,0,0,B_{11i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\beta_{12}}(\alpha) &= \left(0,0,0,-V_i\pi^{-1},0,0,B_{12i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\beta_{22}}(\alpha) &= \left(0,0,0,0,-V_i\pi^{-1},0,B_{22i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha}g_{i,\mathrm{IPW}}^{\beta_{23}}(\alpha) &= \left(0,0,0,0,0,-V_i\pi^{-1},B_{23i}^\top\right)^\top,\\
\frac{\partial}{\partial\alpha^\top}g_i^{\tau_\pi}(\alpha) &= \left(0,0,0,0,0,0,C_i\right).
\end{aligned}$$

Summarizing

$$\frac{\partial}{\partial \alpha^\top} g_{i,\mathrm{IPW}}(\alpha) = \begin{pmatrix} -V_i\pi_i^{-1} & 0 & 0 & 0 & 0 & 0 & A_{1i}^\top \\ 0 & -V_i\pi_i^{-1} & 0 & 0 & 0 & 0 & A_{2i}^\top \\ 0 & 0 & -V_i\pi_i^{-1} & 0 & 0 & 0 & B_{11i}^\top \\ 0 & 0 & 0 & -V_i\pi_i^{-1} & 0 & 0 & B_{12i}^\top \\ 0 & 0 & 0 & 0 & -V_i\pi_i^{-1} & 0 & B_{22i}^\top \\ 0 & 0 & 0 & 0 & 0 & -V_i\pi_i^{-1} & B_{23i}^\top \\ 0 & 0 & 0 & 0 & 0 & 0 & C_i \end{pmatrix}.$$

Finally, we consider the SPE estimators. We have

$$\begin{aligned}
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\theta_1}(\alpha) &= \left(-1,0,0,0,0,0,H_{11i}^\top,H_{21i}^\top,D_{1i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\theta_2}(\alpha) &= \left(0,-1,0,0,0,0,H_{12i}^\top,H_{22i}^\top,D_{2i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\beta_{11}}(\alpha) &= \left(0,0,-1,0,0,0,G_{111i}^\top,G_{121i}^\top,E_{11i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\beta_{12}}(\alpha) &= \left(0,0,0,-1,0,0,G_{112i}^\top,G_{122i}^\top,E_{12i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\beta_{22}}(\alpha) &= \left(0,0,0,0,-1,0,G_{212i}^\top,G_{222i}^\top,E_{22i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha} g_{i,\mathrm{SPE}}^{\beta_{23}}(\alpha) &= \left(0,0,0,0,0,-1,G_{213i}^\top,G_{223i}^\top,E_{23i}^\top\right)^\top, \\
\frac{\partial}{\partial \alpha^\top} g_{i,\mathrm{SPE}}^{\tau_{\rho_1}}(\alpha) &= \left(0,0,0,0,0,0,C_{11i},C_{21i},0\right), \\
\frac{\partial}{\partial \alpha^\top} g_{i,\mathrm{SPE}}^{\tau_{\rho_2}}(\alpha) &= \left(0,0,0,0,0,0,C_{12i},C_{22i},0\right), \\
\frac{\partial}{\partial \alpha^\top} g_{i,\mathrm{SPE}}^{\tau_{\pi}}(\alpha) &= \left(0,0,0,0,0,0,0,0,C_i\right),
\end{aligned}$$

where

$$\begin{aligned}
H_{lki} &= -\frac{V_i-\pi_i}{\pi_i}\frac{\partial}{\partial \tau_{\rho_l}}\rho_{ki}, & G_{jlki} &= -\frac{V_i-\pi_i}{\pi_i}\mathrm{I}(T_i \geq c_j)\frac{\partial}{\partial \tau_{\rho_l}}\rho_{ki}, \\
D_{si} &= \frac{\partial}{\partial \tau_\pi} g_{i,\mathrm{SPE}}^{\theta_s}(\alpha), & E_{jki} &= \frac{\partial}{\partial \tau_\pi} g_{i,\mathrm{SPE}}^{\beta_{jk}}(\alpha),
\end{aligned}$$

and C_{lsi} and C_i are defined above. Therefore

$$\frac{\partial}{\partial \alpha^\top} g_{i,\mathrm{SPE}}(\alpha) = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & H_{11i}^\top & H_{21i}^\top & D_{1i}^\top \\ 0 & -1 & 0 & 0 & 0 & 0 & H_{12i}^\top & H_{22i}^\top & D_{2i}^\top \\ 0 & 0 & -1 & 0 & 0 & 0 & G_{111i}^\top & G_{121i}^\top & E_{11i}^\top \\ 0 & 0 & 0 & -1 & 0 & 0 & G_{112i}^\top & G_{122i}^\top & E_{12i}^\top \\ 0 & 0 & 0 & 0 & -1 & 0 & G_{212i}^\top & G_{222i}^\top & E_{22i}^\top \\ 0 & 0 & 0 & 0 & 0 & -1 & G_{213i}^\top & G_{223i}^\top & E_{23i}^\top \\ 0 & 0 & 0 & 0 & 0 & 0 & C_{11i} & C_{21i} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & C_{12i} & C_{22i} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & C_i \end{pmatrix}.$$

3.2 Nonparametric estimation

All the verification bias-corrected estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1,c_2)$ and $\mathrm{TCF}_3(c_2)$ revised in the previous section belong to the class of (partially) parametric estimators, i.e., they need regression models to estimate $\rho_{ki} = \Pr(D_{ki}=1|T_i,A_i)$ and/or $\pi_i = \Pr(V_i=1|T_i,A_i)$. In what follows, we propose a fully nonparametric approach to the estimation of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1,c_2)$ and $\mathrm{TCF}_3(c_2)$. Our approach is based on the K-nearest neighbor (KNN) imputation method.

3.2.1 The proposed method

Hereafter, we shall assume that A is a continuous random variable. Recall that the true disease status is a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$ such that D_k is a n Bernoulli trials with success probability $\theta_k = \Pr(D_k = 1)$, $k = 1, 2, 3$. Note that $\theta_1 + \theta_2 + \theta_3 = 1$. Let $\beta_{jk} = \Pr(T \geq c_j, D_k = 1)$ with $j = 1, 2$ and $k = 1, 2, 3$. Since parameters θ_k are the means of the random variables D_k, we can use the KNN imputation discussed in Ning and Cheng (2012) to obtain nonparametric estimates $\hat{\theta}_{k,\text{KNN}}$. More precisely, we define

$$\hat{\theta}_{k,\text{KNN}} = \frac{1}{n}\sum_{i=1}^{n}\left[V_i D_{ki} + (1 - V_i)\hat{\rho}_{ki,K}\right], \qquad K \in \mathbb{N},$$

where $\hat{\rho}_{ki,K} = \dfrac{1}{K}\sum_{l=1}^{K} D_{ki(l)}$, and $\left\{(T_{i(l)}, A_{i(l)}, D_{ki(l)}) : V_{i(l)} = 1, l = 1, \ldots, K\right\}$ is a set of K observed data pairs and $(T_{i(l)}, A_{i(l)})$ denotes the j-th nearest neighbor to (T_i, A_i) among all (T, A)'s corresponding to the verified patients, i.e., to those D_{kh}'s with $V_h = 1$. Similarly, we can define the KNN estimates of β_{jk} as follows

$$\hat{\beta}_{jk,\text{KNN}} = \frac{1}{n}\sum_{i=1}^{n}\text{I}(T_i \geq c_j)\left[V_i D_{ki} + (1 - V_i)\hat{\rho}_{ki,K}\right],$$

each j, k. Therefore, the KNN imputation estimators for TCF_k are

$$\begin{aligned}
\widehat{\text{TCF}}_{1,\text{KNN}}(c_1) &= \frac{\sum_{i=1}^{n}\text{I}(T_i < c_1)\left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i,K}\right]}{\sum_{i=1}^{n}\left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i,K}\right]}, \\
\widehat{\text{TCF}}_{2,\text{KNN}}(c_1, c_2) &= \frac{\sum_{i=1}^{n}\text{I}(c_1 \leq T_i < c_2)\left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i,K}\right]}{\sum_{i=1}^{n}\left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i,K}\right]}, \\
\widehat{\text{TCF}}_{3,\text{KNN}}(c_2) &= \frac{\sum_{i=1}^{n}\text{I}(T_i \geq c_2)\left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i,K}\right]}{\sum_{i=1}^{n}\left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i,K}\right]}.
\end{aligned} \tag{3.20}$$

3.2.2 Asymptotic distribution

Let $\rho_k(t, a) = \Pr(D_k = 1 | T = t, A = a)$ and $\pi(t, a) = \Pr(V = 1 | T = t, A = a)$. The KNN imputation estimators of $\text{TCF}_1(c_1)$, $\text{TCF}_2(c_1, c_2)$ and $\text{TCF}_3(c_2)$ are consistent and asymptotically normal. In fact, we have the following theorems.

Theorem 3.2.1. *Assume the functions $\rho_k(t, a)$ and $\pi(t, a)$ are finite and first-order differentiable. Moreover, assume that the expectation of $1/\pi(T, A)$ exists. Then, for a fixed pair cut of points (c_1, c_2) such that $c_1 < c_2$, the KNN imputation estimators $\widehat{\text{TCF}}_{1,\text{KNN}}(c_1)$, $\widehat{\text{TCF}}_{2,\text{KNN}}(c_1, c_2)$ and $\widehat{\text{TCF}}_{3,\text{KNN}}(c_2)$ are consistent.*

Proof. Since the disease status D_k is a Bernoulli random variable, its second-order moment, $\mathbb{E}(D_k^2)$, is finite. According to the first assumption, we can show that the conditional variance of D_k given the test results T and A, $\mathbb{V}\text{ar}(D_k | T = t, A = a)$ is equal to $\rho_k(t, a)\left[1 - \rho_k(t, a)\right]$ and is clearly finite. Thus, by an application of Theorem 1 in Ning and Cheng (2012), the KNN imputation estimators $\hat{\theta}_{k,\text{KNN}}$ are consistent.

Now, observe that,

$$\begin{aligned}
&\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)\left[V_i D_{ki} + (1-V_i)\rho_{ki}\right] + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) - \beta_{jk} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)V_i\left[D_{ki} - \rho_{ki}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) \\
&= S_{jk} + R_{jk} + T_{jk}.
\end{aligned}$$

Here, the quantities R_{jk}, S_{jk} and T_{jk} are similar to the quantities R, S and T in the proof of Theorem 2.1 in Cheng (1994) and Theorem 1 in Ning and Cheng (2012). Thus, we have that

$$\sqrt{n}R_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right) \qquad \text{and} \qquad \sqrt{n}S_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right),$$

where $\delta_{jk}^2(T,A)$ is the conditional variance of $\mathrm{I}(T \geq c_j, D_k = 1)$ given T, A. Also, by using a similar technique to that of proof of Theorem 1 in Ning and Cheng (2012), we get $T_{jk} = W_{jk} + o_p(n^{-1/2})$, where

$$W_{jk} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{ki(l)} - \rho_{ki(l)}\right)\right].$$

Moreover, $\mathbb{E}(W_{jk}) = 0$ and the asymptotic variance is:

$$\mathrm{as}\mathbb{V}\mathrm{ar}(\sqrt{n}W_{jk}) = \frac{1}{K}\mathbb{E}\left[(1-\pi(T,A))\delta_{jk}^2(T,A)\right] + \mathbb{E}\left[\frac{(1-\pi(T,A))^2\delta_{jk}^2(T,A)}{\pi(T,A)}\right].$$

Then, the Markov's inequality implies that $W_{jk} \xrightarrow{p} 0$ as n goes to infinity. This, together with the fact that R_{jk} and S_{jk} converge in probability to zero, leads to the consistency of $\hat{\beta}_{jk,\mathrm{KNN}}$, i.e, $\hat{\beta}_{jk,\mathrm{KNN}} \xrightarrow{p} \beta_{jk}$. It follows that $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) = 1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) = \frac{\hat{\beta}_{12}-\hat{\beta}_{22}}{\hat{\theta}_2}$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) = \frac{\hat{\beta}_{23}}{\hat{\theta}_3}$ are consistent. □

Theorem 3.2.2. *Assume that the conditions in Theorem 3.2.1 hold, we get*

$$\sqrt{n}\left[\begin{pmatrix} \widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) \\ \widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) \\ \widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) \end{pmatrix} - \begin{pmatrix} \mathrm{TCF}_1(c_1) \\ \mathrm{TCF}_2(c_1,c_2) \\ \mathrm{TCF}_3(c_2) \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}(0,\Xi), \tag{3.21}$$

where Ξ *is a suitable matrix.*

Proof. A direct application of Theorem 1 in Ning and Cheng (2012) gives the result that the quantity $\sqrt{n}(\hat{\theta}_{k,\mathrm{KNN}} - \theta_k)$ converges to a normal random variable with mean 0 and variance $\sigma_k^2 = \left[\theta_k(1-\theta_k) + \omega_k^2\right]$. Here,

$$\begin{aligned}
\omega_k^2 &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left[\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))\right] \\
&\quad + \mathbb{E}\left[\frac{\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))^2}{\pi(T,A)}\right].
\end{aligned} \tag{3.22}$$

In addition, from the proof of Theorem 3.2.1, we have

$$\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} \simeq S_{jk} + R_{jk} + W_{jk} + o_p(n^{-1/2}),$$

with

$$\sqrt{n}R_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right), \quad \sqrt{n}S_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right)$$

and

$$\sqrt{n}W_{jk} \xrightarrow{d} \mathcal{N}(0, \sigma_{W_{jk}}^2).$$

Therefore, $\sqrt{n}(\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk}) \xrightarrow{d} \mathcal{N}(0, \sigma_{jk}^2)$. Here, the asymptotic variance σ_{jk}^2 is obtained by

$$\sigma_{jk}^2 = \beta_{jk}\left(1 - \beta_{jk}\right) + \omega_{jk}^2,$$

with

$$\begin{aligned} \omega_{jk}^2 &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)(1 - \rho_k(T,A))(1 - \pi(T,A))\right] \\ &\quad + \mathbb{E}\left[\frac{\mathrm{I}(T \geq c_j)\rho_k(T,A)(1 - \rho_k(T,A))(1 - \pi(T,A))^2}{\pi(T,A)}\right]. \end{aligned} \tag{3.23}$$

This result follows by the fact that R_{jk} and $S_{jk} + W_{jk}$ are uncorrelated and the asymptotic covariance between S_{jk} and W_{jk} is obtained by

$$\mathrm{as}\mathbb{C}\mathrm{ov}\left(S_{jk}, W_{jk}\right) = \mathbb{E}\left[(1 - \pi(T,A))\delta_{jk}^2(T,A)\right].$$

Moreover, we get that the vector $\sqrt{n}(\hat{\theta}_{1,\mathrm{KNN}}, \hat{\theta}_{2,\mathrm{KNN}}, \hat{\beta}_{11,\mathrm{KNN}}, \hat{\beta}_{12,\mathrm{KNN}}, \hat{\beta}_{22,\mathrm{KNN}}, \hat{\beta}_{23,\mathrm{KNN}})^\top$ is jointly asymptotically normally distributed with mean vector $(\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}, \beta_{23})^\top$ and suitable covariance matrix Ξ^*. Then, result (3.21) follows by applying the multivariate delta method to

$$h(\hat{\theta}_1, \hat{\theta}_2, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{22}, \hat{\beta}_{23}) = \left(1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}, \frac{(\hat{\beta}_{12} - \hat{\beta}_{22})}{\hat{\theta}_2}, \frac{\hat{\beta}_{23}}{(1 - \hat{\theta}_1 - \hat{\theta}_2)}\right)^\top.$$

The asymptotic covariance matrix of $\sqrt{n}(\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}, \widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}, \widehat{\mathrm{TCF}}_{3,\mathrm{KNN}})^\top$, Ξ, is obtained by

$$\Xi = h'\Xi^* h'^\top, \tag{3.24}$$

where h' is the first-order derivative of h, i.e.,

$$h' = \begin{pmatrix} \frac{\beta_{11}}{\theta_1^2} & 0 & -\frac{1}{\theta_1} & 0 & 0 & 0 \\ 0 & -\frac{(\beta_{12}-\beta_{22})}{\theta_2^2} & 0 & \frac{1}{\theta_2} & -\frac{1}{\theta_2} & 0 \\ \frac{\beta_{23}}{(1-\theta_1-\theta_2)^2} & \frac{\beta_{23}}{(1-\theta_1-\theta_2)^2} & 0 & 0 & 0 & \frac{1}{(1-\theta_1-\theta_2)} \end{pmatrix}. \tag{3.25}$$

□

3.2.3 The asymptotic covariance matrix

Let

$$\Xi = \begin{pmatrix} \xi_1^2 & \xi_{12} & \xi_{13} \\ \xi_{12} & \xi_2^2 & \xi_{23} \\ \xi_{13} & \xi_{23} & \xi_3^2 \end{pmatrix}.$$

The asymptotic covariance matrix Ξ^* is a 6×6 matrix such that its diagonal elements are the asymptotic variances of $\sqrt{n}\hat{\theta}_{k,\mathrm{KNN}}$ and $\sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}}$. Let us define $\sigma_{12}^* = \mathrm{as}\mathbb{C}\mathrm{ov}(\sqrt{n}\hat{\theta}_{1,\mathrm{KNN}}, \sqrt{n}\hat{\theta}_{2,\mathrm{KNN}})$, $\sigma_{sjk} = \mathrm{as}\mathbb{C}\mathrm{ov}(\sqrt{n}\hat{\theta}_{s,\mathrm{KNN}}, \sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}})$ and $\sigma_{jkls} = \mathrm{as}\mathbb{C}\mathrm{ov}(\sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}}, \sqrt{n}\hat{\beta}_{ls,\mathrm{KNN}})$. We write

$$\Xi^* = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^* & \sigma_{111} & \sigma_{112} & \sigma_{122} & \sigma_{123} \\ \sigma_{12}^* & \sigma_2^2 & \sigma_{211} & \sigma_{212} & \sigma_{222} & \sigma_{223} \\ \sigma_{111} & \sigma_{211} & \sigma_{11}^2 & \sigma_{1112} & \sigma_{1122} & \sigma_{1123} \\ \sigma_{112} & \sigma_{212} & \sigma_{1112} & \sigma_{12}^2 & \sigma_{1222} & \sigma_{1223} \\ \sigma_{122} & \sigma_{222} & \sigma_{1122} & \sigma_{1222} & \sigma_{22}^2 & \sigma_{2223} \\ \sigma_{123} & \sigma_{223} & \sigma_{1123} & \sigma_{1223} & \sigma_{2223} & \sigma_{23}^2 \end{pmatrix}.$$

Hence, from (3.24) and (3.25),

$$\begin{aligned}
\xi_1^2 = \text{asVar}\left(\sqrt{n}\widehat{\text{TCF}}_{1,\text{KNN}}(c_1)\right) &= \frac{\beta_{11}^2}{\theta_1^4}\sigma_1^2 + \frac{\sigma_{11}^2}{\theta_1^2} - 2\frac{\beta_{11}}{\theta_1^3}\sigma_{111}, \\
\xi_2^2 = \text{asVar}\left(\sqrt{n}\widehat{\text{TCF}}_{2,\text{KNN}}(c_1,c_2)\right) &= \sigma_2^2\frac{(\beta_{12}-\beta_{22})^2}{\theta_2^4} + \frac{\sigma_{12}^2+\sigma_{22}^2-2\sigma_{1222}}{\theta_2^2} \\
&\quad - 2\frac{\beta_{12}-\beta_{22}}{\theta_2^3}(\sigma_{212}-\sigma_{222}), \\
\xi_3^2 = \text{asVar}\left(\sqrt{n}\widehat{\text{TCF}}_{3,\text{KNN}}(c_2)\right) &= \frac{\beta_{23}^2}{(1-\theta_1-\theta_2)^4}\left(\sigma_1^2+2\sigma_{12}^*+\sigma_2^2\right) + \frac{\sigma_{23}^2}{(1-\theta_1-\theta_2)^2} \\
&\quad + 2\frac{\beta_{23}}{(1-\theta_1-\theta_2)^3}\left(\sigma_{123}+\sigma_{223}\right). \qquad (3.26)
\end{aligned}$$

Let $\lambda^2 = \text{asVar}(\sqrt{n}\hat{\beta}_{12,\text{KNN}} - \sqrt{n}\hat{\beta}_{22,\text{KNN}})$. Hence, $\sigma_{12}^2+\sigma_{22}^2-2\sigma_{1222} = \lambda^2$, and

$$\xi_2^2 = \sigma_2^2\frac{(\beta_{12}-\beta_{22})^2}{\theta_2^4} + \frac{\lambda^2}{\theta_2^2} - 2\frac{\beta_{12}-\beta_{22}}{\theta_2^3}(\sigma_{212}-\sigma_{222}).$$

Observe that $\hat{\theta}_{3,\text{KNN}} = 1-(\hat{\theta}_{1,\text{KNN}}+\hat{\theta}_{2,\text{KNN}})$. Thus,

$$\begin{aligned}
\text{asVar}(\sqrt{n}\hat{\theta}_{3,\text{KNN}}) &= \text{asVar}(\sqrt{n}\hat{\theta}_{1,\text{KNN}} + \sqrt{n}\hat{\theta}_{2,\text{KNN}}) \\
&= \text{asVar}(\sqrt{n}\hat{\theta}_{1,\text{KNN}}) + \text{asVar}(\sqrt{n}\hat{\theta}_{2,\text{KNN}}) + 2\text{asCov}(\sqrt{n}\hat{\theta}_{1,\text{KNN}}, \sqrt{n}\hat{\theta}_{2,\text{KNN}}).
\end{aligned}$$

This leads to the expression $\sigma_3^2 = \sigma_1^2 + 2\sigma_{12}^* + \sigma_2^2$. In addition,

$$\begin{aligned}
\sigma_{123}+\sigma_{223} &= \text{asCov}(\sqrt{n}\hat{\theta}_{1,\text{KNN}}, \sqrt{n}\hat{\beta}_{23,\text{KNN}}) + \text{asCov}(\sqrt{n}\hat{\theta}_{2,\text{KNN}}, \sqrt{n}\hat{\beta}_{23,\text{KNN}}) \\
&= \text{asCov}(\sqrt{n}\hat{\theta}_{1,\text{KNN}} + \sqrt{n}\hat{\theta}_{2,\text{KNN}}, \sqrt{n}\hat{\beta}_{23,\text{KNN}}) \\
&= -\text{asCov}(\sqrt{n} - (\sqrt{n}\hat{\theta}_{1,\text{KNN}} + \sqrt{n}\hat{\theta}_{2,\text{KNN}}), \sqrt{n}\hat{\beta}_{23,\text{KNN}}) \\
&= -\sigma_{323}.
\end{aligned}$$

Therefore, from (3.26), the asymptotic variance of $\sqrt{n}\widehat{\text{TCF}}_{3,\text{KNN}}(c_2)$ is

$$\xi_3^2 = \frac{\beta_{23}^2\sigma_3^2}{(1-\theta_1-\theta_2)^4} + \frac{\sigma_{23}^2}{(1-\theta_1-\theta_2)^2} - 2\frac{\beta_{23}\sigma_{323}}{(1-\theta_1-\theta_2)^3}.$$

Recall that $\sigma_k^2 = \left[\theta_k(1-\theta_k)+\omega_k^2\right]$ and $\sigma_{jk}^2 = \left[\beta_{jk}(1-\beta_{jk})+\omega_{jk}^2\right]$, where ω_k^2 and ω_{jk}^2 are given in (3.22) and (3.23), respectively. To obtain σ_{kjk}, we observe that

$$\begin{aligned}
\beta_{jk} &= \Pr(T \geq c_j, D_k = 1) = \Pr(D_k = 1)\Pr(T \geq c_j | D_k = 1) \\
&= \Pr(D_k = 1)\left[1 - \Pr(T < c_j | D_k = 1)\right] \\
&= \Pr(D_k = 1) - \Pr(D_k = 1)\Pr(T < c_j | D_k = 1) \\
&= \Pr(D_k = 1) - \Pr(T < c_j, D_k = 1) \\
&= \theta_k - \gamma_{jk},
\end{aligned}$$

for $k = 1,2,3$ and $j = 1,2$. Then, we define

$$\hat{\gamma}_{jk,\text{KNN}} = \frac{1}{n}\sum_{i=1}^{n}\text{I}(T_i < c_j)\left[V_i D'_{ki} + (1-V_i)\hat{\rho}_{ki,K}\right].$$

The asymptotic variance of $\sqrt{n}\hat{\gamma}_{jk,\text{KNN}}$, ζ_{jk}^2, is obtained as that of $\sqrt{n}\hat{\beta}_{jk,\text{KNN}}$. In fact, we get $\zeta_{jk}^2 = \left[\gamma_{jk}(1-\gamma_{jk})+\eta_{jk}^2\right]$, where

$$\begin{aligned}
\eta_{jk}^2 &= \frac{K+1}{K}\mathbb{E}\left[\text{I}(T<c_j)\rho_k(T,A)\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}\right] \\
&\quad + \mathbb{E}\left[\frac{\text{I}(T<c_j)\rho_k(T,A)\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}^2}{\pi(T,A)}\right].
\end{aligned}$$

It is straightforward to see that $\hat{\gamma}_{jk,\mathrm{KNN}} = \hat{\theta}_{k,\mathrm{KNN}} - \hat{\beta}_{jk,\mathrm{KNN}}$. Thus, we can compute the asymptotic covariances σ_{kjk} for $j = 1, 2$ and $k = 1, 2, 3$, using the fact that

$$\begin{aligned}\mathrm{as}\mathbb{V}\mathrm{ar}(\sqrt{n}\hat{\gamma}_{jk,\mathrm{KNN}}) &= \mathrm{as}\mathbb{V}\mathrm{ar}(\sqrt{n}\hat{\theta}_{k,\mathrm{KNN}} - \sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}}) \\ &= \mathrm{as}\mathbb{V}\mathrm{ar}(\sqrt{n}\hat{\theta}_{k,\mathrm{KNN}}) + \mathrm{as}\mathbb{V}\mathrm{ar}(\sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}}) - 2\mathrm{as}\mathbb{C}\mathrm{ov}(\sqrt{n}\hat{\theta}_{k,\mathrm{KNN}}, \sqrt{n}\hat{\beta}_{jk,\mathrm{KNN}}).\end{aligned}$$

This leads to

$$\sigma_{kjk} = \frac{1}{2}\left(\sigma_k^2 + \sigma_{jk}^2 - \zeta_{jk}^2\right).$$

Hence,

$$\begin{aligned}\sigma_{111} &= \frac{1}{2}\left(\sigma_1^2 + \sigma_{11}^2 - \zeta_{11}^2\right); & \sigma_{212} &= \frac{1}{2}\left(\sigma_2^2 + \sigma_{12}^2 - \zeta_{12}^2\right); \\ \sigma_{222} &= \frac{1}{2}\left(\sigma_2^2 + \sigma_{22}^2 - \zeta_{22}^2\right); & \sigma_{323} &= \frac{1}{2}\left(\sigma_3^2 + \sigma_{23}^2 - \zeta_{23}^2\right).\end{aligned}$$

As for λ^2, one can show that

$$\lambda^2 = (\beta_{12} - \beta_{22})\left[1 - (\beta_{12} - \beta_{22})\right] + \omega_{12}^2 - \omega_{22}^2. \tag{3.27}$$

In fact, according the proof of Theorem 3.2.2, we have

$$\left(\hat{\beta}_{12,\mathrm{KNN}} - \hat{\beta}_{22,\mathrm{KNN}}\right) - (\beta_{12} - \beta_{22}) \simeq (S_{12} - S_{22}) + (R_{12} - R_{22}) + (W_{12} - W_{22}) + o_p(n^{-1/2}). \tag{3.28}$$

Here, we have

$$\begin{aligned}S_{12} - S_{22} &= \frac{1}{n}\sum_{i=1}^{n} V_i \mathrm{I}(c_1 \le T_i < c_2)\left(D_{2i} - \rho_{2i}\right), \\ R_{12} - R_{22} &= \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{I}(c_1 \le T_i < c_2)\rho_{2i} - (\beta_{12} - \beta_{22})\right], \\ W_{12} - W_{22} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(c_1 \le T_i < c_2)(1 - V_i)\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{2i(l)} - \rho_{2i(l)}\right)\right].\end{aligned}$$

Under that, we realize that quantities $S_{12} - S_{22}$ and S_{jk}, so as $R_{12} - R_{22}$ and R_{jk}, and $W_{12} - W_{22}$ and W_{jk}, play, in essence, a similar role. Therefore, the quantities in the right hand side of equation (3.28) have approximately normal distributions with mean 0 and variances

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\sqrt{n}(S_{12} - S_{22})\right) &= \mathbb{E}\left\{\pi(T, A)\delta^2(T, A)\right\}, \\ \mathbb{V}\mathrm{ar}\left(\sqrt{n}(R_{12} - R_{22})\right) &= \mathbb{V}\mathrm{ar}\left[\mathrm{I}(c_1 \le T_i < c_2)\rho_2(T, A)\right], \\ \mathbb{V}\mathrm{ar}\left(\sqrt{n}(W_{12} - W_{22})\right) &= \frac{1}{K}\mathbb{E}\left[(1 - \pi(T, A))\delta^2(T, A)\right] + \mathbb{E}\left[\frac{(1 - \pi(T, A))^2\delta^2(T, A)}{\pi(T, A)}\right].\end{aligned}$$

where, $\delta^2(T, A)$ is the conditional variance of $\mathrm{I}(c_1 \le T_i < c_2, D_{2i} = 1)$ given T, A. Then, we get

$$\sqrt{n}\left[\left(\hat{\beta}_{12,\mathrm{KNN}} - \hat{\beta}_{22,\mathrm{KNN}}\right) - (\beta_{12} - \beta_{22})\right] \xrightarrow{d} \mathcal{N}(0, \lambda^2).$$

To obtain λ^2, we notice that the quantities $R_{12} - R_{22}$ and $(S_{12} - S_{22}) + (W_{12} - W_{22})$ are uncorrelated and the asymptotic covariance of $S_{12} - S_{22}$ and $W_{12} - W_{22}$ equals to $\mathbb{E}\left[(1 - \pi(T, A))\delta^2(T, A)\right]$. Taking the sum of this covariance and the above variances, the desired asymptotic variance λ^2 is approximately as (3.27).

Therefore, suitable explicit expressions for the asymptotic variances of KNN estimators can be found. Such expressions will depend on quantities as θ_k, β_{jk} ω_k^2, ω_{jk}^2, γ_{jk} and η_{jk}^2 only. As

a consequence, to obtain consistent estimates of the asymptotic variances, ultimately we need to estimate the quantities $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2.

In the last paragraph, we show that suitable expressions can be obtained also for the elements ξ_{12}, ξ_{13} and ξ_{23} of the covariance matrix Ξ. Such expressions will depend, among others, on certain quantities ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2 similar to ω_k^2, ω_{jk}^2 or η_{jk}^2.

Here, we focus on the elements ξ_{12}, ξ_{13} and ξ_{23} of the covariance matrix Ξ. We can write

$$\begin{aligned}
\xi_{12} &= -\frac{1}{\theta_1\theta_2}\left(\sigma_{1112}-\sigma_{1122}\right)+\frac{\beta_{11}}{\theta_1^2\theta_2}\left(\sigma_{112}-\sigma_{122}\right)-\frac{\beta_{12}-\beta_{22}}{\theta_2^2}\left(\frac{\beta_{11}}{\theta_1^2}\sigma_{12}^*-\frac{\sigma_{211}}{\theta_1}\right), \quad (3.29)\\
\xi_{13} &= \frac{1}{1-\theta_1-\theta_2}\left(\frac{\beta_{11}}{\theta_1^2}\sigma_{123}-\frac{\sigma_{1123}}{\theta_1}\right)\\
&\quad+\frac{\beta_{23}}{\theta_1(1-\theta_1-\theta_2)^2}\left[\frac{\beta_{11}}{\theta_1}\left(\sigma_1^2+\sigma_{12}^*\right)-\left(\sigma_{111}+\sigma_{211}\right)\right], \quad (3.30)
\end{aligned}$$

and

$$\begin{aligned}
\xi_{23} &= \frac{1}{\theta_2(1-\theta_1-\theta_2)}\left[\left(\sigma_{1223}-\sigma_{2223}\right)-\frac{\beta_{12}-\beta_{22}}{\theta_2}\sigma_{223}\right]\\
&\quad+\frac{\beta_{23}}{\theta_2(1-\theta_1-\theta_2)^2}\left[\left(\sigma_{112}-\sigma_{122}+\sigma_{212}-\sigma_{222}\right)-\frac{\beta_{12}-\beta_{22}}{\theta_2}\left(\sigma_2^2+\sigma_{12}^*\right)\right]. \quad (3.31)
\end{aligned}$$

Recall that

$$\begin{aligned}
\hat{\theta}_{k,\mathrm{KNN}}-\theta_k &= \frac{1}{n}\sum_{i=1}^{n}\left[V_iD_{ki}+(1-V_i)\rho_{ki}\right]+\frac{1}{n}\sum_{i=1}^{n}(1-V_i)(\hat{\rho}_{ki,K}-\rho_{ki})-\theta_k\\
&= \frac{1}{n}\sum_{i=1}^{n}V_i\left[D_{ki}-\rho_{ki}\right]+\frac{1}{n}\sum_{i=1}^{n}\left[\rho_{ki}-\theta_k\right]\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{ki(l)}-\rho_{ki(l)}\right)\right]+o_p\left(n^{-1/2}\right)\\
&= S_k+R_k+W_k+o_p\left(n^{-1/2}\right);
\end{aligned}$$

and

$$\begin{aligned}
&\hat{\beta}_{jk,\mathrm{KNN}}-\beta_{jk}\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)\left[V_iD_{ki}+(1-V_i)\rho_{ki}\right]+\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)(1-V_i)(\hat{\rho}_{ki,K}-\rho_{ki})-\beta_{jk}\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)V_i\left[D_{ki}-\rho_{ki}\right]+\frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{I}(T_i\geq c_j)\rho_{ki}-\beta_{jk}\right]\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)(1-V_i)\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{ki(l)}-\rho_{ki(l)}\right)\right]+o_p\left(n^{-1/2}\right)\\
&= S_{jk}+R_{jk}+W_{jk}+o_p\left(n^{-1/2}\right).
\end{aligned}$$

Considering the term $\sigma_{1112} - \sigma_{1122}$ in (3.29), we have

$$\begin{aligned}
\sigma_{1112} - \sigma_{1122} &= \text{asCov}(\sqrt{n}\hat{\beta}_{11,\text{KNN}}, \sqrt{n}\hat{\beta}_{12,\text{KNN}}) - \text{asCov}(\sqrt{n}\hat{\beta}_{11,\text{KNN}}, \sqrt{n}\hat{\beta}_{22,\text{KNN}}) \\
&= \text{asCov}(\sqrt{n}\hat{\beta}_{11,\text{KNN}}, \sqrt{n}\hat{\beta}_{12,\text{KNN}} - \sqrt{n}\hat{\beta}_{22,\text{KNN}}) \\
&= \text{asCov}\big(\sqrt{n}(S_{11} + R_{11} + W_{11}), \sqrt{n}(S_{12} - S_{22}) + \sqrt{n}(R_{12} - R_{22}) \\
&\quad + \sqrt{n}(W_{12} - W_{22})\big) \\
&= \text{asCov}\left(\sqrt{n}S_{11}, \sqrt{n}(S_{12} - S_{22})\right) + \text{asCov}\left(\sqrt{n}S_{11}, \sqrt{n}(W_{12} - W_{22})\right) \\
&\quad + \text{asCov}\left(\sqrt{n}R_{11}, \sqrt{n}(R_{12} - R_{22})\right) + \text{asCov}\left(\sqrt{n}W_{11}, \sqrt{n}(S_{12} - S_{22})\right) \\
&\quad + \text{asCov}\left(\sqrt{n}W_{11}, \sqrt{n}(W_{12} - W_{22})\right).
\end{aligned}$$

This result follows from the fact that $\sqrt{n}R_{11}$ and $\sqrt{n}(S_{12} - S_{22}) + \sqrt{n}(W_{12} - W_{22})$, and $\sqrt{n}(S_{11} + W_{11})$ and $\sqrt{n}(R_{12} - R_{22})$ are uncorrelated (see also Cheng, 1994). By arguments similar to those used in Ning and Cheng (2012), we also obtain

$$\begin{aligned}
\text{asCov}\left(\sqrt{n}S_{11}, \sqrt{n}(S_{12} - S_{22})\right) &= \mathbb{E}\left\{\pi(T,A)\text{Cov}(\text{I}(T \geq c_1)D_1, \text{I}(c_1 \leq T < c_2)D_2|T,A)\right\} \\
&= \mathbb{E}\left\{\pi(T,A)\text{I}(c_1 \leq T < c_2)\text{Cov}(D_1, D_2|T,A)\right\} \\
&= -\mathbb{E}\left\{\pi(T,A)\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}.
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
\text{asCov}\left(\sqrt{n}S_{11}, \sqrt{n}(W_{12} - W_{22})\right) &= -\mathbb{E}\left\{[1 - \pi(T,A)]\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}R_{11}, \sqrt{n}(R_{12} - R_{22})\right) &= -\beta_{11}(\beta_{12} - \beta_{22}) + \mathbb{E}\left\{\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}W_{11}, \sqrt{n}(S_{12} - S_{22})\right) &= -\mathbb{E}\left\{[1 - \pi(T,A)]\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}W_{11}, \sqrt{n}(W_{12} - W_{22})\right) &= -\frac{1}{K}\mathbb{E}\left\{[1 - \pi(T,A)]\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\} \\
&\quad -\mathbb{E}\left\{\frac{[1 - \pi(T,A)]^2\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\}.
\end{aligned}$$

This leads to

$$\sigma_{1112} - \sigma_{1122} = -\psi_{1212}^2 - \beta_{11}(\beta_{12} - \beta_{22}), \tag{3.32}$$

where

$$\begin{aligned}
\psi_{1212}^2 &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1 - \pi(T,A)]\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\} \\
&\quad + \mathbb{E}\left\{\frac{[1 - \pi(T,A)]^2\text{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\}.
\end{aligned}$$

Considering $\sigma_{112} - \sigma_{122}$ in (3.29), we have

$$\begin{aligned}
\sigma_{112} - \sigma_{122} &= \text{asCov}(\sqrt{n}\hat{\theta}_{1,\text{KNN}}, \sqrt{n}\hat{\beta}_{12,\text{KNN}}) - \text{asCov}(\sqrt{n}\hat{\theta}_{1,\text{KNN}}, \sqrt{n}\hat{\beta}_{22,\text{KNN}}) \\
&= \text{asCov}\left(\sqrt{n}\hat{\theta}_{1,\text{KNN}}, \sqrt{n}(\hat{\beta}_{12,\text{KNN}} - \hat{\beta}_{22,\text{KNN}})\right) \\
&= \text{asCov}\left(\sqrt{n}(S_1 + R_1 + W_1), \sqrt{n}(S_{12} - S_{22}) + \sqrt{n}(R_{12} - R_{22}) + \sqrt{n}(W_{12} - W_{22})\right) \\
&= \text{asCov}\left(\sqrt{n}S_1, \sqrt{n}(S_{12} - S_{22})\right) + \text{asCov}\left(\sqrt{n}S_1, \sqrt{n}(W_{12} - W_{22})\right) \\
&\quad + \text{asCov}\left(\sqrt{n}R_1, \sqrt{n}(R_{12} - R_{22})\right) + \text{asCov}\left(\sqrt{n}W_1, \sqrt{n}(S_{12} - S_{22})\right) \\
&\quad + \text{asCov}\left(\sqrt{n}W_1, \sqrt{n}(W_{12} - W_{22})\right).
\end{aligned}$$

We obtain

$$\begin{aligned}
\text{asCov}\left(\sqrt{n}S_1, \sqrt{n}(S_{12} - S_{22})\right) &= -\mathbb{E}\left\{\pi(T,A)\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}S_1, \sqrt{n}(W_{12} - W_{22})\right) &= -\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}R_1, \sqrt{n}(R_{12} - R_{22})\right) &= -\theta_1(\beta_{12} - \beta_{22}) + \mathbb{E}\left\{\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}W_1, \sqrt{n}(S_{12} - S_{22})\right) &= -\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\}, \\
\text{asCov}\left(\sqrt{n}W_1, \sqrt{n}(W_{12} - W_{22})\right) &= -\frac{1}{K}\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)\right\} \\
&\quad -\mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(c_1 \leq T < c_2)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},
\end{aligned}$$

and then

$$\sigma_{112} - \sigma_{122} = -\psi^2_{1212} - \theta_1(\beta_{12} - \beta_{22}). \tag{3.33}$$

Similarly, it is straightforward to obtain

$$\sigma_{211} = -\psi^2_{112} - \theta_2\beta_{11} \tag{3.34}$$

and

$$\sigma_{123} = -\psi^2_{213} - \theta_1\beta_{23}, \tag{3.35}$$

with

$$\begin{aligned}
\psi^2_{112} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \geq c_1)\rho_1(T,A)\rho_2(T,A)\right\} \\
&\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \geq c_1)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\}
\end{aligned}$$

and

$$\begin{aligned}
\psi^2_{213} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \geq c_2)\rho_1(T,A)\rho_3(T,A)\right\} \\
&\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \geq c_2)\rho_1(T,A)\rho_3(T,A)}{\pi(T,A)}\right\}.
\end{aligned}$$

The covariance between $\sqrt{n}\hat{\theta}_{1,\text{KNN}}$ and $\sqrt{n}\hat{\theta}_{2,\text{KNN}}$ is computed analogously, i.e.,

$$\sigma^*_{12} = -\theta_1\theta_2 - \psi^2_{12}, \tag{3.36}$$

where

$$\begin{aligned}
\psi^2_{12} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\rho_1(T,A)\rho_2(T,A)\right\} \\
&\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\}.
\end{aligned}$$

By using results (3.32), (3.33), (3.34) and (3.36) into (3.29), we can obtain a suitable expression for $\text{asCov}\left(\sqrt{n}\widehat{\text{TCF}}_{1,\text{KNN}}(c_1), \sqrt{n}\widehat{\text{TCF}}_{2,\text{KNN}}(c_1,c_2)\right)$, which depends on easily estimable quanties.

Clearly, a similar approach can be used to get suitable expressions for ξ_{13} and ξ_{23} too. In particular, the estimable version of ξ_{13} can be obtained by using suitable expressions for σ_{123}, σ_{1123} and $\sigma_{111} + \sigma_{211}$. The quantity σ_{123} is already computed in (3.35), and the formula for σ_{1123} can be obtained as

$$\sigma_{1123} = -\psi^2_{213} - \beta_{11}\beta_{23}.$$

To compute $\sigma_{111} + \sigma_{211}$, we notice that

$$\begin{aligned} \text{as}\mathbb{C}\text{ov}\left(\sqrt{n}\hat{\theta}_{3,\text{KNN}}, \sqrt{n}\hat{\beta}_{11,\text{KNN}}\right) &= \text{as}\mathbb{C}\text{ov}\left(\sqrt{n} - \sqrt{n}(\hat{\theta}_{1,\text{KNN}} + \hat{\theta}_{1,\text{KNN}}), \sqrt{n}\hat{\beta}_{11,\text{KNN}}\right) \\ &= -\text{as}\mathbb{C}\text{ov}\left(\sqrt{n}(\hat{\theta}_{1,\text{KNN}} + \hat{\theta}_{1,\text{KNN}}), \sqrt{n}\hat{\beta}_{11,\text{KNN}}\right). \end{aligned}$$

It leads to $\sigma_{111} + \sigma_{211} = -\sigma_{311}$. Similarly to (3.34), we have that

$$\sigma_{311} = -\psi^2_{113} - \theta_3\beta_{11},$$

where

$$\begin{aligned} \psi^2_{113} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1 - \pi(T, A)]\text{I}(T \geq c_1)\rho_1(T, A)\rho_3(T, A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1 - \pi(T, A)]^2\text{I}(T \geq c_1)\rho_1(T, A)\rho_3(T, A)}{\pi(T, A)}\right\}. \end{aligned}$$

For the last term ξ_{23}, we need to make some other calculations. First, the quantity $\sigma_{1223} - \sigma_{2223}$ is obtained as $\sigma_{1112} - \sigma_{1122}$. We have

$$\sigma_{1223} - \sigma_{2223} = -\beta_{23}(\beta_{12} - \beta_{22}),$$

because $\text{I}(c_1 \leq T < c_2)\text{I}(T \geq c_2) = 0$. Second, the term σ_{223} is obtained as

$$\sigma_{223} = -\psi^2_{223} - \theta_2\beta_{23},$$

where

$$\begin{aligned} \psi^2_{223} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1 - \pi(T, A)]\text{I}(T \geq c_2)\rho_2(T, A)\rho_3(T, A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1 - \pi(T, A)]^2\text{I}(T \geq c_2)\rho_2(T, A)\rho_3(T, A)}{\pi(T, A)}\right\}. \end{aligned}$$

Moreover, it is straightforward to show that

$$-(\sigma_{312} - \sigma_{322}) = \sigma_{112} - \sigma_{122} + \sigma_{212} - \sigma_{222},$$

and that

$$\sigma_{312} - \sigma_{322} = -\psi^2_{1223} - \theta_3(\beta_{12} - \beta_{22}),$$

with

$$\begin{aligned} \psi^2_{1223} &= \left(1 + \frac{1}{K}\right)\mathbb{E}\left\{[1 - \pi(T, A)]\text{I}(c_1 \leq T < c_2)\rho_2(T, A)\rho_3(T, A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1 - \pi(T, A)]^2\text{I}(c_1 \leq T < c_2)\rho_2(T, A)\rho_3(T, A)}{\pi(T, A)}\right\}. \end{aligned}$$

3.2.4 Choice of K and the distance measure

The proposed method is based on nearest-neighbor imputation, which requires the choice of a value for K as well as a distance measure.

In practice, the selection of a suitable distance is typically dictated by features of the data and possible subjective evaluations; thus, a general indication about an adequate choice is difficult to express. In many cases, the simple Euclidean distance may be appropriate. Other times,

the researcher may wish to consider specific characteristics of data at hand, and then make a different choice. For example, the diagnostic test result T and the auxiliary covariates A could be heterogeneous with respect to their variances (which is particularly true when the variables are measured on heterogeneous scales). In this case, the choice of the Mahalanobis distance may be suitable.

As for the choice of the size of the neighborhood, Ning and Cheng (2012) argue that nearest-neighbor imputation whit a small value of K typically yields negligible bias of the estimators, but a large variance; the opposite happens with a large value of K. The authors suggest that the choice of $K \in \{1, 2\}$ is generally adequate when the aim is to estimate an average. A similar comment is also raised by Adimari and Chiogna (2015, 2016), i.e., a small value of K, within the range 1–3, may be a good choice to estimate ROC curves and AUC. However, the authors stress that, in general, the choice of K may depend on the dimension of the feature space, and propose to use cross–validation to find K in case of high–dimensional covariates. Specifically, the authors indicate that a suitable value for the number of neighbors could be found by

$$K^* = \underset{K=1,\ldots,n_{ver}}{\arg\min} \frac{1}{n_{ver}} \left\| D - \hat{\rho}_K \right\|_1 ,$$

where D is the binary disease status, $\|\cdot\|_1$ denotes the L_1 norm for vectors and n_{ver} is the number of verified subjects. The formula above can be generalized to our multi–class case. In fact, when the disease status $\mathcal{D}$ has q categories ($q \geq 3$), the difference between $\mathcal{D}$ and $\hat{\rho}_K$ is a $n_{ver} \times (q-1)$ matrix. In such situations, the selection rule could be

$$K^* = \underset{K=1,\ldots,n_{ver}}{\arg\min} \frac{1}{n_{ver}(q-1)} \left\| \mathcal{D} - \hat{\rho}_K \right\|_{1,1} , \tag{3.37}$$

where $\|\mathcal{A}\|_{1,1}$ denotes $L_{1,1}$ norm of matrix $\mathcal{A}$, i.e.,

$$\|\mathcal{A}\|_{1,1} = \sum_{j=1}^{q-1} \left(\sum_{i=1}^{n_{ver}} |a_{ij}| \right) .$$

3.2.5 Variance-covariance estimation

Consider first the problem of estimating of the variances of $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}$. In a nonparametric framework, quantities as $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2 can be estimated by their empirical counterparts, using also the plug–in method. Here, we consider an approach that uses a nearest-neighbor rule to estimate both the functions $\rho_k(T, A)$ and the propensity score $\pi(T, A)$, that are present in the expressions of $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. In particular, for the conditional probabilities of disease, we can use KNN estimates $\tilde{\rho}_{ki} = \hat{\rho}_{ki,\bar{K}}$, where the integer $\bar{K}$ must be greater than one to avoid estimates equal to zero. For the conditional probabilities of verification, we can resort to the KNN procedure proposed in Adimari and Chiogna (2015), which considers the estimates

$$\tilde{\pi}_i = \frac{1}{K_i^*} \sum_{l=1}^{K_i^*} V_{i(l)},$$

where $\left\{ (T_{i(l)}, A_{i(l)}, V_{i(l)}) : l = 1, \ldots, K_i^* \right\}$ is a set of K_i^* observed pairs and $(T_{i(l)}, A_{i(l)})$ denotes the j-th nearest neighbor to (T_i, A_i) among all (T, A)'s. When V_i equals 0, K_i^* is set equal to the rank of the first verified nearest neighbor to the unit i, i.e., K_i^* is such that $V_{i(K_i^*)} = 1$ and $V_i = V_{i(1)} = V_{i(2)} = \ldots = V_{i(K_i^*-1)} = 0$. In case of $V_i = 1$, K_i^* is such that $V_i = V_{i(1)} = V_{i(2)} =$

$\ldots = V_{i(K_i^*-1)} = 1$, and $V_{i(K_i^*)} = 0$, i.e., K_i^* is set equal to the rank of the first non–verified nearest neighbor to the unit i. Such a procedure automatically avoids zero values for the $\tilde{\pi}_i$'s.

Then, based on the $\tilde{\rho}_{ki}$'s and $\tilde{\pi}_i$'s, we obtain the estimates

$$\begin{aligned}
\hat{\omega}_k^2 &= \frac{K+1}{nK}\sum_{i=1}^{n}\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)+\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)^2}{\tilde{\pi}_i},\\
\hat{\omega}_{jk}^2 &= \frac{K+1}{nK}\sum_{i=1}^{n}\mathrm{I}(T_i\geq c_j)\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{I}(T_i\geq c_j)\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)^2}{\tilde{\pi}_i},\\
\hat{\eta}_{jk}^2 &= \frac{K+1}{nK}\sum_{i=1}^{n}\mathrm{I}(T_i< c_j)\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{I}(T_i< c_j)\tilde{\rho}_{ki}\left(1-\tilde{\rho}_{ki}\right)\left(1-\tilde{\pi}_i\right)^2}{\tilde{\pi}_i},
\end{aligned}$$

from which, along with $\hat{\theta}_{k,\mathrm{KNN}}$, $\hat{\beta}_{jk,\mathrm{KNN}}$ and $\hat{\gamma}_{jk,\mathrm{KNN}}$, one derives the estimates of the variances of the proposed KNN imputation estimators.

To obtain estimates of covariances, we need to estimate also the quantities ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2 given in Appendix 2. However, estimates of such quantities are similar to those given above for $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. For example,

$$\begin{aligned}
\hat{\psi}_{1212}^2 &= \frac{K+1}{nK}\sum_{i=1}^{n}\mathrm{I}(c_1\leq T_i< c_2)\tilde{\rho}_{1i}\tilde{\rho}_{2i}\left(1-\tilde{\pi}_i\right)\\
&\quad+\frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{I}(c_1\leq T_i< c_2)\tilde{\rho}_{1i}\tilde{\rho}_{2i}\left(1-\tilde{\pi}_i\right)^2}{\tilde{\pi}_i}.
\end{aligned}$$

Of course, there are other possible approaches to obtain variance and covariance estimates. For instance, one could resort to a standard bootstrap procedure. From the original observations $(T_i, A_i, D_{1i}, D_{2i}, D_{3i}, V_i)$, $i = 1, \ldots, n$, consider B bootstrap samples $(T_i^{*b}, A_i^{*b}, D_{1i}^{*b}, D_{2i}^{*b}, D_{3i}^{*b}, V_i^{*b})$, $b = 1, \ldots, B$, and $i = 1, \ldots, n$. For the b-th sample, compute the bootstrap estimates $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}^{*b}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}^{*b}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}^{*b}(c_2)$ as

$$\begin{aligned}
\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}^{*b}(c_1) &= \frac{\sum_{i=1}^{n}\mathrm{I}(T_i^{*b}< c_1)\left[V_i^{*b}D_{1i}^{*b}+(1-V_i^{*b})\hat{\rho}_{1i,K}^{*b}\right]}{\sum_{i=1}^{n}\left[V_i^{*b}D_{1i}^{*b}+(1-V_i^{*b})\hat{\rho}_{1i,K}^{*b}\right]},\\
\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}^{*b}(c_1, c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(c_1\leq T_i^{*b}< c_2)\left[V_i^{*b}D_{2i}^{*b}+(1-V_i^{*b})\hat{\rho}_{2i,K}^{*b}\right]}{\sum_{i=1}^{n}\left[V_i^{*b}D_{2i}^{*b}+(1-V_i^{*b})\hat{\rho}_{2i,K}^{*b}\right]},\\
\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}^{*b}(c_2) &= \frac{\sum_{i=1}^{n}\mathrm{I}(T_i^{*b}\geq c_2)\left[V_i^{*b}D_{3i}^{*b}+(1-V_i^{*b})\hat{\rho}_{3i,K}^{*b}\right]}{\sum_{i=1}^{n}\left[V_i^{*b}D_{3i}^{*b}+(1-V_i^{*b})\hat{\rho}_{3i,K}^{*b}\right]},
\end{aligned}$$

where $\hat{\rho}_{ki,K}^{*b}$, $k = 1, 2, 3$, denote the KNN imputation values for missing labels D_{ki}^{*b} in the bootstrap sample. Then, the bootstrap estimator of the variance of $\widehat{\mathrm{TCF}}_{k,\mathrm{KNN}}(c_1, c_2)$ is

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\mathrm{TCF}}_{k,\mathrm{KNN}}(c_1, c_2)) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\mathrm{TCF}}_{k,\mathrm{KNN}}^{*b}(c_1, c_2)-\widehat{\mathrm{TCF}}_{k,\mathrm{KNN}}^{*}(c_1, c_2)\right)^2,$$

where $\widehat{\text{TCF}}^{*}_{k,\text{KNN}}(c_1, c_2)$ is the mean of the B bootstrap estimates $\widehat{\text{TCF}}^{*b}_{k,\text{KNN}}(c_1, c_2)$. More generally, the bootstrap estimate of the covariance matrix Ξ is

$$\widehat{\Xi}_B = \frac{1}{B-1}\left(\widehat{\text{TCF}}^{*B}_{\text{KNN}}(c_1, c_2) - \widehat{\text{TCF}}^{*}_{\text{KNN}}(c_1, c_2)\right)\left(\widehat{\text{TCF}}^{*B}_{\text{KNN}}(c_1, c_2) - \widehat{\text{TCF}}^{*}_{\text{KNN}}(c_1, c_2)\right)^{\top},$$

where $\widehat{\text{TCF}}^{*B}_{\text{KNN}}(c_1, c_2)$ is a $B \times 3$ matrix, whose element in the b–th row and the k–th column corresponds to $\widehat{\text{TCF}}^{*b}_{k,\text{KNN}}(c_1, c_2)$, and $\widehat{\text{TCF}}^{*}_{\text{KNN}}(c_1, c_2)$ is a column vector that consist of the means of the B bootstrap estimates $\widehat{\text{TCF}}^{*b}_{k,\text{KNN}}(c_1, c_2)$, $k = 1, 2, 3$.

3.3 Simulation studies

3.3.1 Simulation studies for the parametric approaches

In this section, the ability of FI, MSI, IPW and SPE methods to estimate TCF_1, TCF_2 and TCF_3 are evaluated by using Monte Carlo experiments. Also, the square root of the estimates of the variances are compared with Monte Carlo and bootstrap standard deviations.

Note that, the bias-corrected estimators of TCF_1, TCF_2 and TCF_3 require a parametric regression model to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$, or $\pi_i = \Pr(V_i = 1|T_i, A_i)$, or both. A wrong specification of such models may affect the estimation. Therefore, in the simulation study we consider four scenarios:

(i) the disease model and the verification model are both correctly specified;

(ii) the verification model is misspecified;

(iii) the disease model is misspecified;

(iv) the disease model and the verification model are both misspecified.

All scenarios allow to evaluate the behavior of the proposed estimators in finite samples. In particular, we consider 5000 Monte Carlo replications, and three sample sizes, i.e., 250, 500 and 1000 in scenario (i) and a sample size equal to 1000 in scenarios (ii)–(iv). The choice of such sample size in scenarios (ii)–(iv) allows to dig up expected bad behaviors of the estimators under misspecification, when a great amount of information is available, i.e., in large samples.

Study 1

The true disease $\mathcal{D}$ is generated by a trinomial random vector (D_1, D_2, D_3), such that D_k is a Bernoulli random variable with mean θ_k, $k = 1, 2, 3$. We set $\theta_1 = 0.4, \theta_2 = 0.35$ and $\theta_3 = 0.25$. The continuous test results T and A are generated from the following conditional models

$$T, A|D_k \sim \mathcal{N}_2\left(\mu_k, \Lambda\right), \qquad k = 1, 2, 3,$$

where $\mu_k = (2k, k)^{\top}$. We consider three different values for Λ, specifically

$$\begin{pmatrix} 1.75 & 0.1 \\ 0.1 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 5.5 & 3 \\ 3 & 2.5 \end{pmatrix},$$

giving rise to a correlation between T and A equal to $0.36, 0.69$ and 0.84, respectively.

In this scenario -and also in the next one- we consider six pairs for cut points (c_1, c_2), i.e., $(2,4), (2,5), (2,7), (4,5), (4,7)$ and $(5,7)$. Since the conditional distribution of T given D_k is the normal distribution, the true values of TCF's are obtained as

$$\begin{aligned} \mathrm{TCF}_1(c_1) &= \Phi\left(\frac{c_1-2}{\sigma_{T|D}}\right), \\ \mathrm{TCF}_2(c_1,c_2) &= \Phi\left(\frac{c_2-4}{\sigma_{T|D}}\right) - \Phi\left(\frac{c_1-4}{\sigma_{T|D_2}}\right), \\ \mathrm{TCF}_3(c_2) &= 1-\Phi\left(\frac{c_2-6}{\sigma_{T|D}}\right), \end{aligned}$$

where $\sigma_{T|D}$ denotes the entry in the 1-st row and 1-st column of Λ and $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal random variable, respectively.

Under our data–generating process, the true conditional disease model is a multinomial logistic model

$$\Pr(D_k = 1|T,A) = \frac{\exp\left(\tau_{\rho_{1k}} + \tau_{\rho_{2k}} T + \tau_{\rho_{3k}} A\right)}{1 + \exp\left(\tau_{\rho_{11}} + \tau_{\rho_{21}} T + \tau_{\rho_{31}} A\right) + \exp\left(\tau_{\rho_{12}} + \tau_{\rho_{22}} T + \tau_{\rho_{32}} A\right)},$$

for suitable $\tau_{\rho_{1k}}, \tau_{\rho_{2k}}, \tau_{\rho_{3k}}$, where $k = 1, 2$. The verification status V is generated by the following model

$$\mathrm{logit}\left\{\Pr(V = 1|T,A)\right\} = 0.5 - 0.3T + 0.75A.$$

This choice corresponds to a verification rate of about 0.65. In this study, the FI, MSI, IPW and SPE estimators are computed under correct working models for both the disease and the verification status. Therefore, in particular, the conditional verification probabilities π_i are estimated from a logistic model for V given T and A.

Tables 3.1–3.9 show Monte Carlo means, Monte Carlo standard deviations (MC.sd), the square roots of the variance estimated via asymptotic results (asy.sd) and bootstrap standard deviations (boot.sd) of $\widehat{\mathrm{TCF}}_1$, $\widehat{\mathrm{TCF}}_2$ and $\widehat{\mathrm{TCF}}_3$. Here, and in the following, bootstrap estimates are obtained from 250 bootstrap replications. Overall, the estimators FI, MSI, IPW and SPE behave similarly in this scenario, with the IPW estimator showing a slightly larger standard deviation. Simulation results, in this and in the following scenarios, also show that, excluding the SPE approach, bootstrap estimates of standard deviations are generally more accurate than estimates obtained via asymptotic theory.

Study 2

In this study, the true disease status $\mathcal{D}$ and the test results T and A are generated in the same way as in the first scenario. The true conditional verification process π, instead, is chosen to be the following function of T and A

$$\pi(T,A) = 0.35 + 0.3\mathrm{I}\left(T > t^{(0.8)}\right) + 0.35\mathrm{I}\left(A > a^{(0.8)}\right),$$

where $t^{(0.8)}$ and $a^{(0.8)}$ correspond to the 80-th percentile of distribution of T and A, respectively. In this case, the verification probabilities are 1 for subjects with $T > t^{(0.8)}$ and $A > a^{(0.8)}$; 0.7 for subjects with $T \leq t^{(0.8)}$ and $A > a^{(0.8)}$; 0.65 for subjects with $T > t^{(0.8)}$ and $A \leq a^{(0.8)}$; 0.35 otherwise. In our setting, the verification rate is approximately 0.48.

The aim in this scenario is to evaluate the behavior of the estimators, in particular that of IPW and SPE, under misspecification of the verification process. Therefore, $\hat{\pi}_i$ is estimated from

Table 3.1: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the first value of Λ is considered. "True" denotes the true parameter value. Sample size = 250.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
							cut point = (2,4)					
True	0.5000	0.4347	0.9347									
FI	0.5008	0.4357	0.9342	0.0529	0.0472	0.0272	0.0486	0.0440	0.0512	0.0530	0.0488	0.0276
MSI	0.5007	0.4353	0.9341	0.0544	0.0536	0.0318	0.0500	0.0501	0.0538	0.0542	0.0542	0.0324
IPW	0.5017	0.4352	0.9341	0.0714	0.0721	0.0371	0.0687	0.0697	0.0398	0.0704	0.0711	0.0373
SPE	0.5008	0.4352	0.9343	0.0574	0.0648	0.0364	0.0562	0.0632	0.0340	0.0596	0.1497	0.0425
							cut point = (2,5)					
True	0.5000	0.7099	0.7752									
FI	0.5008	0.7122	0.7756	0.0529	0.0464	0.0537	0.0486	0.0454	0.0618	0.0530	0.0467	0.0533
MSI	0.5007	0.7112	0.7747	0.0544	0.0511	0.0568	0.0500	0.0503	0.0644	0.0542	0.0514	0.0563
IPW	0.5017	0.7123	0.7739	0.0714	0.0683	0.0666	0.0687	0.0658	0.0704	0.0704	0.0677	0.0655
SPE	0.5008	0.7116	0.7751	0.0574	0.0619	0.0630	0.0562	0.0597	0.0603	0.0596	0.1219	0.1033
							cut point = (2,7)					
True	0.5000	0.9230	0.2248									
FI	0.5008	0.9231	0.2229	0.0529	0.0236	0.0520	0.0486	0.0327	0.0437	0.0530	0.0243	0.0525
MSI	0.5007	0.9230	0.2230	0.0544	0.0285	0.0530	0.0500	0.0361	0.0447	0.0542	0.0287	0.0534
IPW	0.5017	0.9234	0.2216	0.0714	0.0376	0.0748	0.0687	0.0341	0.0706	0.0704	0.0368	0.0727
SPE	0.5008	0.9234	0.2236	0.0574	0.0361	0.0571	0.0562	0.0334	0.0559	0.0596	0.0474	0.4185
							cut point = (4,5)					
True	0.9347	0.2752	0.7752									
FI	0.9350	0.2765	0.7756	0.0244	0.0408	0.0537	0.0224	0.0350	0.0618	0.0247	0.0415	0.0533
MSI	0.9351	0.2759	0.7747	0.0270	0.0467	0.0568	0.0247	0.0411	0.0644	0.0271	0.0467	0.0563
IPW	0.9356	0.2770	0.7739	0.0413	0.0690	0.0666	0.0343	0.0645	0.0704	0.0395	0.0663	0.0655
SPE	0.9356	0.2764	0.7751	0.0378	0.0587	0.0630	0.0333	0.0560	0.0603	0.0471	0.1566	0.1033
							cut point = (4,7)					
True	0.9347	0.4883	0.2248									
FI	0.9350	0.4874	0.2229	0.0244	0.0523	0.0520	0.0224	0.0494	0.0437	0.0247	0.0537	0.0525
MSI	0.9351	0.4877	0.2230	0.0270	0.0559	0.0530	0.0247	0.0528	0.0447	0.0271	0.0567	0.0534
IPW	0.9356	0.4881	0.2216	0.0413	0.0741	0.0748	0.0343	0.0708	0.0706	0.0395	0.0723	0.0727
SPE	0.9356	0.4882	0.2236	0.0378	0.0661	0.0571	0.0333	0.0640	0.0559	0.0471	0.1745	0.4185
							cut point = (5,7)					
True	0.9883	0.2132	0.2248									
FI	0.9880	0.2109	0.2229	0.0075	0.0432	0.0520	0.0066	0.0387	0.0437	0.0081	0.0436	0.0525
MSI	0.9881	0.2118	0.2230	0.0098	0.0460	0.0530	0.0075	0.0423	0.0447	0.0101	0.0468	0.0534
IPW	0.9885	0.2111	0.2216	0.0203	0.0634	0.0748	0.0097	0.0601	0.0706	0.0185	0.0625	0.0727
SPE	0.9883	0.2117	0.2236	0.0191	0.0569	0.0571	0.0117	0.0542	0.0559	0.0180	0.1382	0.4185

Table 3.2: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the first value of Λ is considered. "True" denotes the true parameter value. Sample size = 500.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.4347	0.9347									
FI	0.5007	0.4357	0.9343	0.0373	0.0339	0.0190	0.0343	0.0309	0.0360	0.0372	0.0340	0.0193
MSI	0.5008	0.4357	0.9343	0.0382	0.0378	0.0225	0.0353	0.0354	0.0380	0.0381	0.0380	0.0227
IPW	0.5020	0.4357	0.9345	0.0506	0.0508	0.0260	0.0493	0.0502	0.0280	0.0499	0.0507	0.0262
SPE	0.5012	0.4357	0.9345	0.0401	0.0456	0.0256	0.0399	0.0450	0.0249	0.0409	0.0457	0.0259
						cut point = (2,5)						
True	0.5000	0.7099	0.7752									
FI	0.5007	0.7115	0.7747	0.0373	0.0329	0.0372	0.0343	0.0321	0.0436	0.0372	0.0329	0.0374
MSI	0.5008	0.7111	0.7743	0.0382	0.0361	0.0395	0.0353	0.0355	0.0455	0.0381	0.0362	0.0396
IPW	0.5020	0.7111	0.7743	0.0506	0.0496	0.0464	0.0493	0.0479	0.0500	0.0499	0.0487	0.0463
SPE	0.5012	0.7112	0.7744	0.0401	0.0434	0.0442	0.0399	0.0425	0.0433	0.0409	0.0436	0.0448
						cut point = (2,7)						
True	0.5000	0.9230	0.2248									
FI	0.5007	0.9228	0.2241	0.0373	0.0167	0.0377	0.0343	0.0229	0.0310	0.0372	0.0169	0.0370
MSI	0.5008	0.9230	0.2242	0.0382	0.0199	0.0382	0.0353	0.0253	0.0317	0.0381	0.0202	0.0376
IPW	0.5020	0.9232	0.2242	0.0506	0.0266	0.0534	0.0493	0.0251	0.0520	0.0499	0.0263	0.0525
SPE	0.5012	0.9235	0.2245	0.0401	0.0255	0.0416	0.0399	0.0244	0.0403	0.0409	0.0253	0.0545
						cut point = (4,5)						
True	0.9347	0.2752	0.7752									
FI	0.9349	0.2758	0.7747	0.0176	0.0285	0.0372	0.0161	0.0246	0.0436	0.0174	0.0289	0.0374
MSI	0.9348	0.2754	0.7743	0.0194	0.0326	0.0395	0.0179	0.0291	0.0455	0.0191	0.0328	0.0396
IPW	0.9352	0.2754	0.7743	0.0299	0.0472	0.0464	0.0263	0.0466	0.0500	0.0284	0.0472	0.0463
SPE	0.9353	0.2755	0.7744	0.0270	0.0407	0.0442	0.0249	0.0399	0.0433	0.0291	0.0404	0.0448
						cut point = (4,7)						
True	0.9347	0.4883	0.2248									
FI	0.9349	0.4872	0.2241	0.0176	0.0375	0.0377	0.0161	0.0347	0.0310	0.0174	0.0375	0.0370
MSI	0.9348	0.4872	0.2242	0.0194	0.0396	0.0382	0.0179	0.0373	0.0317	0.0191	0.0398	0.0376
IPW	0.9352	0.4876	0.2242	0.0299	0.0520	0.0534	0.0263	0.0511	0.0520	0.0284	0.0516	0.0525
SPE	0.9353	0.4877	0.2245	0.0270	0.0462	0.0416	0.0249	0.0456	0.0403	0.0291	0.0463	0.0545
						cut point = (5,7)						
True	0.9883	0.2132	0.2248									
FI	0.9882	0.2114	0.2241	0.0051	0.0310	0.0377	0.0047	0.0274	0.0310	0.0054	0.0306	0.0370
MSI	0.9882	0.2118	0.2242	0.0069	0.0330	0.0382	0.0058	0.0299	0.0317	0.0069	0.0329	0.0376
IPW	0.9886	0.2121	0.2242	0.0137	0.0467	0.0534	0.0088	0.0441	0.0520	0.0130	0.0452	0.0525
SPE	0.9886	0.2123	0.2245	0.0133	0.0398	0.0416	0.0097	0.0387	0.0403	0.0127	0.0398	0.0545

Table 3.3: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the first value of Λ is considered. "True" denotes the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
							cut point = (2,4)					
True	0.5000	0.4347	0.9347									
FI	0.5001	0.4346	0.9348	0.0265	0.0235	0.0133	0.0242	0.0217	0.0254	0.0262	0.0238	0.0135
MSI	0.5002	0.4349	0.9349	0.0273	0.0264	0.0157	0.0250	0.0250	0.0268	0.0269	0.0268	0.0160
IPW	0.5006	0.4357	0.9349	0.0362	0.0357	0.0184	0.0352	0.0358	0.0198	0.0353	0.0360	0.0185
SPE	0.5004	0.4353	0.9349	0.0287	0.0321	0.0180	0.0282	0.0320	0.0180	0.0283	0.0322	0.0182
							cut point = (2,5)					
True	0.5000	0.7099	0.7752									
FI	0.5001	0.7096	0.7758	0.0265	0.0232	0.0260	0.0242	0.0227	0.0308	0.0262	0.0232	0.0263
MSI	0.5002	0.7095	0.7756	0.0273	0.0256	0.0276	0.0250	0.0251	0.0321	0.0269	0.0256	0.0279
IPW	0.5006	0.7104	0.7756	0.0362	0.0349	0.0325	0.0352	0.0342	0.0354	0.0353	0.0345	0.0327
SPE	0.5004	0.7100	0.7757	0.0287	0.0309	0.0307	0.0282	0.0303	0.0308	0.0283	0.0305	0.0310
							cut point = (2,7)					
True	0.5000	0.9230	0.2248									
FI	0.5001	0.9228	0.2250	0.0265	0.0117	0.0260	0.0242	0.0160	0.0220	0.0262	0.0119	0.0262
MSI	0.5002	0.9230	0.2252	0.0273	0.0141	0.0265	0.0250	0.0178	0.0226	0.0269	0.0142	0.0266
IPW	0.5006	0.9233	0.2258	0.0362	0.0187	0.0383	0.0352	0.0181	0.0374	0.0353	0.0186	0.0375
SPE	0.5004	0.9235	0.2256	0.0287	0.0180	0.0286	0.0282	0.0176	0.0286	0.0283	0.0180	0.0291
							cut point = (4,5)					
True	0.9347	0.2752	0.7752									
FI	0.9346	0.2749	0.7758	0.0124	0.0203	0.0260	0.0115	0.0173	0.0308	0.0123	0.0203	0.0263
MSI	0.9345	0.2746	0.7756	0.0137	0.0232	0.0276	0.0128	0.0205	0.0321	0.0136	0.0231	0.0279
IPW	0.9346	0.2748	0.7756	0.0213	0.0337	0.0325	0.0196	0.0332	0.0354	0.0205	0.0335	0.0327
SPE	0.9344	0.2747	0.7757	0.0190	0.0286	0.0307	0.0183	0.0283	0.0308	0.0187	0.0285	0.0310
							cut point = (4,7)					
True	0.9347	0.4883	0.2248									
FI	0.9346	0.4882	0.2250	0.0124	0.0262	0.0260	0.0115	0.0245	0.0220	0.0123	0.0264	0.0262
MSI	0.9345	0.4881	0.2252	0.0137	0.0279	0.0265	0.0128	0.0263	0.0226	0.0136	0.0280	0.0266
IPW	0.9346	0.4876	0.2258	0.0213	0.0365	0.0383	0.0196	0.0364	0.0374	0.0205	0.0366	0.0375
SPE	0.9344	0.4882	0.2256	0.0190	0.0325	0.0286	0.0183	0.0324	0.0286	0.0187	0.0326	0.0291
							cut point = (5,7)					
True	0.9883	0.2132	0.2248									
FI	0.9881	0.2132	0.2250	0.0036	0.0217	0.0260	0.0033	0.0194	0.0220	0.0037	0.0216	0.0262
MSI	0.9881	0.2135	0.2252	0.0048	0.0234	0.0265	0.0044	0.0212	0.0226	0.0049	0.0232	0.0266
IPW	0.9882	0.2129	0.2258	0.0100	0.0325	0.0383	0.0077	0.0317	0.0374	0.0097	0.0320	0.0375
SPE	0.9880	0.2135	0.2256	0.0097	0.0282	0.0286	0.0080	0.0276	0.0286	0.0094	0.0278	0.0291

Table 3.4: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the second value of Λ is considered. "True" denotes the true parameter value. Sample size = 250.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3970	0.8970									
FI	0.4995	0.3966	0.8972	0.0502	0.0419	0.0357	0.0461	0.0375	0.0502	0.0506	0.0429	0.0362
MSI	0.4996	0.3966	0.8970	0.0519	0.0498	0.0409	0.0479	0.0463	0.0536	0.0522	0.0506	0.0410
IPW	0.5001	0.3972	0.8979	0.0659	0.0700	0.0523	0.0646	0.0677	0.0504	0.0658	0.0687	0.0510
SPE	0.4996	0.3968	0.8980	0.0565	0.0623	0.0508	0.0559	0.0617	0.0469	0.0576	0.1619	0.0502
						cut point = (2,5)						
True	0.5000	0.6335	0.7365									
FI	0.4995	0.6340	0.7378	0.0502	0.0431	0.0580	0.0461	0.0410	0.0619	0.0506	0.0440	0.0580
MSI	0.4996	0.6335	0.7370	0.0519	0.0502	0.0617	0.0479	0.0485	0.0653	0.0522	0.0510	0.0616
IPW	0.5001	0.6330	0.7379	0.0659	0.0679	0.0733	0.0646	0.0660	0.0737	0.0658	0.0671	0.0721
SPE	0.4996	0.6335	0.7377	0.0565	0.0616	0.0686	0.0559	0.0610	0.0665	0.0576	0.1438	0.0686
						cut point = (2,7)						
True	0.5000	0.8682	0.2635									
FI	0.4995	0.8679	0.2640	0.0502	0.0307	0.0559	0.0461	0.0333	0.0499	0.0506	0.0314	0.0558
MSI	0.4996	0.8680	0.2644	0.0519	0.0362	0.0588	0.0479	0.0387	0.0523	0.0522	0.0372	0.0580
IPW	0.5001	0.8678	0.2659	0.0659	0.0492	0.0695	0.0646	0.0472	0.0682	0.0658	0.0492	0.0690
SPE	0.4996	0.8684	0.2649	0.0565	0.0467	0.0615	0.0559	0.0451	0.0591	0.0576	0.0593	0.0610
						cut point = (4,5)						
True	0.8970	0.2365	0.7365									
FI	0.8974	0.2374	0.7378	0.0284	0.0368	0.0580	0.0274	0.0318	0.0619	0.0288	0.0371	0.0580
MSI	0.8972	0.2369	0.7370	0.0320	0.0441	0.0617	0.0306	0.0395	0.0653	0.0320	0.0439	0.0616
IPW	0.8978	0.2358	0.7379	0.0377	0.0603	0.0733	0.0361	0.0574	0.0737	0.0372	0.0586	0.0721
SPE	0.8975	0.2368	0.7377	0.0364	0.0538	0.0686	0.0352	0.0519	0.0665	0.0363	0.2833	0.0686
						cut point = (4,7)						
True	0.8970	0.4711	0.2635									
FI	0.8974	0.4713	0.2640	0.0284	0.0504	0.0559	0.0274	0.0467	0.0499	0.0288	0.0510	0.0558
MSI	0.8972	0.4714	0.2644	0.0320	0.0554	0.0588	0.0306	0.0525	0.0523	0.0320	0.0562	0.0580
IPW	0.8978	0.4706	0.2659	0.0377	0.0693	0.0695	0.0361	0.0677	0.0682	0.0372	0.0687	0.0690
SPE	0.8975	0.4716	0.2649	0.0364	0.0635	0.0615	0.0352	0.0627	0.0591	0.0363	0.1949	0.0610
						cut point = (5,7)						
True	0.9711	0.2347	0.2635									
FI	0.9710	0.2339	0.2640	0.0121	0.0404	0.0559	0.0118	0.0369	0.0499	0.0127	0.0409	0.0558
MSI	0.9708	0.2345	0.2644	0.0165	0.0458	0.0588	0.0151	0.0431	0.0523	0.0167	0.0465	0.0580
IPW	0.9710	0.2348	0.2659	0.0203	0.0569	0.0695	0.0178	0.0556	0.0682	0.0201	0.0569	0.0690
SPE	0.9710	0.2348	0.2649	0.0201	0.0526	0.0615	0.0179	0.0521	0.0591	0.0199	0.1084	0.0610

Table 3.5: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the second value of Λ is considered. "True" denotes the true parameter value. Sample size = 500.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
							cut point = (2,4)					
True	0.5000	0.3970	0.8970									
FI	0.4999	0.3974	0.8965	0.0356	0.0294	0.0253	0.0326	0.0263	0.0355	0.0355	0.0298	0.0256
MSI	0.4999	0.3975	0.8961	0.0368	0.0355	0.0291	0.0339	0.0326	0.0380	0.0367	0.0355	0.0291
IPW	0.5000	0.3977	0.8962	0.0470	0.0492	0.0373	0.0460	0.0484	0.0369	0.0464	0.0487	0.0368
SPE	0.5000	0.3976	0.8963	0.0402	0.0446	0.0363	0.0397	0.0438	0.0348	0.0400	0.0442	0.0356
							cut point = (2,5)					
True	0.5000	0.6335	0.7365									
FI	0.4999	0.6342	0.7360	0.0356	0.0303	0.0410	0.0326	0.0287	0.0439	0.0355	0.0306	0.0409
MSI	0.4999	0.6339	0.7358	0.0368	0.0356	0.0437	0.0339	0.0342	0.0463	0.0367	0.0357	0.0435
IPW	0.5000	0.6336	0.7363	0.0470	0.0477	0.0528	0.0460	0.0471	0.0529	0.0464	0.0474	0.0514
SPE	0.5000	0.6341	0.7362	0.0402	0.0440	0.0494	0.0397	0.0434	0.0479	0.0400	0.0437	0.0483
							cut point = (2,7)					
True	0.5000	0.8682	0.2635									
FI	0.4999	0.8677	0.2631	0.0356	0.0222	0.0388	0.0326	0.0233	0.0352	0.0355	0.0219	0.0391
MSI	0.4999	0.8678	0.2633	0.0368	0.0263	0.0401	0.0339	0.0272	0.0370	0.0367	0.0261	0.0407
IPW	0.5000	0.8677	0.2638	0.0470	0.0354	0.0477	0.0460	0.0341	0.0486	0.0464	0.0349	0.0484
SPE	0.5000	0.8679	0.2635	0.0402	0.0336	0.0420	0.0397	0.0326	0.0420	0.0400	0.0331	0.0424
							cut point = (4,5)					
True	0.8970	0.2365	0.7365									
FI	0.8972	0.2368	0.7360	0.0205	0.0257	0.0410	0.0195	0.0223	0.0439	0.0203	0.0258	0.0409
MSI	0.8968	0.2364	0.7358	0.0229	0.0310	0.0437	0.0219	0.0279	0.0463	0.0226	0.0308	0.0435
IPW	0.8969	0.2359	0.7363	0.0268	0.0421	0.0528	0.0261	0.0411	0.0529	0.0265	0.0415	0.0514
SPE	0.8967	0.2365	0.7362	0.0260	0.0374	0.0494	0.0254	0.0370	0.0479	0.0257	0.0373	0.0483
							cut point = (4,7)					
True	0.8970	0.4711	0.2635									
FI	0.8972	0.4703	0.2631	0.0205	0.0356	0.0388	0.0195	0.0328	0.0352	0.0203	0.0356	0.0391
MSI	0.8968	0.4703	0.2633	0.0229	0.0398	0.0401	0.0219	0.0370	0.0370	0.0226	0.0394	0.0407
IPW	0.8969	0.4699	0.2638	0.0268	0.0492	0.0477	0.0261	0.0483	0.0486	0.0265	0.0486	0.0484
SPE	0.8967	0.4703	0.2635	0.0260	0.0454	0.0420	0.0254	0.0445	0.0420	0.0257	0.0449	0.0424
							cut point = (5,7)					
True	0.9711	0.2347	0.2635									
FI	0.9710	0.2335	0.2631	0.0086	0.0283	0.0388	0.0084	0.0260	0.0352	0.0088	0.0284	0.0391
MSI	0.9711	0.2339	0.2633	0.0116	0.0327	0.0401	0.0111	0.0304	0.0370	0.0117	0.0325	0.0407
IPW	0.9711	0.2341	0.2638	0.0144	0.0402	0.0477	0.0136	0.0397	0.0486	0.0143	0.0400	0.0484
SPE	0.9711	0.2339	0.2635	0.0142	0.0376	0.0420	0.0135	0.0370	0.0420	0.0141	0.0373	0.0424

Table 3.6: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the second value of Λ is considered. "True" denotes the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3970	0.8970									
FI	0.4997	0.3967	0.8966	0.0248	0.0208	0.0177	0.0230	0.0185	0.0250	0.0250	0.0209	0.0180
MSI	0.4997	0.3965	0.8966	0.0257	0.0251	0.0202	0.0240	0.0230	0.0268	0.0259	0.0250	0.0205
IPW	0.4994	0.3967	0.8967	0.0323	0.0349	0.0259	0.0327	0.0343	0.0263	0.0327	0.0344	0.0260
SPE	0.4997	0.3966	0.8967	0.0279	0.0317	0.0251	0.0281	0.0311	0.0250	0.0282	0.0311	0.0252
						cut point = (2,5)						
True	0.5000	0.6335	0.7365									
FI	0.4997	0.6330	0.7364	0.0248	0.0215	0.0286	0.0230	0.0203	0.0310	0.0250	0.0216	0.0288
MSI	0.4997	0.6327	0.7361	0.0257	0.0253	0.0304	0.0240	0.0241	0.0327	0.0259	0.0252	0.0307
IPW	0.4994	0.6326	0.7365	0.0323	0.0339	0.0360	0.0327	0.0335	0.0375	0.0327	0.0335	0.0363
SPE	0.4997	0.6328	0.7362	0.0279	0.0314	0.0338	0.0281	0.0308	0.0340	0.0282	0.0309	0.0341
						cut point = (2,7)						
True	0.5000	0.8682	0.2635									
FI	0.4997	0.8679	0.2640	0.0248	0.0153	0.0274	0.0230	0.0164	0.0249	0.0250	0.0154	0.0275
MSI	0.4997	0.8680	0.2643	0.0257	0.0183	0.0286	0.0240	0.0192	0.0262	0.0259	0.0184	0.0287
IPW	0.4994	0.8682	0.2645	0.0323	0.0248	0.0343	0.0327	0.0244	0.0345	0.0327	0.0246	0.0341
SPE	0.4997	0.8682	0.2644	0.0279	0.0236	0.0299	0.0281	0.0232	0.0297	0.0282	0.0234	0.0298
						cut point = (4,5)						
True	0.8970	0.2365	0.7365									
FI	0.8971	0.2363	0.7364	0.0144	0.0180	0.0286	0.0138	0.0157	0.0310	0.0143	0.0182	0.0288
MSI	0.8971	0.2362	0.7361	0.0160	0.0217	0.0304	0.0155	0.0197	0.0327	0.0160	0.0217	0.0307
IPW	0.8972	0.2359	0.7365	0.0188	0.0297	0.0360	0.0186	0.0291	0.0375	0.0187	0.0293	0.0363
SPE	0.8972	0.2362	0.7362	0.0183	0.0264	0.0338	0.0181	0.0262	0.0340	0.0182	0.0262	0.0341
						cut point = (4,7)						
True	0.8970	0.4711	0.2635									
FI	0.8971	0.4712	0.2640	0.0144	0.0252	0.0274	0.0138	0.0232	0.0249	0.0143	0.0250	0.0275
MSI	0.8971	0.4715	0.2643	0.0160	0.0280	0.0286	0.0155	0.0261	0.0262	0.0160	0.0278	0.0287
IPW	0.8972	0.4715	0.2645	0.0188	0.0348	0.0343	0.0186	0.0342	0.0345	0.0187	0.0343	0.0341
SPE	0.8972	0.4717	0.2644	0.0183	0.0321	0.0299	0.0181	0.0316	0.0297	0.0182	0.0316	0.0298
						cut point = (5,7)						
True	0.9711	0.2347	0.2635									
FI	0.9709	0.2350	0.2640	0.0061	0.0201	0.0274	0.0060	0.0184	0.0249	0.0062	0.0200	0.0275
MSI	0.9709	0.2353	0.2643	0.0082	0.0229	0.0286	0.0080	0.0216	0.0262	0.0082	0.0229	0.0287
IPW	0.9709	0.2356	0.2645	0.0101	0.0285	0.0343	0.0099	0.0282	0.0345	0.0102	0.0283	0.0341
SPE	0.9710	0.2354	0.2644	0.0100	0.0266	0.0299	0.0098	0.0263	0.0297	0.0100	0.0264	0.0298

Table 3.7: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the third value of Λ is considered. "True" denotes the true parameter value. Sample size = 250.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3031	0.8031									
FI	0.4997	0.3037	0.8055	0.0498	0.0338	0.0498	0.0453	0.0294	0.0529	0.0498	0.0348	0.0489
MSI	0.4999	0.3035	0.8046	0.0519	0.0453	0.0554	0.0480	0.0412	0.0578	0.0522	0.0453	0.0542
IPW	0.5003	0.3033	0.8042	0.0617	0.0632	0.0655	0.0616	0.0617	0.0633	0.0624	0.0627	0.0639
SPE	0.5001	0.3034	0.8044	0.0564	0.0588	0.0636	0.0564	0.0573	0.0615	0.0570	0.0579	0.0624
						cut point = (2,5)						
True	0.5000	0.4682	0.6651									
FI	0.4997	0.4697	0.6684	0.0498	0.0381	0.0617	0.0453	0.0339	0.0608	0.0498	0.0390	0.0608
MSI	0.4999	0.4691	0.6675	0.0519	0.0503	0.0670	0.0480	0.0460	0.0654	0.0522	0.0499	0.0653
IPW	0.5003	0.4687	0.6675	0.0617	0.0688	0.0763	0.0616	0.0668	0.0741	0.0624	0.0676	0.0742
SPE	0.5001	0.4690	0.6674	0.0564	0.0641	0.0735	0.0564	0.0621	0.0706	0.0570	0.0627	0.0715
						cut point = (2,7)						
True	0.5000	0.7027	0.3349									
FI	0.4997	0.7037	0.3370	0.0498	0.0378	0.0591	0.0453	0.0353	0.0545	0.0498	0.0384	0.0592
MSI	0.4999	0.7037	0.3367	0.0519	0.0482	0.0626	0.0480	0.0451	0.0588	0.0522	0.0476	0.0632
IPW	0.5003	0.7033	0.3371	0.0617	0.0642	0.0709	0.0616	0.0614	0.0715	0.0624	0.0624	0.0721
SPE	0.5001	0.7038	0.3367	0.0564	0.0603	0.0660	0.0564	0.0581	0.0661	0.0570	0.0587	0.0670
						cut point = (4,5)						
True	0.8031	0.1651	0.6651									
FI	0.8037	0.1660	0.6684	0.0393	0.0277	0.0617	0.0366	0.0236	0.0608	0.0388	0.0282	0.0608
MSI	0.8033	0.1656	0.6675	0.0425	0.0369	0.0670	0.0400	0.0333	0.0654	0.0420	0.0369	0.0653
IPW	0.8033	0.1654	0.6675	0.0486	0.0497	0.0763	0.0469	0.0486	0.0741	0.0475	0.0496	0.0742
SPE	0.8032	0.1656	0.6674	0.0469	0.0460	0.0735	0.0457	0.0454	0.0706	0.0460	0.0458	0.0715
						cut point = (4,7)						
True	0.8031	0.3996	0.3349									
FI	0.8037	0.4000	0.3370	0.0393	0.0419	0.0591	0.0366	0.0383	0.0545	0.0388	0.0430	0.0592
MSI	0.8033	0.4002	0.3367	0.0425	0.0513	0.0626	0.0400	0.0480	0.0588	0.0420	0.0519	0.0632
IPW	0.8033	0.4000	0.3371	0.0486	0.0639	0.0709	0.0469	0.0643	0.0715	0.0475	0.0652	0.0721
SPE	0.8032	0.4004	0.3367	0.0469	0.0604	0.0660	0.0457	0.0605	0.0661	0.0460	0.0612	0.0670
						cut point = (5,7)						
True	0.8996	0.2345	0.3349									
FI	0.9000	0.2340	0.3370	0.0271	0.0348	0.0591	0.0255	0.0313	0.0545	0.0269	0.0356	0.0592
MSI	0.8998	0.2346	0.3367	0.0312	0.0441	0.0626	0.0296	0.0407	0.0588	0.0310	0.0441	0.0632
IPW	0.8998	0.2347	0.3371	0.0359	0.0553	0.0709	0.0347	0.0545	0.0715	0.0355	0.0555	0.0721
SPE	0.8998	0.2348	0.3367	0.0351	0.0521	0.0660	0.0347	0.0516	0.0661	0.0348	0.0521	0.0670

Table 3.8: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the third value of Λ is considered. "True" denotes the true parameter value. Sample size = 500.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3031	0.8031									
FI	0.5001	0.3027	0.8034	0.0356	0.0240	0.0348	0.0320	0.0206	0.0373	0.0350	0.0242	0.0346
MSI	0.5001	0.3031	0.8034	0.0375	0.0322	0.0384	0.0340	0.0291	0.0408	0.0367	0.0318	0.0383
IPW	0.5004	0.3031	0.8037	0.0454	0.0444	0.0454	0.0438	0.0439	0.0451	0.0440	0.0441	0.0452
SPE	0.5002	0.3032	0.8036	0.0410	0.0411	0.0441	0.0400	0.0406	0.0438	0.0401	0.0407	0.0440
						cut point = (2,5)						
True	0.5000	0.4682	0.6651									
FI	0.5001	0.4681	0.6656	0.0356	0.0266	0.0438	0.0320	0.0237	0.0430	0.0350	0.0271	0.0428
MSI	0.5001	0.4679	0.6654	0.0375	0.0348	0.0469	0.0340	0.0325	0.0461	0.0367	0.0350	0.0459
IPW	0.5004	0.4676	0.6655	0.0454	0.0475	0.0538	0.0438	0.0474	0.0526	0.0440	0.0476	0.0524
SPE	0.5002	0.4678	0.6654	0.0410	0.0440	0.0513	0.0400	0.0440	0.0500	0.0401	0.0442	0.0503
						cut point = (2,7)						
True	0.5000	0.7027	0.3349									
FI	0.5001	0.7033	0.3346	0.0356	0.0268	0.0424	0.0320	0.0246	0.0383	0.0350	0.0267	0.0412
MSI	0.5001	0.7033	0.3346	0.0375	0.0336	0.0455	0.0340	0.0318	0.0414	0.0367	0.0334	0.0441
IPW	0.5004	0.7031	0.3352	0.0454	0.0439	0.0515	0.0438	0.0437	0.0505	0.0440	0.0440	0.0504
SPE	0.5002	0.7034	0.3347	0.0410	0.0416	0.0481	0.0400	0.0413	0.0465	0.0401	0.0414	0.0468
						cut point = (4,5)						
True	0.8031	0.1651	0.6651									
FI	0.8033	0.1654	0.6656	0.0278	0.0196	0.0438	0.0260	0.0166	0.0430	0.0274	0.0196	0.0428
MSI	0.8030	0.1648	0.6654	0.0303	0.0256	0.0469	0.0284	0.0236	0.0461	0.0297	0.0259	0.0459
IPW	0.8030	0.1645	0.6655	0.0344	0.0346	0.0538	0.0335	0.0346	0.0526	0.0337	0.0349	0.0524
SPE	0.8030	0.1645	0.6654	0.0334	0.0317	0.0513	0.0325	0.0321	0.0500	0.0326	0.0322	0.0503
						cut point = (4,7)						
True	0.8031	0.3996	0.3349									
FI	0.8033	0.4007	0.3346	0.0278	0.0300	0.0424	0.0260	0.0268	0.0383	0.0274	0.0299	0.0412
MSI	0.8030	0.4002	0.3346	0.0303	0.0367	0.0455	0.0284	0.0339	0.0414	0.0297	0.0364	0.0441
IPW	0.8030	0.4000	0.3352	0.0344	0.0458	0.0515	0.0335	0.0456	0.0505	0.0337	0.0458	0.0504
SPE	0.8030	0.4002	0.3347	0.0334	0.0431	0.0481	0.0325	0.0429	0.0465	0.0326	0.0430	0.0468
						cut point = (5,7)						
True	0.8996	0.2345	0.3349									
FI	0.8996	0.2353	0.3346	0.0192	0.0245	0.0424	0.0182	0.0220	0.0383	0.0190	0.0248	0.0412
MSI	0.8996	0.2354	0.3346	0.0221	0.0307	0.0455	0.0212	0.0288	0.0414	0.0219	0.0310	0.0441
IPW	0.8997	0.2355	0.3352	0.0253	0.0384	0.0515	0.0249	0.0388	0.0505	0.0252	0.0391	0.0504
SPE	0.8997	0.2356	0.3347	0.0249	0.0364	0.0481	0.0246	0.0366	0.0465	0.0247	0.0367	0.0468

Table 3.9: Simulation results from 5000 replications when both models for ρ_k and π are correctly specified (Study 1) and the third value of Λ is considered. "True" denotes the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3031	0.8031									
FI	0.5003	0.3030	0.8040	0.0242	0.0169	0.0243	0.0226	0.0145	0.0264	0.0247	0.0170	0.0243
MSI	0.5001	0.3030	0.8038	0.0256	0.0222	0.0270	0.0240	0.0206	0.0288	0.0259	0.0224	0.0270
IPW	0.5001	0.3032	0.8038	0.0310	0.0310	0.0320	0.0310	0.0311	0.0320	0.0311	0.0311	0.0319
SPE	0.5001	0.3030	0.8040	0.0281	0.0285	0.0312	0.0283	0.0288	0.0310	0.0283	0.0287	0.0311
						cut point = (2,5)						
True	0.5000	0.4682	0.6651									
FI	0.5003	0.4682	0.6663	0.0242	0.0193	0.0301	0.0226	0.0167	0.0304	0.0247	0.0191	0.0301
MSI	0.5001	0.4681	0.6663	0.0256	0.0248	0.0320	0.0240	0.0230	0.0326	0.0259	0.0247	0.0324
IPW	0.5001	0.4683	0.6664	0.0310	0.0337	0.0368	0.0310	0.0336	0.0373	0.0311	0.0336	0.0370
SPE	0.5001	0.4682	0.6665	0.0281	0.0311	0.0350	0.0283	0.0312	0.0355	0.0283	0.0312	0.0355
						cut point = (2,7)						
True	0.5000	0.7027	0.3349									
FI	0.5003	0.7028	0.3359	0.0242	0.0188	0.0289	0.0226	0.0173	0.0271	0.0247	0.0188	0.0290
MSI	0.5001	0.7025	0.3359	0.0256	0.0236	0.0307	0.0240	0.0225	0.0293	0.0259	0.0236	0.0311
IPW	0.5001	0.7023	0.3360	0.0310	0.0311	0.0350	0.0310	0.0310	0.0358	0.0311	0.0311	0.0356
SPE	0.5001	0.7024	0.3358	0.0281	0.0292	0.0324	0.0283	0.0293	0.0329	0.0283	0.0293	0.0330
						cut point = (4,5)						
True	0.8031	0.1651	0.6651									
FI	0.8034	0.1652	0.6663	0.0193	0.0139	0.0301	0.0184	0.0117	0.0304	0.0193	0.0138	0.0301
MSI	0.8032	0.1652	0.6663	0.0211	0.0184	0.0320	0.0201	0.0167	0.0326	0.0209	0.0183	0.0324
IPW	0.8034	0.1651	0.6664	0.0241	0.0248	0.0368	0.0237	0.0246	0.0373	0.0237	0.0247	0.0370
SPE	0.8032	0.1653	0.6665	0.0233	0.0229	0.0350	0.0230	0.0228	0.0355	0.0230	0.0228	0.0355
						cut point = (4,7)						
True	0.8031	0.3996	0.3349									
FI	0.8034	0.3998	0.3359	0.0193	0.0207	0.0289	0.0184	0.0189	0.0271	0.0193	0.0210	0.0290
MSI	0.8032	0.3995	0.3359	0.0211	0.0253	0.0307	0.0201	0.0240	0.0293	0.0209	0.0256	0.0311
IPW	0.8034	0.3991	0.3360	0.0241	0.0319	0.0350	0.0237	0.0323	0.0358	0.0237	0.0323	0.0356
SPE	0.8032	0.3994	0.3358	0.0233	0.0299	0.0324	0.0230	0.0303	0.0329	0.0230	0.0304	0.0330
						cut point = (5,7)						
True	0.8996	0.2345	0.3349									
FI	0.8998	0.2346	0.3359	0.0134	0.0172	0.0289	0.0129	0.0155	0.0271	0.0134	0.0174	0.0290
MSI	0.8997	0.2343	0.3359	0.0157	0.0216	0.0307	0.0150	0.0204	0.0293	0.0155	0.0218	0.0311
IPW	0.8998	0.2340	0.3360	0.0180	0.0273	0.0350	0.0177	0.0274	0.0358	0.0177	0.0275	0.0356
SPE	0.8997	0.2342	0.3358	0.0178	0.0256	0.0324	0.0174	0.0258	0.0329	0.0174	0.0259	0.0330

a logistic regression model with V as the response and T as predictor, while $\hat{\rho}_{ki}$ is still obtained from the multinomial logistic model (similarly to the first scenario). Clearly, the model used for verification status is misspecified.

Table 3.11–3.12 show Monte Carlo means and standard deviations for the estimators of the true class fractions TCF_1, TCF_2 and TCF_3. Moreover, estimated standard deviations (via asymptotic theory) and bootstrap standard deviations are also presented. The results clearly show the effect of misspecification on IPW estimates, despite the high sample size. In particular, in terms of bias, the IPW method performs almost always poorly, with high distortion in some cases (values highlighted in bold). On the other hand, the SPE estimator behaves well, due to its doubly robustness property.

Study 3

Starting from two independent random variables $Z_1 \sim \mathcal{N}(0, 0.5)$ and $Z_2 \sim \mathcal{N}(0, 0.5)$, the true conditional disease $\mathcal{D}$ is generated by a trinomial random vector (D_1, D_2, D_3) such that

$$D_1 = \begin{cases} 1 & \text{if } Z_1 + Z_2 \leq h_1 \\ 0 & \text{otherwise} \end{cases}, \; D_2 = \begin{cases} 1 & \text{if } h_1 < Z_1 + Z_2 \leq h_2 \\ 0 & \text{otherwise} \end{cases},$$

$$D_3 = \begin{cases} 1 & \text{if } Z_1 + Z_2 > h_2 \\ 0 & \text{otherwise} \end{cases}.$$

Here, h_1 and h_2 are two thresholds. We choose h_1 and h_2 to make $\theta_1 = 0.4$ and $\theta_3 = 0.25$. The continuous test result T and the covariate A are generated to be related to $\mathcal{D}$ through Z_1 and Z_2. More precisely,

$$T = 0.5(Z_1 + Z_2) + \varepsilon_1, \qquad A = Z_1 + Z_2 + \varepsilon_2,$$

where ε_1 and ε_2 are two independent normal random variables with mean 0 and the common variance 0.25, independent also from Z_1 and Z_2. The verification status V is simulated by the following logistic model

$$\text{logit}\left\{\Pr(V = 1|T, A)\right\} = 0.1 - 1.53T + A.$$

Under this model, the verification rate is roughly 0.52. We consider the cut points as the pairs $(-1, -0.5), (-1, 0.7), (-1, 1.3), (-0.5, 0.7), (-0.5, 1.3)$ and $(0.7, 1.3)$. In this set–up, we determine the true values of TCF's as

$$\begin{aligned}
\text{TCF}_1(c_1) &= \frac{1}{\Phi(h_1)} \int_{-\infty}^{h_1} \Phi\left(\frac{c_1 - 0.5z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z, \\
\text{TCF}_2(c_1, c_2) &= \frac{1}{\Phi(h_2) - \Phi(h_1)} \int_{h_1}^{h_2} \left[\Phi\left(\frac{c_2 - 0.5z}{\sqrt{0.25}}\right) - \Phi\left(\frac{c_1 - 0.5z}{\sqrt{0.25}}\right)\right] \phi(z)\mathrm{d}z, \\
\text{TCF}_3(c_2) &= 1 - \frac{1}{1 - \Phi(h_2)} \int_{h_2}^{\infty} \Phi\left(\frac{c_2 - 0.5z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z.
\end{aligned}$$

The aim in this scenario is to evaluate the behavior of the estimators, in particular that of FI, MSI and SPE, when the estimators $\hat{\rho}_{ki}$ are inconsistent, whereas $\hat{\pi}_i$ are consistent. Therefore, $\hat{\rho}_{ki}$ are obtained from a multinomial logistic regression model with (D_1, D_2, D_3) as the response and T as predictor. As the correct process is a multinomial probit process, the chosen model is clearly misspecified. To estimate the conditional verification process π, we use a generalized linear model for V given T and A with logit link. Clearly, this model is correctly specified.

Table 3.13 shows Monte Carlo means and standard deviations for the estimators of the true class fractions TCF_1, TCF_2 and TCF_3. Moreover, estimated standard deviations (via asymptotic

Table 3.10: Simulation results from 5000 replications when the model for the verification process is misspecified (Study 2) and the first value of Λ is used. "True" indicates the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
							cut point = (2,4)					
True	0.5000	0.4347	0.9347									
FI	0.5011	0.4345	0.9351	0.0277	0.0238	0.0152	0.0239	0.0203	0.0262	0.0275	0.0241	0.0153
MSI	0.5011	0.4344	0.9351	0.0280	0.0259	0.0169	0.0243	0.0226	0.0271	0.0279	0.0260	0.0169
IPW	**0.5822**	**0.4436**	0.9375	0.0381	0.0407	0.0213	0.0380	0.0400	0.0255	0.0381	0.0401	0.0212
SPE	0.5011	0.4345	0.9352	0.0304	0.0334	0.0218	0.0304	0.0330	0.0214	0.0305	0.0331	0.0216
							cut point = (2,5)					
True	0.5000	0.7099	0.7752									
FI	0.5011	0.7105	0.7765	0.0277	0.0227	0.0297	0.0239	0.0223	0.0334	0.0275	0.0228	0.0298
MSI	0.5011	0.7101	0.7762	0.0280	0.0245	0.0305	0.0243	0.0241	0.0343	0.0279	0.0245	0.0309
IPW	**0.5822**	**0.6815**	**0.8046**	0.0381	0.0376	0.0325	0.0380	0.0370	0.0381	0.0381	0.0371	0.0327
SPE	0.5011	0.7099	0.7760	0.0304	0.0309	0.0328	0.0304	0.0306	0.0330	0.0305	0.0307	0.0331
							cut point = (2,7)					
True	0.5000	0.9230	0.2248									
FI	0.5011	0.9233	0.2256	0.0277	0.0144	0.0270	0.0239	0.0193	0.0250	0.0275	0.0143	0.0270
MSI	0.5011	0.9234	0.2258	0.0280	0.0161	0.0275	0.0243	0.0204	0.0256	0.0279	0.0158	0.0275
IPW	**0.5822**	**0.9009**	0.2306	0.0381	0.0276	0.0306	0.0380	0.0268	0.0316	0.0381	0.0270	0.0308
SPE	0.5011	0.9234	0.2258	0.0304	0.0225	0.0279	0.0304	0.0218	0.0280	0.0305	0.0220	0.0281
							cut point = (4,5)					
True	0.9347	0.2752	0.7752									
FI	0.9352	0.2760	0.7765	0.0135	0.0218	0.0297	0.0127	0.0168	0.0334	0.0135	0.0215	0.0298
MSI	0.9352	0.2757	0.7762	0.0143	0.0237	0.0305	0.0135	0.0191	0.0343	0.0143	0.0233	0.0309
IPW	**0.9540**	**0.2379**	**0.8046**	0.0139	0.0335	0.0325	0.0138	0.0330	0.0381	0.0139	0.0331	0.0327
SPE	0.9352	0.2754	0.7760	0.0161	0.0279	0.0328	0.0160	0.0275	0.0330	0.0161	0.0275	0.0331
							cut point = (4,7)					
True	0.9347	0.4883	0.2248									
FI	0.9352	0.4888	0.2256	0.0135	0.0290	0.0270	0.0127	0.0259	0.0250	0.0135	0.0287	0.0270
MSI	0.9352	0.4889	0.2258	0.0143	0.0302	0.0275	0.0135	0.0273	0.0256	0.0143	0.0300	0.0275
IPW	**0.9540**	**0.4574**	0.2306	0.0139	0.0391	0.0306	0.0138	0.0387	0.0316	0.0139	0.0388	0.0308
SPE	0.9352	0.4890	0.2258	0.0161	0.0328	0.0279	0.0160	0.0327	0.0280	0.0161	0.0328	0.0281
							cut point = (5,7)					
True	0.9883	0.2132	0.2248									
FI	0.9883	0.2128	0.2256	0.0040	0.0216	0.0270	0.0038	0.0190	0.0250	0.0040	0.0215	0.0270
MSI	0.9884	0.2133	0.2258	0.0050	0.0231	0.0275	0.0046	0.0208	0.0256	0.0050	0.0231	0.0275
IPW	**0.9912**	0.2195	0.2306	0.0060	0.0305	0.0306	0.0054	0.0301	0.0316	0.0059	0.0302	0.0308
SPE	0.9885	0.2135	0.2258	0.0065	0.0256	0.0279	0.0060	0.0256	0.0280	0.0064	0.0257	0.0281

Table 3.11: Simulation results from 5000 replications when the model for the verification process is misspecified (Study 2) and the second value of Λ is used. "True" indicates the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3970	0.8970									
FI	0.4998	0.3970	0.8977	0.0267	0.0211	0.0207	0.0231	0.0172	0.0268	0.0268	0.0213	0.0204
MSI	0.4997	0.3970	0.8978	0.0272	0.0240	0.0220	0.0237	0.0202	0.0279	0.0274	0.0239	0.0219
IPW	**0.5983**	**0.3743**	**0.9150**	0.0364	0.0407	0.0257	0.0368	0.0399	0.0284	0.0369	0.0399	0.0258
SPE	0.4996	0.3971	0.8980	0.0314	0.0342	0.0268	0.0314	0.0336	0.0267	0.0315	0.0336	0.0268
						cut point = (2,5)						
True	0.5000	0.6335	0.7365									
FI	0.4998	0.6340	0.7381	0.0267	0.0218	0.0332	0.0231	0.0198	0.0344	0.0268	0.0219	0.0326
MSI	0.4997	0.6338	0.7381	0.0272	0.0246	0.0344	0.0237	0.0226	0.0356	0.0274	0.0243	0.0338
IPW	**0.5983**	**0.5749**	**0.7965**	0.0364	0.0406	0.0348	0.0368	0.0401	0.0387	0.0369	0.0402	0.0346
SPE	0.4996	0.6338	0.7383	0.0314	0.0341	0.0368	0.0314	0.0335	0.0364	0.0315	0.0336	0.0364
						cut point = (2,7)						
True	0.5000	0.8682	0.2635									
FI	0.4998	0.8690	0.2639	0.0267	0.0175	0.0295	0.0231	0.0190	0.0280	0.0268	0.0176	0.0290
MSI	0.4997	0.8689	0.2639	0.0272	0.0197	0.0308	0.0237	0.0211	0.0292	0.0274	0.0198	0.0302
IPW	**0.5983**	**0.8307**	**0.3054**	0.0364	0.0342	0.0347	0.0368	0.0341	0.0358	0.0369	0.0343	0.0343
SPE	0.4996	0.8688	0.2639	0.0314	0.0283	0.0316	0.0314	0.0282	0.0308	0.0315	0.0284	0.0310
						cut point = (4,5)						
True	0.8970	0.2365	0.7365									
FI	0.8975	0.2370	0.7381	0.0159	0.0191	0.0332	0.0153	0.0147	0.0344	0.0162	0.0191	0.0326
MSI	0.8974	0.2368	0.7381	0.0168	0.0215	0.0344	0.0162	0.0175	0.0356	0.0171	0.0213	0.0338
IPW	**0.9216**	**0.2006**	**0.7965**	0.0165	0.0315	0.0348	0.0167	0.0308	0.0387	0.0168	0.0309	0.0346
SPE	0.8974	0.2367	0.7383	0.0189	0.0266	0.0368	0.0191	0.0261	0.0364	0.0191	0.0262	0.0364
						cut point = (4,7)						
True	0.8970	0.4711	0.2635									
FI	0.8975	0.4721	0.2639	0.0159	0.0276	0.0295	0.0153	0.0247	0.0280	0.0162	0.0276	0.0290
MSI	0.8974	0.4719	0.2639	0.0168	0.0300	0.0308	0.0162	0.0269	0.0292	0.0171	0.0296	0.0302
IPW	**0.9216**	**0.4564**	**0.3054**	0.0165	0.0395	0.0347	0.0167	0.0387	0.0358	0.0168	0.0388	0.0343
SPE	0.8974	0.4717	0.2639	0.0189	0.0339	0.0316	0.0191	0.0331	0.0308	0.0191	0.0333	0.0310
						cut point = (5,7)						
True	0.9711	0.2347	0.2635									
FI	0.9712	0.2351	0.2639	0.0069	0.0208	0.0295	0.0067	0.0185	0.0280	0.0070	0.0209	0.0290
MSI	0.9712	0.2351	0.2639	0.0083	0.0237	0.0308	0.0080	0.0214	0.0292	0.0084	0.0234	0.0302
IPW	**0.9752**	**0.2558**	**0.3054**	0.0092	0.0319	0.0347	0.0091	0.0315	0.0358	0.0092	0.0316	0.0343
SPE	0.9713	0.2350	0.2639	0.0101	0.0273	0.0316	0.0100	0.0266	0.0308	0.0101	0.0267	0.0310

Table 3.12: Simulation results from 5000 replications when the model for the verification process is misspecified (Study 2) and the third value of Λ is used. "True" indicates the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (2,4)						
True	0.5000	0.3031	0.8031									
FI	0.4998	0.3026	0.8043	0.0257	0.0172	0.0280	0.0221	0.0124	0.0293	0.0259	0.0171	0.0278
MSI	0.4999	0.3027	0.8044	0.0264	0.0204	0.0297	0.0230	0.0166	0.0308	0.0267	0.0204	0.0293
IPW	**0.6267**	**0.2614**	**0.8259**	0.0345	0.0371	0.0371	0.0346	0.0364	0.0372	0.0348	0.0366	0.0365
SPE	0.5000	0.3031	0.8047	0.0322	0.0323	0.0361	0.0324	0.0321	0.0352	0.0326	0.0322	0.0354
						cut point = (2,5)						
True	0.5000	0.4682	0.6651									
FI	0.4998	0.4681	0.6667	0.0257	0.0192	0.0341	0.0221	0.0151	0.0342	0.0259	0.0192	0.0343
MSI	0.4999	0.4681	0.6664	0.0264	0.0227	0.0354	0.0230	0.0195	0.0357	0.0267	0.0229	0.0358
IPW	**0.6267**	**0.3884**	**0.7253**	0.0345	0.0403	0.0396	0.0346	0.0400	0.0413	0.0348	0.0402	0.0401
SPE	0.5000	0.4684	0.6665	0.0322	0.0352	0.0389	0.0324	0.0353	0.0391	0.0326	0.0355	0.0393
						cut point = (2,7)						
True	0.5000	0.7027	0.3349									
FI	0.4998	0.7035	0.3360	0.0257	0.0201	0.0318	0.0221	0.0184	0.0311	0.0259	0.0203	0.0320
MSI	0.4999	0.7035	0.3360	0.0264	0.0237	0.0337	0.0230	0.0224	0.0331	0.0267	0.0240	0.0339
IPW	**0.6267**	**0.6157**	**0.4102**	0.0345	0.0417	0.0386	0.0346	0.0416	0.0398	0.0348	0.0417	0.0386
SPE	0.5000	0.7038	0.3360	0.0322	0.0360	0.0350	0.0324	0.0361	0.0350	0.0326	0.0364	0.0352
						cut point = (4,5)						
True	0.8031	0.1651	0.6651									
FI	0.8032	0.1655	0.6667	0.0207	0.0139	0.0341	0.0189	0.0099	0.0342	0.0207	0.0141	0.0343
MSI	0.8031	0.1654	0.6664	0.0217	0.0165	0.0354	0.0200	0.0135	0.0357	0.0216	0.0169	0.0358
IPW	**0.8512**	**0.1270**	**0.7253**	0.0217	0.0245	0.0396	0.0215	0.0251	0.0413	0.0215	0.0253	0.0401
SPE	0.8030	0.1653	0.6665	0.0239	0.0225	0.0389	0.0237	0.0228	0.0391	0.0238	0.0229	0.0393
						cut point = (4,7)						
True	0.8031	0.3996	0.3349									
FI	0.8032	0.4009	0.3360	0.0207	0.0226	0.0318	0.0189	0.0194	0.0311	0.0207	0.0227	0.0320
MSI	0.8031	0.4008	0.3360	0.0217	0.0261	0.0337	0.0200	0.0234	0.0331	0.0216	0.0262	0.0339
IPW	**0.8512**	**0.3544**	**0.4102**	0.0217	0.0358	0.0386	0.0215	0.0362	0.0398	0.0215	0.0363	0.0386
SPE	0.8030	0.4008	0.3360	0.0239	0.0326	0.0350	0.0237	0.0325	0.0350	0.0238	0.0327	0.0352
						cut point = (5,7)						
True	0.8996	0.2345	0.3349									
FI	0.8997	0.2354	0.3360	0.0144	0.0183	0.0318	0.0135	0.0156	0.0311	0.0144	0.0184	0.0320
MSI	0.8995	0.2354	0.3360	0.0158	0.0223	0.0337	0.0149	0.0197	0.0331	0.0157	0.0220	0.0339
IPW	**0.9149**	**0.2274**	**0.4102**	0.0163	0.0303	0.0386	0.0160	0.0299	0.0398	0.0161	0.0301	0.0386
SPE	0.8995	0.2355	0.3360	0.0175	0.0273	0.0350	0.0173	0.0268	0.0350	0.0174	0.0269	0.0352

theory) and bootstrap standard deviations are also presented. The results clearly show the effect of misspecification on FI and MSI estimates, despite the high sample size. In particular, in terms of bias, the two methods performs almost always poorly, with high distortion in some cases (values highlighted in bold). Again, the SPE estimator behaves well due to its doubly robustness property.

Study 4

We generate data exactly as in Study 3. The aim in this scenario is to evaluate the behavior of FI, MSI, IPW and SPE estimators when the estimates $\hat{\rho}_{ki}$ and $\hat{\pi}_i$ are inconsistent. Therefore, $\hat{\rho}_{ki}$ are obtained from a multinomial logistic regression model with (D_1, D_2, D_3) as the response and T as predictor. This model is misspecified. To estimate the conditional verification disease π, we use a generalized linear model for V given T and $A^{2/3}$ with logit link. Clearly, this model is misspecified.

Table 3.14 shows Monte Carlo means and standard deviations for the estimators of the true class fractions TCF_1, TCF_2 and TCF_3. Moreover, estimated standard deviations (via asymptotic theory) and bootstrap standard deviations are also presented. The results clearly show that when both the disease and verification models are misspecified, all estimators may behave poorly, with high distortion in some cases (values highlighted in bold).

3.3.2 Simulation studies for the KNN estimator

To assess the performance of the KNN approach in finite samples under MAR assumption, we conducted the following simulation experiments. The first study uses the same setting as in Study 1 of the parametric approaches, and aims to investigate the effect of the choice of K for the bias and variance of the estimates. In particular, we use $K = 1, 3, 5, 10, 20$ and the Euclidean distance. The asymptotic variances are estimated by using the procedure discussed in Section 3.2.3, with $\bar{K} = 2$. The results obtained with the three different values of Λ are respectively presented in Tables 3.15–3.17, 3.18–3.20 and 3.21–3.23. As shown in the results, the KNN estimators seem to be working well, and also the estimates of asymptotic variances. It is easy to realize that the bias of estimates are increasing when K changes from 1 to 20, whereas, the variances are decreasing. In this study, a choice of a small value of K (within the range 1 to 3) seems a good choice; nevertheless, it is worth noting that, in practice, the best choice of K might depend upon the dimension of the feature space as we already mentioned.

The second study concerns the advantage of the KNN estimator in the setting of misspecification models. The set–up of the disease status $\mathcal{D}$, diagnostic test T and covariate A presented in Study 3 of previous part are repeated. The verification status V is simulated through the following logistic model

$$\text{logit}\{\Pr(V = 1|T, A)\} = -1.5 - 0.35T - 1.5A.$$

Under this model, the verification rate is roughly 0.276. This has led us to the choice of $n = 1000$. For the cut point, we still consider the six pairs that employed in Study 3, i.e., $(-1.0, -0.5)$, $(-1.0, 0.7)$, $(-1.0, 1.3)$, $(-0.5, 0.7)$, $(-0.5, 1.3)$ and $(0.7, 1.3)$. Note that, in this study, both the estimates for $\hat{\rho}_{ki}$ and $\hat{\pi}_i$ obtained through parametric approaches are inconsistent, as $\hat{\rho}_{ki}$ could be obtained by from a multinomial logistic regression model with $\mathcal{D} = (D_1, D_2, D_3)$ as the response and T as predictor; and $\hat{\pi}_i$ could be estimated by a generalized linear model for V given T and $A^{2/3}$ with logit link. Clearly, the two fitted models are misspecified. The KNN estimators are obtained by using $K = 1$ and $K = 3$ and the Euclidean distance. Again, we use $\bar{K} = 2$ in the KNN procedure to estimate standard deviations of KNN estimators. Table 3.24 presents Monte

Table 3.13: Simulation results from 5000 replications when only model for ρ_k is misspecified (Study 3). "True" indicates the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (-1,-0.5)						
True	0.1812	0.1070	0.9817									
FI	**0.2144**	**0.1318**	0.9813	0.0230	0.0152	0.0051	0.0243	0.0135	0.0200	0.0230	0.0150	0.0052
MSI	**0.2172**	**0.1328**	0.9800	0.0237	0.0182	0.0074	0.0250	0.0166	0.0207	0.0237	0.0179	0.0075
IPW	0.1819	0.1072	0.9817	0.0258	0.0197	0.0091	0.0258	0.0194	0.0135	0.0260	0.0196	0.0092
SPE	0.1818	0.1073	0.9816	0.0208	0.0206	0.0093	0.0207	0.0202	0.0090	0.0208	0.0204	0.0094
						cut point = (-1,0.7)						
True	0.1812	0.8609	0.4469									
FI	**0.2144**	**0.8879**	**0.4010**	0.0230	0.0149	0.0284	0.0243	0.0153	0.0242	0.0230	0.0146	0.0284
MSI	**0.2172**	**0.8931**	**0.4035**	0.0237	0.0165	0.0292	0.0250	0.0172	0.0251	0.0237	0.0165	0.0292
IPW	0.1819	0.8606	0.4462	0.0258	0.0350	0.0437	0.0258	0.0342	0.0447	0.0260	0.0348	0.0437
SPE	0.1818	0.8608	0.4462	0.0208	0.0311	0.0455	0.0207	0.0305	0.0449	0.0208	0.0310	0.0482
						cut point = (-1,1.3)						
True	0.1812	0.9732	0.1171									
FI	**0.2144**	0.9672	**0.0949**	0.0230	0.0063	0.0161	0.0243	0.0099	0.0104	0.0230	0.0062	0.0161
MSI	**0.2172**	0.9708	**0.0960**	0.0237	0.0079	0.0164	0.0250	0.0110	0.0109	0.0237	0.0078	0.0164
IPW	0.1819	0.9734	0.1164	0.0258	0.0167	0.0358	0.0258	0.0130	0.0347	0.0260	0.0160	0.0354
SPE	0.1818	0.9734	0.1169	0.0208	0.0158	0.0281	0.0207	0.0128	0.0263	0.0208	0.0151	0.0333
						cut point = (-0.5,0.7)						
True	0.4796	0.7539	0.4469									
FI	**0.5497**	0.7561	**0.4010**	0.0302	0.0196	0.0284	0.0284	0.0183	0.0242	0.0301	0.0192	0.0284
MSI	**0.5502**	0.7603	**0.4035**	0.0312	0.0220	0.0292	0.0295	0.0211	0.0251	0.0310	0.0219	0.0292
IPW	0.4801	0.7534	0.4462	0.0390	0.0373	0.0437	0.0384	0.0371	0.0447	0.0387	0.0374	0.0437
SPE	0.4801	0.7535	0.4462	0.0327	0.0344	0.0455	0.0322	0.0339	0.0449	0.0324	0.0343	0.0482
						cut point = (-0.5,1.3)						
True	0.4796	0.8661	0.1171									
FI	**0.5497**	**0.8354**	**0.0949**	0.0302	0.0189	0.0161	0.0284	0.0185	0.0104	0.0301	0.0186	0.0161
MSI	**0.5502**	**0.8380**	**0.0960**	0.0312	0.0207	0.0164	0.0295	0.0204	0.0109	0.0310	0.0204	0.0164
IPW	0.4801	0.8661	0.1164	0.0390	0.0248	0.0358	0.0384	0.0238	0.0347	0.0387	0.0245	0.0354
SPE	0.4801	0.8660	0.1169	0.0327	0.0250	0.0281	0.0322	0.0239	0.0263	0.0324	0.0245	0.0333
						cut point = (0.7,1.3)						
True	0.9836	0.1122	0.1171									
FI	**0.9933**	**0.0793**	**0.0949**	0.0023	0.0133	0.0161	0.0021	0.0119	0.0104	0.0023	0.0131	0.0161
MSI	**0.9930**	**0.0777**	**0.0960**	0.0038	0.0145	0.0164	0.0032	0.0135	0.0109	0.0038	0.0145	0.0164
IPW	0.9839	0.1128	0.1164	0.0183	0.0324	0.0358	0.0122	0.0319	0.0347	0.0173	0.0325	0.0354
SPE	0.9839	0.1125	0.1169	0.0180	0.0283	0.0281	0.0122	0.0280	0.0263	0.0170	0.0285	0.0333

Table 3.14: Simulation results from 5000 replications when both models for ρ_k and π are misspecified (Study 4). "True" indicates the true parameter value. Sample size = 1000.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$	$boot.sd_1$	$boot.sd_2$	$boot.sd_3$
						cut point = (-1,-0.5)						
True	0.1812	0.1070	0.9817									
FI	**0.2143**	**0.1320**	0.9814	0.0231	0.0149	0.0051	0.0243	0.0135	0.0200	0.0230	0.0150	0.0052
MSI	**0.2170**	**0.1330**	0.9801	0.0238	0.0179	0.0074	0.0250	0.0166	0.0207	0.0237	0.0179	0.0075
IPW	**0.2185**	**0.1339**	0.9792	0.0284	0.0234	0.0102	0.0282	0.0232	0.0105	0.0283	0.0233	0.0102
SPE	**0.2183**	**0.1339**	0.9792	0.0247	0.0220	0.0101	0.0245	0.0219	0.0098	0.0246	0.0219	0.0102
						cut point = (-1,0.7)						
True	0.1812	0.8609	0.4469									
FI	**0.2143**	**0.8887**	**0.4002**	0.0231	0.0143	0.0285	0.0243	0.0153	0.0242	0.0230	0.0146	0.0285
MSI	**0.2170**	**0.8940**	**0.4029**	0.0238	0.0164	0.0290	0.0250	0.0171	0.0251	0.0237	0.0165	0.0292
IPW	**0.2185**	**0.8994**	**0.4078**	0.0284	0.0237	0.0397	0.0282	0.0232	0.0410	0.0283	0.0234	0.0397
SPE	**0.2183**	**0.8998**	**0.4071**	0.0247	0.0223	0.0323	0.0245	0.0219	0.0325	0.0246	0.0220	0.0326
						cut point = (-1,1.3)						
True	0.1812	0.9732	0.1171									
FI	**0.2143**	0.9675	**0.0947**	0.0231	0.0061	0.0160	0.0243	0.0099	0.0104	0.0230	0.0062	0.0161
MSI	**0.2170**	0.9711	**0.0958**	0.0238	0.0078	0.0163	0.0250	0.0110	0.0108	0.0237	0.0078	0.0164
IPW	**0.2185**	0.9742	**0.0977**	0.0284	0.0112	0.0269	0.0282	0.0107	0.0270	0.0283	0.0111	0.0273
SPE	**0.2183**	0.9742	**0.0978**	0.0247	0.0110	0.0174	0.0245	0.0105	0.0175	0.0246	0.0108	0.0177
						cut point = (-0.5,0.7)						
True	0.4796	0.7539	0.4469									
FI	**0.5510**	0.7567	**0.4002**	0.0306	0.0190	0.0285	0.0285	0.0183	0.0242	0.0301	0.0192	0.0285
MSI	**0.5514**	0.7610	**0.4029**	0.0316	0.0219	0.0290	0.0295	0.0211	0.0251	0.0310	0.0219	0.0292
IPW	**0.5509**	0.7655	**0.4078**	0.0360	0.0313	0.0397	0.0357	0.0310	0.0410	0.0358	0.0311	0.0397
SPE	**0.5509**	0.7659	**0.4071**	0.0336	0.0286	0.0323	0.0329	0.0286	0.0325	0.0329	0.0287	0.0326
						cut point = (-0.5,1.3)						
True	0.4796	0.8661	0.1171									
FI	**0.5510**	**0.8355**	**0.0947**	0.0306	0.0186	0.0160	0.0285	0.0186	0.0104	0.0301	0.0186	0.0161
MSI	**0.5514**	**0.8380**	**0.0958**	0.0316	0.0205	0.0163	0.0295	0.0204	0.0108	0.0310	0.0204	0.0164
IPW	**0.5509**	**0.8403**	**0.0977**	0.0360	0.0255	0.0269	0.0357	0.0251	0.0270	0.0358	0.0252	0.0273
SPE	**0.5509**	**0.8403**	**0.0978**	0.0336	0.0240	0.0174	0.0329	0.0237	0.0175	0.0329	0.0238	0.0177
						cut point = (0.7,1.3)						
True	0.9836	0.1122	0.1171									
FI	**0.9934**	**0.0788**	**0.0947**	0.0022	0.0129	0.0160	0.0021	0.0119	0.0104	0.0023	0.0131	0.0161
MSI	**0.9930**	**0.0771**	**0.0958**	0.0038	0.0145	0.0163	0.0032	0.0134	0.0108	0.0037	0.0145	0.0164
IPW	**0.9925**	**0.0748**	**0.0977**	0.0075	0.0213	0.0269	0.0057	0.0208	0.0270	0.0073	0.0211	0.0273
SPE	**0.9925**	**0.0744**	**0.0978**	0.0074	0.0201	0.0174	0.0058	0.0196	0.0175	0.0073	0.0198	0.0177

Table 3.15: Simulation results of the KNN estimators for TCFs. The sample size equals to 250 and the first value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.4347	0.9347						
1NN	0.4989	0.4334	0.9331	0.0592	0.0665	0.0387	0.0555	0.0626	0.0382
3NN	0.4975	0.4325	0.9322	0.0567	0.0617	0.0364	0.0545	0.0608	0.0372
5NN	0.4965	0.4320	0.9315	0.0559	0.0600	0.0360	0.0543	0.0604	0.0372
10NN	0.4943	0.4306	0.9297	0.0551	0.0580	0.0357	0.0542	0.0600	0.0376
20NN	0.4902	0.4278	0.9258	0.0542	0.0557	0.0358	0.0541	0.0595	0.0384
					cut point $= (2, 5)$				
True	0.5000	0.7099	0.7752						
1NN	0.4989	0.7068	0.7738	0.0592	0.0627	0.0652	0.0555	0.0591	0.0625
3NN	0.4975	0.7038	0.7714	0.0567	0.0576	0.0615	0.0545	0.0574	0.0610
5NN	0.4965	0.7016	0.7698	0.0559	0.0558	0.0607	0.0543	0.0571	0.0609
10NN	0.4943	0.6967	0.7662	0.0551	0.0535	0.0599	0.0542	0.0568	0.0612
20NN	0.4902	0.6881	0.7595	0.0542	0.0535	0.0594	0.0541	0.0567	0.0612
					cut point $= (2, 7)$				
True	0.5000	0.9230	0.2248						
1NN	0.4989	0.9201	0.2233	0.0592	0.0372	0.0577	0.0555	0.0366	0.0570
3NN	0.4975	0.9177	0.2216	0.0567	0.0340	0.0558	0.0545	0.0355	0.0563
5NN	0.4965	0.9157	0.2205	0.0559	0.0330	0.0550	0.0543	0.0355	0.0561
10NN	0.4943	0.9112	0.2184	0.0551	0.0318	0.0542	0.0542	0.0358	0.0560
20NN	0.4902	0.9031	0.2145	0.0542	0.0318	0.0531	0.0541	0.0366	0.0559
					cut point $= (4, 5)$				
True	0.9347	0.2752	0.7752						
1NN	0.9322	0.2734	0.7738	0.0374	0.0572	0.0652	0.0342	0.0553	0.0625
3NN	0.9303	0.2712	0.7714	0.0328	0.0526	0.0615	0.0332	0.0538	0.0610
5NN	0.9288	0.2696	0.7698	0.0315	0.0512	0.0607	0.0332	0.0534	0.0609
10NN	0.9255	0.2662	0.7662	0.0301	0.0489	0.0599	0.0335	0.0529	0.0612
20NN	0.9196	0.2603	0.7595	0.0291	0.0467	0.0594	0.0342	0.0522	0.0612
					cut point $= (4, 7)$				
True	0.9347	0.4883	0.2248						
1NN	0.9322	0.4867	0.2233	0.0374	0.0680	0.0577	0.0342	0.0633	0.0570
3NN	0.9303	0.4852	0.2216	0.0328	0.0630	0.0558	0.0332	0.0615	0.0563
5NN	0.9288	0.4837	0.2205	0.0315	0.0615	0.0550	0.0332	0.0611	0.0561
10NN	0.9255	0.4807	0.2184	0.0301	0.0597	0.0542	0.0335	0.0606	0.0560
20NN	0.9196	0.4753	0.2145	0.0291	0.0577	0.0531	0.0342	0.0602	0.0559
					cut point $= (5, 7)$				
True	0.9883	0.2132	0.2248						
1NN	0.9868	0.2133	0.2233	0.0177	0.0567	0.0577	0.0172	0.0532	0.0570
3NN	0.9860	0.2139	0.2216	0.0151	0.0519	0.0558	0.0168	0.0516	0.0563
5NN	0.9851	0.2141	0.2205	0.0142	0.0502	0.0550	0.0170	0.0512	0.0561
10NN	0.9833	0.2145	0.2184	0.0135	0.0479	0.0542	0.0174	0.0508	0.0560
20NN	0.9800	0.2150	0.2145	0.0131	0.0453	0.0531	0.0183	0.0505	0.0559

Table 3.16: Simulation results of the KNN estimators for TCFs. The sample size equals to 500 and the first value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.4347	0.9347						
1NN	0.5000	0.4343	0.9333	0.0419	0.0464	0.0285	0.0393	0.0443	0.0269
3NN	0.4991	0.4343	0.9329	0.0401	0.0430	0.0266	0.0386	0.0431	0.0262
5NN	0.4984	0.4339	0.9324	0.0396	0.0419	0.0262	0.0385	0.0429	0.0261
10NN	0.4971	0.4332	0.9315	0.0391	0.0407	0.0258	0.0384	0.0426	0.0262
20NN	0.4948	0.4317	0.9296	0.0386	0.0394	0.0256	0.0383	0.0424	0.0265
					cut point $= (2, 5)$				
True	0.5000	0.7099	0.7752						
1NN	0.5000	0.7077	0.7732	0.0419	0.0450	0.0466	0.0393	0.0417	0.0442
3NN	0.4991	0.7066	0.7722	0.0401	0.0415	0.0437	0.0386	0.0406	0.0431
5NN	0.4984	0.7053	0.7712	0.0396	0.0404	0.0430	0.0385	0.0403	0.0429
10NN	0.4971	0.7025	0.7693	0.0391	0.0391	0.0423	0.0384	0.0402	0.0430
20NN	0.4948	0.6975	0.7656	0.0386	0.0391	0.0417	0.0383	0.0401	0.0430
					cut point $= (2, 7)$				
True	0.5000	0.9230	0.2248						
1NN	0.5000	0.9214	0.2235	0.0419	0.0261	0.0407	0.0393	0.0256	0.0401
3NN	0.4991	0.9199	0.2225	0.0401	0.0240	0.0392	0.0386	0.0248	0.0396
5NN	0.4984	0.9186	0.2218	0.0396	0.0235	0.0388	0.0385	0.0248	0.0395
10NN	0.4971	0.9160	0.2205	0.0391	0.0229	0.0384	0.0384	0.0248	0.0394
20NN	0.4948	0.9115	0.2183	0.0386	0.0229	0.0379	0.0383	0.0251	0.0394
					cut point $= (4, 5)$				
True	0.9347	0.2752	0.7752						
1NN	0.9332	0.2734	0.7732	0.0269	0.0398	0.0466	0.0242	0.0391	0.0442
3NN	0.9317	0.2723	0.7722	0.0242	0.0371	0.0437	0.0235	0.0381	0.0431
5NN	0.9308	0.2713	0.7712	0.0233	0.0360	0.0430	0.0234	0.0379	0.0429
10NN	0.9287	0.2693	0.7693	0.0225	0.0350	0.0423	0.0235	0.0376	0.0430
20NN	0.9254	0.2658	0.7656	0.0215	0.0337	0.0417	0.0237	0.0373	0.0430
					cut point $= (4, 7)$				
True	0.9347	0.4883	0.2248						
1NN	0.9332	0.4871	0.2235	0.0269	0.0465	0.0407	0.0242	0.0447	0.0401
3NN	0.9317	0.4855	0.2225	0.0242	0.0434	0.0392	0.0235	0.0436	0.0396
5NN	0.9308	0.4847	0.2218	0.0233	0.0424	0.0388	0.0234	0.0433	0.0395
10NN	0.9287	0.4829	0.2205	0.0225	0.0414	0.0384	0.0235	0.0431	0.0394
20NN	0.9254	0.4798	0.2183	0.0215	0.0403	0.0379	0.0237	0.0428	0.0394
					cut point $= (5, 7)$				
True	0.9883	0.2132	0.2248						
1NN	0.9875	0.2136	0.2235	0.0129	0.0400	0.0407	0.0120	0.0376	0.0401
3NN	0.9867	0.2133	0.2225	0.0113	0.0369	0.0392	0.0118	0.0365	0.0396
5NN	0.9861	0.2133	0.2218	0.0107	0.0359	0.0388	0.0118	0.0363	0.0395
10NN	0.9850	0.2135	0.2205	0.0102	0.0345	0.0384	0.0120	0.0361	0.0394
20NN	0.9833	0.2140	0.2183	0.0098	0.0332	0.0379	0.0123	0.0359	0.0394

Table 3.17: Simulation results of the KNN estimators for TCFs. The sample size equals to 1000 and the first value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.4347	0.9347						
1NN	0.5002	0.4349	0.9340	0.0295	0.0336	0.0194	0.0278	0.0313	0.0189
3NN	0.4997	0.4346	0.9336	0.0283	0.0318	0.0183	0.0273	0.0305	0.0183
5NN	0.4993	0.4344	0.9333	0.0280	0.0313	0.0180	0.0272	0.0304	0.0183
10NN	0.4985	0.4341	0.9328	0.0276	0.0304	0.0178	0.0272	0.0302	0.0182
20NN	0.4973	0.4333	0.9318	0.0273	0.0296	0.0176	0.0271	0.0301	0.0183
					cut point $= (2, 5)$				
True	0.5000	0.7099	0.7752						
1NN	0.5002	0.7093	0.7741	0.0295	0.0319	0.0331	0.0278	0.0295	0.0312
3NN	0.4997	0.7083	0.7733	0.0283	0.0297	0.0312	0.0273	0.0287	0.0304
5NN	0.4993	0.7075	0.7728	0.0280	0.0292	0.0306	0.0272	0.0285	0.0303
10NN	0.4985	0.7057	0.7715	0.0276	0.0282	0.0301	0.0272	0.0284	0.0302
20NN	0.4973	0.7028	0.7694	0.0273	0.0282	0.0298	0.0271	0.0283	0.0302
					cut point $= (2, 7)$				
True	0.5000	0.9230	0.2248						
1NN	0.5002	0.9223	0.2239	0.0295	0.0187	0.0292	0.0278	0.0179	0.0283
3NN	0.4997	0.9213	0.2234	0.0283	0.0173	0.0283	0.0273	0.0174	0.0280
5NN	0.4993	0.9206	0.2229	0.0280	0.0169	0.0281	0.0272	0.0173	0.0279
10NN	0.4985	0.9190	0.2220	0.0276	0.0164	0.0279	0.0272	0.0173	0.0279
20NN	0.4973	0.9163	0.2208	0.0273	0.0164	0.0276	0.0271	0.0174	0.0278
					cut point $= (4, 5)$				
True	0.9347	0.2752	0.7752						
1NN	0.9337	0.2745	0.7741	0.0198	0.0286	0.0331	0.0172	0.0277	0.0312
3NN	0.9329	0.2737	0.7733	0.0179	0.0268	0.0312	0.0166	0.0270	0.0304
5NN	0.9323	0.2731	0.7728	0.0173	0.0263	0.0306	0.0166	0.0269	0.0303
10NN	0.9310	0.2717	0.7715	0.0166	0.0256	0.0301	0.0166	0.0267	0.0302
20NN	0.9289	0.2695	0.7694	0.0159	0.0248	0.0298	0.0166	0.0266	0.0302
					cut point $= (4, 7)$				
True	0.9347	0.4883	0.2248						
1NN	0.9337	0.4874	0.2239	0.0198	0.0342	0.0292	0.0172	0.0316	0.0283
3NN	0.9329	0.4867	0.2234	0.0179	0.0322	0.0283	0.0166	0.0308	0.0280
5NN	0.9323	0.4861	0.2229	0.0173	0.0316	0.0281	0.0166	0.0307	0.0279
10NN	0.9310	0.4849	0.2220	0.0166	0.0308	0.0279	0.0166	0.0305	0.0279
20NN	0.9289	0.4830	0.2208	0.0159	0.0300	0.0276	0.0166	0.0304	0.0278
					cut point $= (5, 7)$				
True	0.9883	0.2132	0.2248						
1NN	0.9876	0.2130	0.2239	0.0096	0.0289	0.0292	0.0085	0.0266	0.0283
3NN	0.9871	0.2130	0.2234	0.0084	0.0268	0.0283	0.0083	0.0259	0.0280
5NN	0.9869	0.2131	0.2229	0.0080	0.0263	0.0281	0.0083	0.0257	0.0279
10NN	0.9862	0.2132	0.2220	0.0075	0.0253	0.0279	0.0083	0.0256	0.0279
20NN	0.9851	0.2135	0.2208	0.0072	0.0244	0.0276	0.0084	0.0255	0.0278

Table 3.18: Simulation results of the KNN estimators for TCFs. The sample size equals to 250 and the second value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3970	0.8970						
1NN	0.4982	0.3953	0.8976	0.0587	0.0642	0.0537	0.0561	0.0618	0.0487
3NN	0.4960	0.3933	0.8970	0.0556	0.0595	0.0494	0.0548	0.0600	0.0472
5NN	0.4941	0.3917	0.8966	0.0548	0.0578	0.0484	0.0545	0.0596	0.0471
10NN	0.4904	0.3885	0.8955	0.0540	0.0561	0.0473	0.0542	0.0591	0.0473
20NN	0.4841	0.3834	0.8926	0.0533	0.0539	0.0467	0.0540	0.0587	0.0479
					cut point $= (2, 5)$				
True	0.5000	0.6335	0.7365						
1NN	0.4982	0.6304	0.7400	0.0587	0.0645	0.0721	0.0561	0.0615	0.0672
3NN	0.4960	0.6283	0.7396	0.0556	0.0600	0.0670	0.0548	0.0597	0.0654
5NN	0.4941	0.6260	0.7392	0.0548	0.0587	0.0661	0.0545	0.0594	0.0652
10NN	0.4904	0.6212	0.7378	0.0540	0.0569	0.0649	0.0542	0.0591	0.0653
20NN	0.4841	0.6133	0.7345	0.0533	0.0569	0.0641	0.0540	0.0590	0.0653
					cut point $= (2, 7)$				
True	0.5000	0.8682	0.2635						
1NN	0.4982	0.8672	0.2672	0.0587	0.0495	0.0629	0.0561	0.0458	0.0609
3NN	0.4960	0.8657	0.2671	0.0556	0.0452	0.0610	0.0548	0.0442	0.0601
5NN	0.4941	0.8642	0.2671	0.0548	0.0442	0.0605	0.0545	0.0440	0.0601
10NN	0.4904	0.8608	0.2667	0.0540	0.0430	0.0602	0.0542	0.0441	0.0602
20NN	0.4841	0.8539	0.2655	0.0533	0.0430	0.0598	0.0540	0.0444	0.0604
					cut point $= (4, 5)$				
True	0.8970	0.2365	0.7365						
1NN	0.8958	0.2352	0.7400	0.0388	0.0540	0.0721	0.0373	0.0524	0.0672
3NN	0.8946	0.2350	0.7396	0.0362	0.0502	0.0670	0.0361	0.0510	0.0654
5NN	0.8933	0.2343	0.7392	0.0355	0.0490	0.0661	0.0360	0.0507	0.0652
10NN	0.8905	0.2328	0.7378	0.0348	0.0474	0.0649	0.0360	0.0503	0.0653
20NN	0.8857	0.2299	0.7345	0.0343	0.0455	0.0641	0.0364	0.0499	0.0653
					cut point $= (4, 7)$				
True	0.8970	0.4711	0.2635						
1NN	0.8958	0.4719	0.2672	0.0388	0.0666	0.0629	0.0373	0.0630	0.0609
3NN	0.8946	0.4724	0.2671	0.0362	0.0627	0.0610	0.0361	0.0611	0.0601
5NN	0.8933	0.4725	0.2671	0.0355	0.0612	0.0605	0.0360	0.0607	0.0601
10NN	0.8905	0.4723	0.2667	0.0348	0.0600	0.0602	0.0360	0.0604	0.0602
20NN	0.8857	0.4705	0.2655	0.0343	0.0584	0.0598	0.0364	0.0600	0.0604
					cut point $= (5, 7)$				
True	0.9711	0.2347	0.2635						
1NN	0.9701	0.2368	0.2672	0.0217	0.0549	0.0629	0.0213	0.0533	0.0609
3NN	0.9695	0.2375	0.2671	0.0200	0.0519	0.0610	0.0206	0.0517	0.0601
5NN	0.9689	0.2382	0.2671	0.0197	0.0507	0.0605	0.0205	0.0515	0.0601
10NN	0.9675	0.2395	0.2667	0.0194	0.0492	0.0602	0.0207	0.0512	0.0602
20NN	0.9648	0.2406	0.2655	0.0192	0.0478	0.0598	0.0212	0.0510	0.0604

Table 3.19: Simulation results of the KNN estimators for TCFs. The sample size equals to 500 and the second value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3970	0.8970						
1NN	0.4987	0.3958	0.8975	0.0419	0.0466	0.0379	0.0396	0.0438	0.0347
3NN	0.4975	0.3947	0.8971	0.0400	0.0431	0.0349	0.0388	0.0425	0.0335
5NN	0.4965	0.3937	0.8968	0.0395	0.0423	0.0342	0.0386	0.0423	0.0334
10NN	0.4945	0.3918	0.8962	0.0391	0.0412	0.0335	0.0384	0.0420	0.0334
20NN	0.4909	0.3886	0.8951	0.0386	0.0400	0.0328	0.0383	0.0418	0.0336
					cut point $= (2, 5)$				
True	0.5000	0.6335	0.7365						
1NN	0.4987	0.6318	0.7376	0.0419	0.0466	0.0510	0.0396	0.0436	0.0477
3NN	0.4975	0.6306	0.7372	0.0400	0.0433	0.0472	0.0388	0.0423	0.0464
5NN	0.4965	0.6294	0.7371	0.0395	0.0426	0.0464	0.0386	0.0420	0.0462
10NN	0.4945	0.6268	0.7366	0.0391	0.0416	0.0456	0.0384	0.0419	0.0461
20NN	0.4909	0.6222	0.7355	0.0386	0.0416	0.0448	0.0383	0.0418	0.0461
					cut point $= (2, 7)$				
True	0.5000	0.8682	0.2635						
1NN	0.4987	0.8675	0.2634	0.0419	0.0352	0.0441	0.0396	0.0325	0.0425
3NN	0.4975	0.8667	0.2633	0.0400	0.0324	0.0427	0.0388	0.0313	0.0420
5NN	0.4965	0.8661	0.2634	0.0395	0.0316	0.0424	0.0386	0.0311	0.0419
10NN	0.4945	0.8644	0.2634	0.0391	0.0307	0.0423	0.0384	0.0311	0.0420
20NN	0.4909	0.8610	0.2632	0.0386	0.0307	0.0422	0.0383	0.0311	0.0421
					cut point $= (4, 5)$				
True	0.8970	0.2365	0.7365						
1NN	0.8964	0.2360	0.7376	0.0279	0.0388	0.0510	0.0263	0.0372	0.0477
3NN	0.8954	0.2359	0.7372	0.0264	0.0362	0.0472	0.0255	0.0362	0.0464
5NN	0.8947	0.2357	0.7371	0.0259	0.0353	0.0464	0.0253	0.0360	0.0462
10NN	0.8932	0.2350	0.7366	0.0254	0.0343	0.0456	0.0253	0.0358	0.0461
20NN	0.8905	0.2335	0.7355	0.0250	0.0332	0.0448	0.0254	0.0356	0.0461
					cut point $= (4, 7)$				
True	0.8970	0.4711	0.2635						
1NN	0.8964	0.4717	0.2634	0.0279	0.0480	0.0441	0.0263	0.0446	0.0425
3NN	0.8954	0.4720	0.2633	0.0264	0.0448	0.0427	0.0255	0.0433	0.0420
5NN	0.8947	0.4723	0.2634	0.0259	0.0439	0.0424	0.0253	0.0431	0.0419
10NN	0.8932	0.4726	0.2634	0.0254	0.0430	0.0423	0.0253	0.0429	0.0420
20NN	0.8905	0.4723	0.2632	0.0250	0.0421	0.0422	0.0254	0.0427	0.0421
					cut point $= (5, 7)$				
True	0.9711	0.2347	0.2635						
1NN	0.9707	0.2357	0.2634	0.0150	0.0400	0.0441	0.0148	0.0376	0.0425
3NN	0.9701	0.2360	0.2633	0.0142	0.0373	0.0427	0.0144	0.0366	0.0420
5NN	0.9697	0.2367	0.2634	0.0139	0.0366	0.0424	0.0143	0.0364	0.0419
10NN	0.9690	0.2376	0.2634	0.0137	0.0358	0.0423	0.0143	0.0362	0.0420
20NN	0.9676	0.2388	0.2632	0.0136	0.0350	0.0422	0.0145	0.0361	0.0421

Table 3.20: Simulation results of the KNN estimators for TCFs. The sample size equals to 500 and the second value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3970	0.8970						
1NN	0.4989	0.3958	0.8974	0.0291	0.0328	0.0279	0.0280	0.0310	0.0246
3NN	0.4981	0.3950	0.8974	0.0279	0.0307	0.0258	0.0274	0.0301	0.0238
5NN	0.4975	0.3943	0.8973	0.0275	0.0301	0.0252	0.0273	0.0300	0.0236
10NN	0.4962	0.3931	0.8972	0.0271	0.0294	0.0246	0.0272	0.0298	0.0236
20NN	0.4940	0.3910	0.8966	0.0268	0.0288	0.0242	0.0271	0.0297	0.0236
					cut point $= (2, 5)$				
True	0.5000	0.6335	0.7365						
1NN	0.4989	0.6322	0.7372	0.0291	0.0325	0.0369	0.0280	0.0308	0.0338
3NN	0.4981	0.6312	0.7372	0.0279	0.0303	0.0345	0.0274	0.0299	0.0328
5NN	0.4975	0.6305	0.7372	0.0275	0.0298	0.0338	0.0273	0.0298	0.0327
10NN	0.4962	0.6289	0.7371	0.0271	0.0291	0.0332	0.0272	0.0296	0.0326
20NN	0.4940	0.6261	0.7367	0.0268	0.0291	0.0328	0.0271	0.0296	0.0326
					cut point $= (2, 7)$				
True	0.5000	0.8682	0.2635						
1NN	0.4989	0.8684	0.2643	0.0291	0.0247	0.0313	0.0280	0.0229	0.0300
3NN	0.4981	0.8678	0.2643	0.0279	0.0228	0.0304	0.0274	0.0221	0.0297
5NN	0.4975	0.8674	0.2644	0.0275	0.0223	0.0301	0.0273	0.0220	0.0296
10NN	0.4962	0.8665	0.2645	0.0271	0.0217	0.0299	0.0272	0.0219	0.0296
20NN	0.4940	0.8648	0.2645	0.0268	0.0217	0.0299	0.0271	0.0219	0.0297
					cut point $= (4, 5)$				
True	0.8970	0.2365	0.7365						
1NN	0.8963	0.2364	0.7372	0.0198	0.0276	0.0369	0.0185	0.0263	0.0338
3NN	0.8958	0.2362	0.7372	0.0186	0.0260	0.0345	0.0179	0.0256	0.0328
5NN	0.8954	0.2361	0.7372	0.0183	0.0254	0.0338	0.0178	0.0255	0.0327
10NN	0.8945	0.2358	0.7371	0.0179	0.0248	0.0332	0.0178	0.0254	0.0326
20NN	0.8930	0.2351	0.7367	0.0177	0.0242	0.0328	0.0178	0.0253	0.0326
					cut point $= (4, 7)$				
True	0.8970	0.4711	0.2635						
1NN	0.8963	0.4726	0.2643	0.0198	0.0342	0.0313	0.0185	0.0316	0.0300
3NN	0.8958	0.4728	0.2643	0.0186	0.0320	0.0304	0.0179	0.0307	0.0297
5NN	0.8954	0.4731	0.2644	0.0183	0.0314	0.0301	0.0178	0.0305	0.0296
10NN	0.8945	0.4734	0.2645	0.0179	0.0307	0.0299	0.0178	0.0304	0.0296
20NN	0.8930	0.4737	0.2645	0.0177	0.0302	0.0299	0.0178	0.0303	0.0297
					cut point $= (5, 7)$				
True	0.9711	0.2347	0.2635						
1NN	0.9709	0.2362	0.2643	0.0111	0.0281	0.0313	0.0104	0.0266	0.0300
3NN	0.9707	0.2366	0.2643	0.0103	0.0262	0.0304	0.0100	0.0259	0.0297
5NN	0.9704	0.2369	0.2644	0.0101	0.0257	0.0301	0.0100	0.0258	0.0296
10NN	0.9700	0.2376	0.2645	0.0099	0.0252	0.0299	0.0100	0.0257	0.0296
20NN	0.9692	0.2386	0.2645	0.0098	0.0247	0.0299	0.0100	0.0256	0.0297

Table 3.21: Simulation results of the KNN estimators for TCFs. The sample size equals to 250 and the third value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3031	0.8031						
1NN	0.4997	0.3021	0.8047	0.0592	0.0602	0.0682	0.0571	0.0584	0.0621
3NN	0.4984	0.3018	0.8043	0.0561	0.0565	0.0632	0.0556	0.0566	0.0601
5NN	0.4975	0.3016	0.8039	0.0554	0.0552	0.0621	0.0553	0.0562	0.0598
10NN	0.4956	0.3006	0.8029	0.0549	0.0539	0.0611	0.0550	0.0558	0.0597
20NN	0.4915	0.2986	0.8005	0.0543	0.0523	0.0605	0.0549	0.0555	0.0600
					cut point $= (2, 5)$				
True	0.5000	0.4682	0.6651						
1NN	0.4997	0.4676	0.6668	0.0592	0.0661	0.0780	0.0571	0.0634	0.0717
3NN	0.4984	0.4670	0.6666	0.0561	0.0619	0.0729	0.0556	0.0614	0.0695
5NN	0.4975	0.4665	0.6664	0.0554	0.0606	0.0717	0.0553	0.0610	0.0692
10NN	0.4956	0.4652	0.6654	0.0549	0.0593	0.0706	0.0550	0.0606	0.0691
20NN	0.4915	0.4623	0.6630	0.0543	0.0593	0.0698	0.0549	0.0604	0.0691
					cut point $= (2, 7)$				
True	0.5000	0.7027	0.3349						
1NN	0.4997	0.7024	0.3366	0.0592	0.0633	0.0712	0.0571	0.0592	0.0675
3NN	0.4984	0.7016	0.3362	0.0561	0.0590	0.0680	0.0556	0.0572	0.0660
5NN	0.4975	0.7011	0.3361	0.0554	0.0578	0.0674	0.0553	0.0568	0.0657
10NN	0.4956	0.6995	0.3353	0.0549	0.0565	0.0666	0.0550	0.0565	0.0656
20NN	0.4915	0.6959	0.3332	0.0543	0.0565	0.0659	0.0549	0.0564	0.0656
					cut point $= (4, 5)$				
True	0.8031	0.1651	0.6651						
1NN	0.8032	0.1655	0.6668	0.0487	0.0481	0.0780	0.0472	0.0466	0.0717
3NN	0.8020	0.1651	0.6666	0.0460	0.0450	0.0729	0.0457	0.0451	0.0695
5NN	0.8011	0.1649	0.6664	0.0453	0.0439	0.0717	0.0455	0.0448	0.0692
10NN	0.7996	0.1646	0.6654	0.0446	0.0427	0.0706	0.0453	0.0446	0.0691
20NN	0.7965	0.1637	0.6630	0.0441	0.0412	0.0698	0.0454	0.0443	0.0691
					cut point $= (4, 7)$				
True	0.8031	0.3996	0.3349						
1NN	0.8032	0.4003	0.3366	0.0487	0.0660	0.0712	0.0472	0.0619	0.0675
3NN	0.8020	0.3998	0.3362	0.0460	0.0617	0.0680	0.0457	0.0600	0.0660
5NN	0.8011	0.3995	0.3361	0.0453	0.0604	0.0674	0.0455	0.0596	0.0657
10NN	0.7996	0.3989	0.3353	0.0446	0.0594	0.0666	0.0453	0.0592	0.0656
20NN	0.7965	0.3973	0.3332	0.0441	0.0581	0.0659	0.0454	0.0589	0.0656
					cut point $= (5, 7)$				
True	0.8996	0.2345	0.3349						
1NN	0.9000	0.2348	0.3366	0.0373	0.0556	0.0712	0.0361	0.0531	0.0675
3NN	0.8992	0.2346	0.3362	0.0349	0.0520	0.0680	0.0349	0.0515	0.0660
5NN	0.8987	0.2346	0.3361	0.0345	0.0510	0.0674	0.0347	0.0511	0.0657
10NN	0.8974	0.2343	0.3353	0.0340	0.0499	0.0666	0.0346	0.0508	0.0656
20NN	0.8952	0.2335	0.3332	0.0336	0.0485	0.0659	0.0348	0.0506	0.0656

Table 3.22: Simulation results of the KNN estimators for TCFs. The sample size equals to 500 and the third value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3031	0.8031						
1NN	0.4994	0.3024	0.8036	0.0428	0.0441	0.0489	0.0403	0.0412	0.0442
3NN	0.4990	0.3020	0.8033	0.0409	0.0412	0.0454	0.0394	0.0399	0.0427
5NN	0.4984	0.3017	0.8032	0.0403	0.0402	0.0447	0.0391	0.0397	0.0425
10NN	0.4971	0.3010	0.8028	0.0398	0.0392	0.0438	0.0390	0.0395	0.0423
20NN	0.4951	0.3000	0.8019	0.0394	0.0384	0.0433	0.0388	0.0393	0.0424
					cut point $= (2, 5)$				
True	0.5000	0.4682	0.6651						
1NN	0.4994	0.4677	0.6653	0.0428	0.0476	0.0545	0.0403	0.0447	0.0509
3NN	0.4990	0.4672	0.6650	0.0409	0.0447	0.0512	0.0394	0.0434	0.0494
5NN	0.4984	0.4668	0.6648	0.0403	0.0437	0.0505	0.0391	0.0431	0.0491
10NN	0.4971	0.4659	0.6645	0.0398	0.0428	0.0498	0.0390	0.0429	0.0490
20NN	0.4951	0.4645	0.6638	0.0394	0.0428	0.0493	0.0388	0.0427	0.0490
					cut point $= (2, 7)$				
True	0.5000	0.7027	0.3349						
1NN	0.4994	0.7018	0.3353	0.0428	0.0448	0.0495	0.0403	0.0418	0.0477
3NN	0.4990	0.7018	0.3353	0.0409	0.0419	0.0473	0.0394	0.0404	0.0466
5NN	0.4984	0.7014	0.3352	0.0403	0.0410	0.0469	0.0391	0.0401	0.0464
10NN	0.4971	0.7004	0.3349	0.0398	0.0403	0.0466	0.0390	0.0399	0.0463
20NN	0.4951	0.6988	0.3342	0.0394	0.0403	0.0463	0.0388	0.0399	0.0464
					cut point $= (4, 5)$				
True	0.8031	0.1651	0.6651						
1NN	0.8027	0.1653	0.6653	0.0357	0.0342	0.0545	0.0334	0.0328	0.0509
3NN	0.8024	0.1652	0.6650	0.0335	0.0319	0.0512	0.0324	0.0319	0.0494
5NN	0.8020	0.1652	0.6648	0.0329	0.0313	0.0505	0.0322	0.0317	0.0491
10NN	0.8010	0.1648	0.6645	0.0325	0.0304	0.0498	0.0320	0.0315	0.0490
20NN	0.7995	0.1645	0.6638	0.0322	0.0295	0.0493	0.0320	0.0314	0.0490
					cut point $= (4, 7)$				
True	0.8031	0.3996	0.3349						
1NN	0.8027	0.3994	0.3353	0.0357	0.0469	0.0495	0.0334	0.0436	0.0477
3NN	0.8024	0.3997	0.3353	0.0335	0.0438	0.0473	0.0324	0.0423	0.0466
5NN	0.8020	0.3997	0.3352	0.0329	0.0432	0.0469	0.0322	0.0421	0.0464
10NN	0.8010	0.3994	0.3349	0.0325	0.0424	0.0466	0.0320	0.0418	0.0463
20NN	0.7995	0.3988	0.3342	0.0322	0.0416	0.0463	0.0320	0.0417	0.0464
					cut point $= (5, 7)$				
True	0.8996	0.2345	0.3349						
1NN	0.8990	0.2341	0.3353	0.0271	0.0397	0.0495	0.0256	0.0374	0.0477
3NN	0.8988	0.2345	0.3353	0.0254	0.0374	0.0473	0.0247	0.0363	0.0466
5NN	0.8985	0.2345	0.3352	0.0249	0.0368	0.0469	0.0246	0.0361	0.0464
10NN	0.8979	0.2345	0.3349	0.0246	0.0361	0.0466	0.0245	0.0359	0.0463
20NN	0.8968	0.2343	0.3342	0.0244	0.0353	0.0463	0.0245	0.0358	0.0464

Table 3.23: Simulation results of the KNN estimators for TCFs. The sample size equals to 1000 and the third value of Λ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.3031	0.8031						
1NN	0.4999	0.3023	0.8027	0.0295	0.0313	0.0341	0.0285	0.0291	0.0313
3NN	0.4994	0.3023	0.8030	0.0281	0.0295	0.0316	0.0278	0.0283	0.0302
5NN	0.4991	0.3021	0.8030	0.0277	0.0290	0.0309	0.0277	0.0281	0.0300
10NN	0.4986	0.3018	0.8029	0.0274	0.0284	0.0303	0.0276	0.0279	0.0299
20NN	0.4975	0.3012	0.8026	0.0271	0.0280	0.0299	0.0275	0.0278	0.0299
					cut point $= (2, 5)$				
True	0.5000	0.4682	0.6651						
1NN	0.4999	0.4677	0.6650	0.0295	0.0344	0.0384	0.0285	0.0316	0.0360
3NN	0.4994	0.4676	0.6653	0.0281	0.0324	0.0363	0.0278	0.0307	0.0349
5NN	0.4991	0.4674	0.6653	0.0277	0.0319	0.0356	0.0277	0.0305	0.0346
10NN	0.4986	0.4670	0.6652	0.0274	0.0313	0.0350	0.0276	0.0303	0.0345
20NN	0.4975	0.4662	0.6649	0.0271	0.0313	0.0346	0.0275	0.0302	0.0345
					cut point $= (2, 7)$				
True	0.5000	0.7027	0.3349						
1NN	0.4999	0.7029	0.3351	0.0295	0.0323	0.0340	0.0285	0.0295	0.0336
3NN	0.4994	0.7027	0.3351	0.0281	0.0302	0.0328	0.0278	0.0286	0.0328
5NN	0.4991	0.7025	0.3351	0.0277	0.0296	0.0325	0.0277	0.0284	0.0327
10NN	0.4986	0.7021	0.3350	0.0274	0.0291	0.0323	0.0276	0.0282	0.0326
20NN	0.4975	0.7014	0.3348	0.0271	0.0291	0.0321	0.0275	0.0282	0.0326
					cut point $= (4, 5)$				
True	0.8031	0.1651	0.6651						
1NN	0.8029	0.1653	0.6650	0.0245	0.0240	0.0384	0.0236	0.0232	0.0360
3NN	0.8025	0.1652	0.6653	0.0233	0.0226	0.0363	0.0229	0.0225	0.0349
5NN	0.8023	0.1653	0.6653	0.0229	0.0222	0.0356	0.0227	0.0224	0.0346
10NN	0.8019	0.1652	0.6652	0.0226	0.0218	0.0350	0.0226	0.0223	0.0345
20NN	0.8011	0.1650	0.6649	0.0224	0.0213	0.0346	0.0226	0.0222	0.0345
					cut point $= (4, 7)$				
True	0.8031	0.3996	0.3349						
1NN	0.8029	0.4006	0.3351	0.0245	0.0323	0.0340	0.0236	0.0308	0.0336
3NN	0.8025	0.4003	0.3351	0.0233	0.0304	0.0328	0.0229	0.0299	0.0328
5NN	0.8023	0.4004	0.3351	0.0229	0.0300	0.0325	0.0227	0.0298	0.0327
10NN	0.8019	0.4003	0.3350	0.0226	0.0295	0.0323	0.0226	0.0296	0.0326
20NN	0.8011	0.4001	0.3348	0.0224	0.0292	0.0321	0.0226	0.0295	0.0326
					cut point $= (5, 7)$				
True	0.8996	0.2345	0.3349						
1NN	0.8997	0.2352	0.3351	0.0186	0.0274	0.0340	0.0181	0.0264	0.0336
3NN	0.8994	0.2351	0.3351	0.0175	0.0258	0.0328	0.0175	0.0257	0.0328
5NN	0.8992	0.2351	0.3351	0.0172	0.0254	0.0325	0.0173	0.0255	0.0327
10NN	0.8989	0.2351	0.3350	0.0170	0.0250	0.0323	0.0173	0.0254	0.0326
20NN	0.8983	0.2351	0.3348	0.0168	0.0247	0.0321	0.0172	0.0253	0.0326

Carlo means and standard deviations (across 5000 replications) for the estimators of the true class fractions, TCF_1, TCF_2 and TCF_3. The table also gives the means of the estimated standard deviations (of the estimators), based on the asymptotic theory. The table clearly shows limitations of the (partially) parametric approaches in case of misspecified models for $\Pr(D_k = 1|T, A)$ and $\Pr(V = 1|T, A)$. More precisely, in term of bias, the FI, MSI, IPW and SPE approaches perform almost always poorly, with high distortion in mostly all cases. As we mentioned in Section 3.1.4, the SPE estimators could fall outside the interval $(0, 1)$. In our simulations, in the worst case, the estimator $\widehat{\text{TCF}}_{3,\text{SPE}}(-1.0, -0.5)$ gives rise to 20% of the values greater than 1. Moreover, the Monte Carlo standard deviations shown in the table indicate that the SPE approach might yield unstable estimates. Finally, the misspecification also has a clear effect on the estimated standard deviations of the estimators. On the other side, the estimators 1NN and 3NN seem to perform well in terms of both bias and standard deviation. In fact, KNN estimators yield estimated values that are near to the true values. In addition, we observe that the estimator 3NN has larger bias than 1NN, but with slightly less variance.

Finally, some results of simulation experiments performed to explore the effect of a multidimensional vector of auxiliary covariates are given. In particular, we consider $A = (A_1, A_2, A_3)^\top$. The data are generated in a similar way as in Study 1 of Section 3.3.1. More precisely, the disease status $\mathcal{D}$ is a trinomial random vector (D_1, D_2, D_3), such that D_k is a Bernoulli random variable with success probability θ_k, $k = 1, 2, 3$. We set $\theta_1 = 0.4, \theta_2 = 0.35$ and $\theta_3 = 0.25$. The continuous test results T and covariates A_1, A_2, A_3 are generated by the following conditional models

$$T, A_1, A_2, A_3|D_k \sim \mathcal{N}_4\left(\mu_k, \Sigma\right), \qquad k = 1, 2, 3,$$

where $\mu_k = (2k, k, 1.5k, 0.5k)^\top$ and

$$\Sigma = \begin{pmatrix} 1.75 & 0.1 & -0.2 & 0.5 \\ 0.10 & 2.5 & 0.5 & -0.3 \\ -0.20 & 0.5 & 1.0 & 0.7 \\ 0.50 & -0.3 & 0.7 & 1.2 \end{pmatrix}.$$

The verification status V is generated by the model

$$\text{logit}\left\{\Pr(V = 1|T, A_1, A_2, A_3)\right\} = -0.7 - 0.35T + 0.2A_1 + 0.8A_2 - 0.6A_3.$$

This choice corresponds to a verification rate of about 0.51. We consider six pairs of cut points (c_1, c_2), i.e., $(2, 4), (2, 5), (2, 7), (4, 5), (4, 7)$ and $(5, 7)$. For the (partially) parametric estimators FI, MSI, IPW and SPE, disease probabilities and verification probabilities are estimated by using correctly specified models. For the KNN estimators, the Mahalanobis distance is employed, because the variability of T and covariates A_1, A_2, A_3 is relatively large. In addition, we use $\bar{K} = 2$ for the estimation of standard deviations. The number of replicates in each simulation experiment is 5000. The sample size is 500.

As expected, results given in Table 3.25, show a certain loss of efficiency of KNN estimators, compared to parametric competitors.

3.4 Real data examples

To illustrate the application of the proposed methods, in this section we consider two quite distinct real data examples, both dealing with epithelial ovarian cancer (EOC). In the first illustration, we consider diagnosis of EOC in one of three classes i.e., benign disease, early stage and late stage

Table 3.24: Simulation results in case where both models for $\rho_k(t,a)$ and $\pi(t,a)$ are misspecified and sample size equals to 1000. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
				cut point $= (-1.0, -0.5)$					
True	0.1812	0.1070	0.9817						
FI	**0.1290**	**0.0588**	0.9888	0.0153	0.0133	0.0118	0.0154	0.0087	0.0412
MSI	**0.1299**	**0.0592**	0.9895	0.0154	0.0153	0.0131	0.0157	0.0110	0.0417
IPW	**0.1231**	**0.0576**	0.9889	0.0178	0.0211	0.0208	0.0175	0.0207	0.3694
SPE	**0.1407**	**0.0649**	0.9877	0.0173	0.0216	0.0231	0.0176	0.0212	0.0432
1NN	0.1809	0.1036	0.9817	0.0224	0.0304	0.0255	0.0211	0.0262	0.0242
3NN	0.1795	0.0991	0.9814	0.0214	0.0258	0.0197	0.0208	0.0244	0.0240
				cut point $= (-1.0, 0.7)$					
True	0.1812	0.8609	0.4469						
FI	**0.1290**	**0.7399**	**0.5850**	0.0153	0.0447	0.1002	0.0154	0.0181	0.0739
MSI	**0.1299**	**0.7423**	**0.5841**	0.0154	0.0453	0.1008	0.0157	0.0188	0.0666
IPW	**0.1231**	**0.7690**	**0.5004**	0.0178	0.0902	0.2049	0.0175	0.0844	0.2018
SPE	**0.1407**	**0.7635**	**0.5350**	0.0173	0.0702	**0.2682**	0.0176	0.0668	**2.0344**
1NN	0.1809	0.8452	0.4406	0.0224	0.0622	0.1114	0.0211	0.0544	0.1079
3NN	0.1795	0.8285	0.4339	0.0214	0.0521	0.0882	0.0208	0.0516	0.1066
				cut point $= (-1.0, 1.3)$					
True	0.1812	0.9732	0.1171						
FI	**0.1290**	**0.9499**	**0.1900**	0.0153	0.0179	0.0550	0.0154	0.0133	0.0422
MSI	**0.1299**	**0.9516**	**0.1902**	0.0154	0.0184	0.0552	0.0157	0.0142	0.0389
IPW	**0.1231**	0.9645	**0.1294**	0.0178	0.0519	0.1795	0.0175	0.0466	0.1344
SPE	**0.1407**	**0.9567**	**0.1760**	0.0173	0.0425	**0.3383**	0.0176	0.0402	**3.4770**
1NN	0.1809	0.9656	0.1124	0.0224	0.0218	0.0448	0.0211	0.0317	0.0710
3NN	0.1795	0.9604	0.1086	0.0214	0.0172	0.0338	0.0208	0.0305	0.0716
				cut point $= (-0.5, 0.7)$					
True	0.4796	0.7539	0.4469						
FI	**0.3715**	**0.6811**	**0.5850**	0.0270	0.0400	0.1002	0.0151	0.0145	0.0739
MSI	**0.3723**	**0.6831**	**0.5841**	0.0271	0.0409	0.1008	0.0162	0.0172	0.0666
IPW	**0.3547**	**0.7114**	**0.5004**	0.0325	0.0883	0.2049	0.0322	0.0831	0.2018
SPE	**0.3949**	**0.6986**	**0.5350**	0.0318	0.0687	**0.2682**	0.0331	0.0657	**2.0344**
1NN	0.4783	0.7416	0.4406	0.0361	0.0610	0.1114	0.0311	0.0551	0.1079
3NN	0.4756	0.7294	0.4339	0.0341	0.0499	0.0882	0.0304	0.0523	0.1066
				cut point $= (-0.5, 1.3)$					
True	0.4796	0.8661	0.1171						
FI	**0.3715**	**0.8910**	**0.1900**	0.0270	0.0202	0.0550	0.0151	0.0142	0.0422
MSI	**0.3723**	**0.8924**	**0.1902**	0.0271	0.0211	0.0552	0.0162	0.0165	0.0389
IPW	**0.3547**	**0.9068**	**0.1294**	0.0325	0.0535	0.1795	0.0322	0.0492	0.1344
SPE	**0.3949**	**0.8918**	**0.1760**	0.0318	0.0451	**0.3383**	0.0331	0.0435	**3.4770**
1NN	0.4783	0.8620	0.1124	0.0361	0.0349	0.0448	0.0311	0.0390	0.0710
3NN	0.4756	0.8613	0.1086	0.0341	0.0285	0.0338	0.0304	0.0371	0.0716
				cut point $= (0.7, 1.3)$					
True	0.9836	0.1122	0.1171						
FI	0.9618	**0.2099**	**0.1900**	0.0122	0.0317	0.0550	0.0043	0.0132	0.0422
MSI	0.9613	**0.2093**	**0.1902**	0.0125	0.0320	0.0552	0.0048	0.0135	0.0389
IPW	0.9548	**0.1955**	**0.1294**	0.0339	0.0831	0.1795	0.0323	0.0784	0.1344
SPE	0.9582	**0.1932**	**0.1760**	0.0332	0.0618	**0.3383**	0.0320	0.0605	**3.4770**
1NN	0.9821	0.1204	0.1124	0.0144	0.0494	0.0448	0.0133	0.0487	0.0710
3NN	0.9804	0.1319	0.1086	0.0138	0.0404	0.0338	0.0131	0.0464	0.0716

Table 3.25: Simulation results in case dimension of covariate A is 3. KNN estimators are based in the Mahalanobis distance."True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut point $= (2, 4)$				
True	0.5000	0.4347	0.9347						
FI	0.5001	0.4361	0.9344	0.0369	0.0383	0.0222	0.0310	0.0333	0.0507
MSI	0.5000	0.4358	0.9344	0.0370	0.0387	0.0227	0.0310	0.0335	0.0508
IPW	0.5017	0.4370	0.9340	0.0625	0.0566	0.0247	0.0600	0.0557	0.0297
SPE	0.4999	0.4355	0.9344	0.0372	0.0404	0.0230	0.0369	0.0403	0.0226
1NN	0.5009	0.4401	0.9304	0.0388	0.0418	0.0239	0.0393	0.0435	0.0262
3NN	0.5006	0.4420	0.9278	0.0384	0.0404	0.0232	0.0390	0.0425	0.0259
					cut point $= (2, 5)$				
True	0.5000	0.7099	0.7752						
FI	0.5001	0.7106	0.7756	0.0369	0.0358	0.0383	0.0310	0.0374	0.0504
MSI	0.5000	0.7101	0.7754	0.0370	0.0362	0.0385	0.0310	0.0376	0.0506
IPW	0.5017	0.7113	0.7739	0.0625	0.0627	0.0453	0.0600	0.0597	0.0528
SPE	0.4999	0.7096	0.7751	0.0372	0.0380	0.0392	0.0369	0.0375	0.0384
1NN	0.5009	0.7085	0.7671	0.0388	0.0401	0.0404	0.0393	0.0417	0.0421
3NN	0.5006	0.7060	0.7620	0.0384	0.0380	0.0394	0.0390	0.0406	0.0417
					cut point $= (2, 7)$				
True	0.5000	0.9230	0.2248						
FI	0.5001	0.9226	0.2246	0.0369	0.0200	0.0377	0.0310	0.0409	0.0275
MSI	0.5000	0.9225	0.2247	0.0370	0.0204	0.0377	0.0310	0.0410	0.0276
IPW	0.5017	0.9230	0.2216	0.0625	0.0318	0.0630	0.0600	0.0300	0.0622
SPE	0.4999	0.9226	0.2250	0.0372	0.0217	0.0388	0.0369	0.0220	0.0385
1NN	0.5009	0.9121	0.2155	0.0388	0.0233	0.0395	0.0393	0.0274	0.0402
3NN	0.5006	0.9054	0.2106	0.0384	0.0220	0.0383	0.0390	0.0270	0.0395
					cut point $= (4, 5)$				
True	0.9347	0.2752	0.7752						
FI	0.9347	0.2745	0.7756	0.0193	0.0327	0.0383	0.0096	0.0262	0.0504
MSI	0.9347	0.2743	0.7754	0.0194	0.0332	0.0385	0.0097	0.0264	0.0506
IPW	0.9373	0.2742	0.7739	0.0518	0.0550	0.0453	0.0460	0.0542	0.0528
SPE	0.9348	0.2742	0.7751	0.0223	0.0355	0.0392	0.0226	0.0364	0.0384
1NN	0.9299	0.2685	0.7671	0.0236	0.0347	0.0404	0.0266	0.0384	0.0421
3NN	0.9261	0.2640	0.7620	0.0214	0.0326	0.0394	0.0263	0.0375	0.0417
					cut point $= (4, 7)$				
True	0.9347	0.4883	0.2248						
FI	0.9347	0.4866	0.2246	0.0193	0.0380	0.0377	0.0096	0.0341	0.0275
MSI	0.9347	0.4867	0.2247	0.0194	0.0383	0.0377	0.0097	0.0343	0.0276
IPW	0.9373	0.4860	0.2216	0.0518	0.0589	0.0630	0.0460	0.0581	0.0622
SPE	0.9348	0.4871	0.2250	0.0223	0.0401	0.0388	0.0226	0.0407	0.0385
1NN	0.9299	0.4720	0.2155	0.0236	0.0416	0.0395	0.0266	0.0441	0.0402
3NN	0.9261	0.4633	0.2106	0.0214	0.0399	0.0383	0.0263	0.0430	0.0395
					cut point $= (5, 7)$				
True	0.9883	0.2132	0.2248						
FI	0.9879	0.2121	0.2246	0.0080	0.0317	0.0377	0.0036	0.0216	0.0275
MSI	0.9880	0.2124	0.2247	0.0080	0.0319	0.0377	0.0037	0.0218	0.0276
IPW	0.9893	0.2117	0.2216	0.0280	0.0599	0.0630	0.0236	0.0575	0.0622
SPE	0.9882	0.2130	0.2250	0.0112	0.0338	0.0388	0.0115	0.0340	0.0385
1NN	0.9831	0.2035	0.2155	0.0120	0.0348	0.0395	0.0162	0.0369	0.0402
3NN	0.9802	0.1994	0.2106	0.0105	0.0322	0.0383	0.0162	0.0358	0.0395

cancer on the basis of a well known tumor marker, i.e., CA125. We make use of a publicly available dataset in which the disease status is known for all subjects. Then, we simulate a verification process and apply our estimators. This allows to compare results obtained in the complete data case with those obtained in the incomplete data case after correcting for verification bias. In the second illustration, we focus on prediction of patients' response to chemotherapy, classified as sensitive, partially sensitive and resistant. Data are available for late stage EOC patients. In this second example, the response is missing for about 25% of the subjects involved in the study.

3.4.1 Diagnosis of EOC

We use data from the Pre-PLCO Phase II Dataset from the SPORE/ Early Detection Network/ Prostate, Lung, Colon, and Ovarian Cancer Ovarian Validation Study. The study protocol and data are publicly available at the address[1], along with descriptions of the study aims and analytic methods. In particular, we consider the following three classes of EOC, i.e., benign disease, early stage (I and II) and late stage (III and IV) cancer, and 2 of the 59 available biomarkers, i.e. CA125 and CA153, measured at Harvard laboratories. In detail, we use CA125 as the test T and CA153 as a covariate. Reasons for using CA153 as a covariate come from the medical literature that suggests that the concomitant measurement of CA153 with CA125 could be advantageous in the pre-operative discrimination of benign and malignant ovarian tumors. In addition, age of patients is also considered. Here, we have 134 patients with benign disease, 67 early stage samples and 77 late stage samples. After performing exploratory analysis, we have the mean of CA125 corresponding to three classes are 0.192, 1.810 and 3.214. Thus, it implies that the order of three classes is monotone ordering, i.e., Benign $<$ Early $<$ Late under the results of CA125 marker. The boxplots are depicted in Figure 3.1.

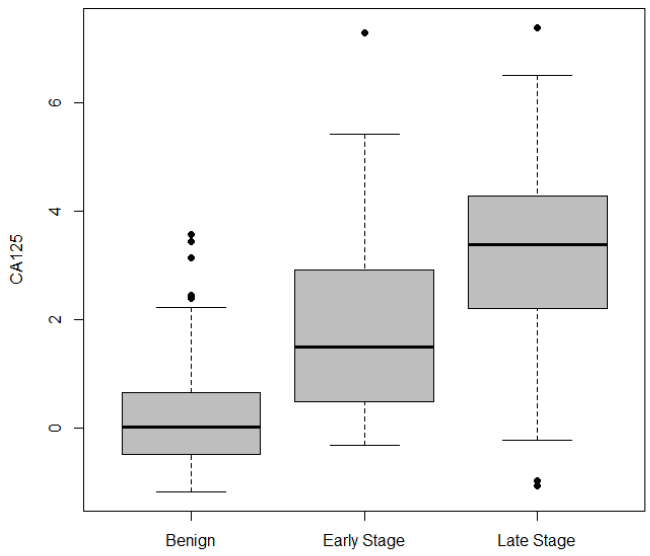


Figure 3.1: Boxplots of CA125 marker measurements for three classes under study of EOC.

[1] https://edrn.nci.nih.gov/protocols/119-spore-edrn-pre-plco-ovarian-phase-ii-validation

To mimic verification bias, a subset of the complete dataset is constructed using the test T and the vector $A = (A_1, A_2)^\top$ of the two covariates, namely the marker CA153 (A_1) and age (A_2). In this subset, T and A are known for all samples, but the true status (benign, early stage or late stage) is available only for some samples, that we select according to the following mechanism. We select all samples having a value for both T and A above their respective medians, i.e. 0.87 and (0.30, 45); as for the others, we apply the following selection process

$$\Pr(V = 1) = 0.05 + \delta_1 \mathrm{I}(T > 0.87) + \delta_2 \mathrm{I}(A_1 > 0.30) + \delta_3 \mathrm{I}(A_2 > 45),$$

with $\delta_1 = 0.35$, $\delta_2 = 0.25$ and $\delta_3 = 0.35$, leading to a marginal probability of selection equal to 0.634. With such a choice, the verification probability is equal to about 0.65 for subjects with $T > 0.87$, $A_1 > 0.30$ and $A_2 < 45$; 0.75 for subjects with $T > 0.87$, $A_1 < 0.30$ and $A_2 > 45$; 0.65 for subjects with $T < 0.87$, $A_1 > 0.30$ and $A_2 > 45$; 0.4 for subjects with $T > 0.87$, $A_1 < 0.30$ and $A_2 < 45$; 0.3 for subjects with $T < 0.87$, $A_1 > 0.30$ and $A_2 < 45$; 0.4 for subjects with $T < 0.87$, $A_1 < 0.30$ and $A_2 > 45$; 0.05 otherwise.

To apply FI, MSI and SPE estimators, we employ a multinomial logistic model to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_{1i}, A_{2i})$, where $D_k = 1$, $k = 1, 2, 3$ refers to benign, early and late, respectively. On the other hand, SPE and IPW methods require estimates of $\pi_i = \Pr(V_i = 1|T_i, A_{1i}, A_{2i})$. For estimating such quantities, we make use, firstly, of a correctly specified model, i.e., a linear threshold regression model and, then, of a misspecified model, i.e., a logistic model. For the KNN estimator, we use the Mahalanobis distance, since the test T and the covariates A_1, A_2 are heterogeneous with respect to their variances. Following discussion in Subsection 3.2.4, we use the selection rule (3.37) to find the size K of the neighborhood. This leads to the choice of $K = 1$ for our data. In addition, we also employ $K = 3$ for the sake of comparison with 1NN result.

The estimated ROC surfaces for the test T (CA125) obtained by applying the proposed methods are shown in Figure 3.3 and 3.4. For the sake of comparison, we also produced the ROC surface with full data (Full estimate), reported in Figure 3.2.

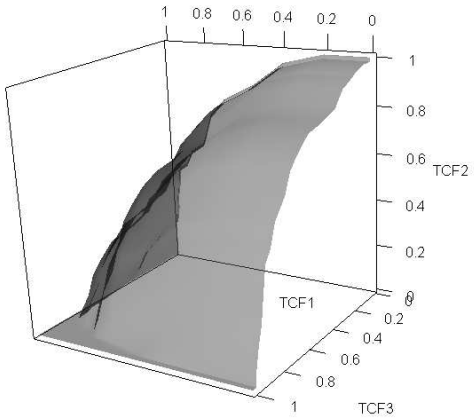
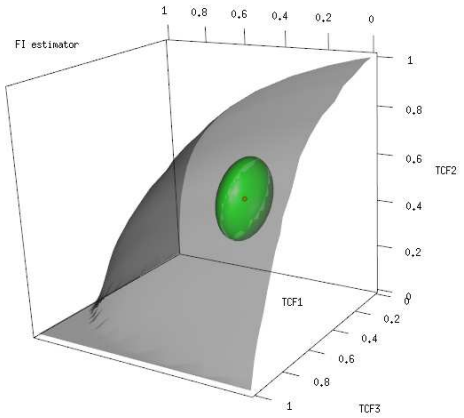
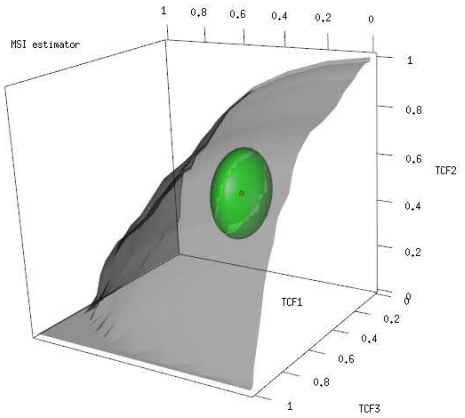


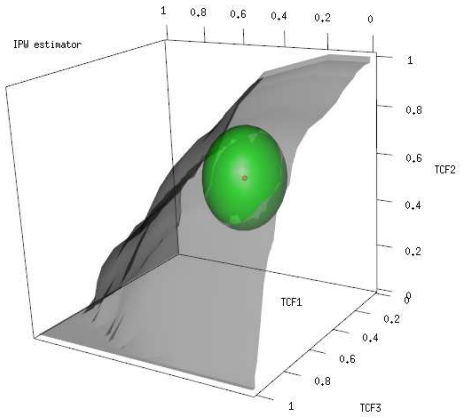
Figure 3.2: Estimated ROC surface for the CA125 marker, based on full data.



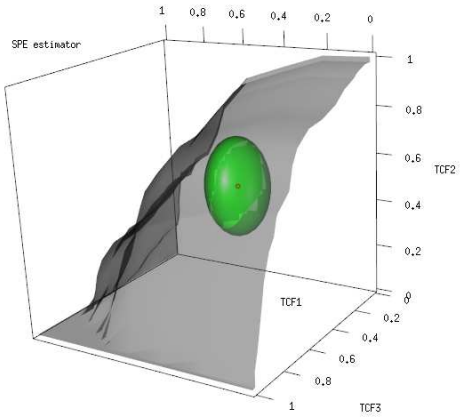
(a) FI



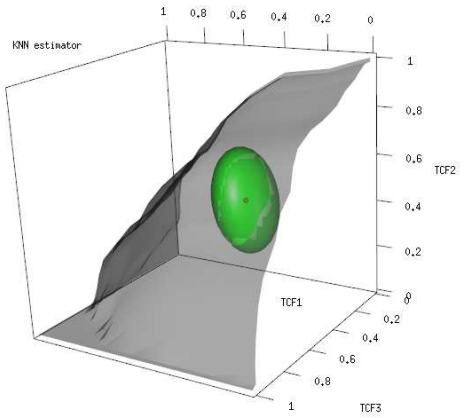
(b) MSI



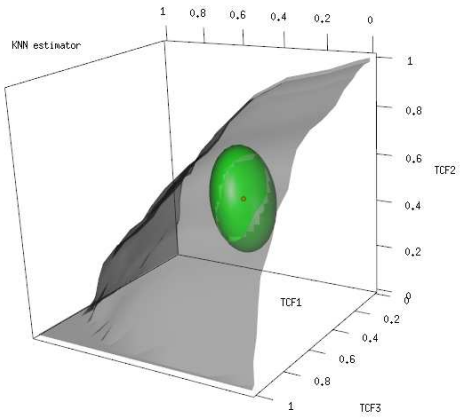
(c) IPW-threshold model



(d) SPE-threshold model

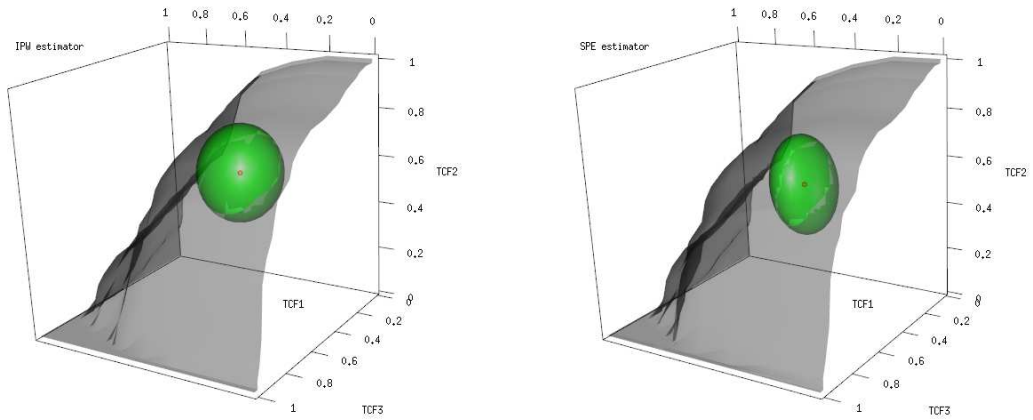


(e) 1NN



(f) 3NN

Figure 3.3: Bias–corrected estimated ROC surfaces for CA125 marker, based on incomplete data. The IPW and SPE estimators are obtained by using the threshold model.



(a) IPW-logit model

(b) SPE-logit model

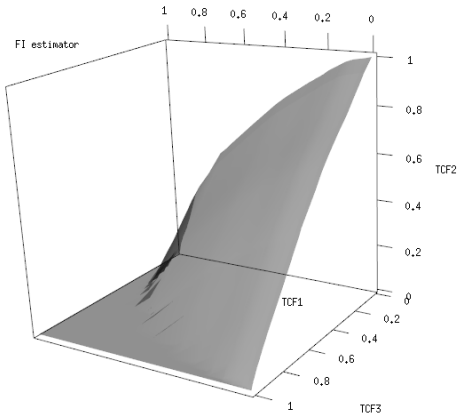
Figure 3.4: IPW and SPE estimated ROC surfaces for CA125 marker using the logistic regression model, based on incomplete data.

In Figures 3.3 and 3.4, we also give the 95% ellipsoidal confidence regions (green color) for true class fractions $(\mathrm{TCF}_1, \mathrm{TCF}_2, \mathrm{TCF}_3)$ at cut point $(-0.56, 2.31)$. These regions are built using the asymptotic normality of the estimators. Compared with the Full estimate, all the bias-corrected methods proposed in this chapter seem to behave well, yielding reasonable estimates of the ROC surface with incomplete data.

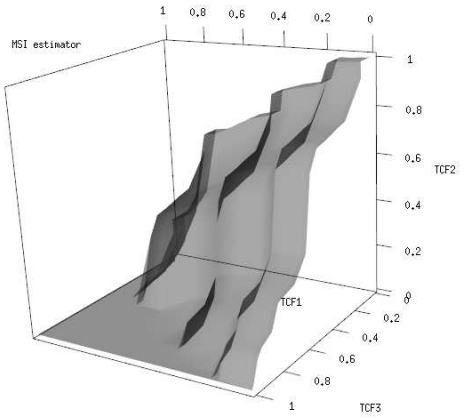
3.4.2 Prediction of response to chemotherapy

A major challenge in advanced-stage EOC is prediction of response to platinum chemotherapy on the basis of markers measured at molecular level. Indeed, several genomic profiling studies have shown that gene expressions relate with different aspects of ovarian cancer (tumor subtype, stage, grade, prognosis, and therapy resistance), although the measured association is usually very low. Here, we consider a cohort of 99 snap-frozen tumor biopsies taken from a frozen tissue bank, located at the Department of Oncology, IRCCS-Mario Negri Institute, Milan, Italy. Biopsies were collected from late stage (III and IV) cancer patients who underwent surgery at the Obstetrics and Gynaecology Department, San Gerardo Hospital, Monza, Italy between September 1992 and March 2010. For 75 of the 99 subjects, the three-class response to platinum therapy is available, being 31 patients sensitive, 11 partially sensitive and 33 resistant. For all the subjects, we consider as test predictive of the response to therapy the marker (T) resulting as a given linear combination of the logarithm of the expression levels of six genes, i.e., Entrez Gene ID: 23513, 7284, 128408, 56996, 2969, 6170. As a covariate, we consider age at onset of patients.

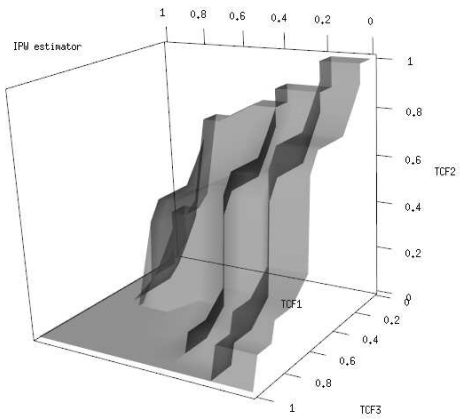
The estimated ROC surfaces for T obtained by applying the proposed methods are shown in Figure 3.5. FI, MSI, IPW and SPE estimators are based on the multinomial logistic model for the disease process and/or the logistic model for the verification process. KNN estimator is obtained by using $K = 1, 3$ and the Mahalanobis distance.



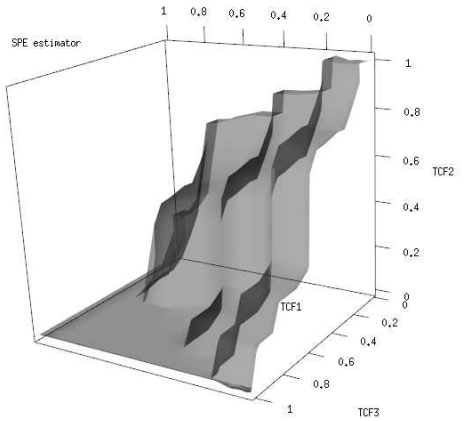
(a) FI



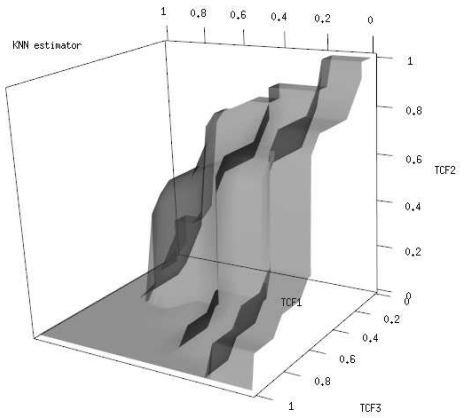
(b) MSI



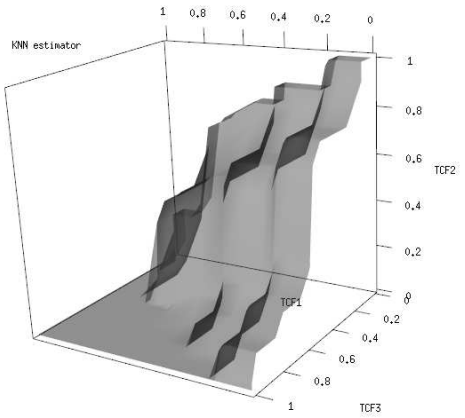
(c) IPW



(d) SPE



(e) 1NN



(f) 3NN

Figure 3.5: Bias–corrected estimated ROC surfaces for the test T predicting the response to therapy of late stage EOC patients.

3.5 Discussion

This chapter proposed five verification bias-corrected estimators of the ROC surface (and the VUS) of a continuous diagnostic test, namely, FI, MSI, IPW, SPE and KNN. The first four estimators, which can be considered an extension to the three-class case of estimators in Alonzo and Pepe (2005), are partially parametric in that they require the choice of a parametric model for the estimation of the disease process, or of the verification process, or of both processes. In some cases, wrong specifications of such models can visibly affect the produced estimates, as highlighted also by our results in the simulation studies. In fact, FI and MSI estimators are inconsistent if the model for the disease process is misspecified. On the contrary, the IPW estimator is inconsistent if the model for the verification process is incorrect. Thanks to the property of doubly robustness, inconsistency of SPE estimators occurs only if both models for the two processes are misspecified. A suitable solution for reducing the effects of model misspecification in statistical inference is to resort to fully nonparametric methods. That is reason why the KNN estimator is proposed. This approach is based on nearest-neighbor imputation and works under MAR assumption.

As in Adimari and Chiogna (2015, 2016), a simple extension of the KNN estimator, that could be used when categorical auxiliary variables are also available, is possible. Without loss of generality, we suppose that a single factor C, with m levels, is observed together with T and A. We also assume that C may be associated with both $\mathcal{D}$ and V. In this case, the sample can be divided into m strata, i.e., m groups of units sharing the same level of C. Then, for example, if the MAR assumption and first-order differentiability of the functions $\rho_k(t,a)$ and $\pi(t,a)$ hold in each stratum, a consistent and asymptotically normally distributed estimator of TCF_1 is

$$\widehat{\mathrm{TCF}}^{S}_{1,\mathrm{KNN}}(c_1) = \frac{1}{n}\sum_{j=1}^{m} n_j \widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1),$$

where n_j denotes the size of the j-th stratum and $\widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1)$ denotes the KNN estimator of the conditional TCF_1, i.e., the KNN estimator in (3.20) obtained from the patients in the j-th stratum. Of course, we must assume that, for every j, ratios n_j/n have finite and nonzero limits as n goes to infinity.

Verification bias occurs not only in the estimation of ROC surface, but also in VUS. Thus, finding the bias–corrected methods for estimation of VUS under the missing data is needed. It is worth noting that the proposed methods for the ROC surface can be valid for VUS, provided that the missingness mechanism is MAR. The details of this work will be presented in the next chapter.

Chapter 4

Estimation of the VUS in presence of verification bias

In this chapter, we develop various methods for estimating the VUS in presence of verification bias. These methods are particularly useful when the missing mechanism is MAR.

4.1 The parametric estimation scheme

4.1.1 Methods

We apply the four partially parametric estimators (FI, MSI, IPW, SPE) to estimate VUS in presence of verification bias. More precisely, FI, MSI, IPW and SPE estimators of VUS take the form

$$\hat{\mu}_* = \frac{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n} \mathrm{I}_{i\ell r}\hat{D}_{1i,*}\hat{D}_{2\ell,*}\hat{D}_{3r,*}}{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n} \hat{D}_{1i,*}\hat{D}_{2\ell,*}\hat{D}_{3r,*}},$$

where the star again stands for FI, MSI, IPW, SPE, and

$$\begin{aligned} \hat{D}_{ki,\mathrm{FI}} &= \hat{\rho}_{ki}, & \hat{D}_{ki,\mathrm{MSI}} &= V_i D_{ki} + (1-V_i)\hat{\rho}_{ki}, \\ \hat{D}_{ki,\mathrm{IPW}} &= V_i D_{ki}\hat{\pi}_i^{-1}, & \hat{D}_{ki,\mathrm{SPE}} &= V_i D_{ki}\hat{\pi}_i^{-1} - \hat{\rho}_{ki}(V_i\hat{\pi}_i^{-1}-1), \end{aligned}$$

for $k = 1, 2, 3$. It is worth noting that the bias–corrected VUS estimators are unbiased. The following remarks support the sentence.

Remark 4.1.1 (Expectation)**.** *We have*

$$\begin{aligned} \mathbb{E}(D_{1i}D_{2\ell}D_{3r}\mathrm{I}_{i\ell r}) &= \mathbb{E}_{T,A}\left\{\mathrm{I}_{i\ell r}\mathbb{E}(D_{1i}D_{2\ell}D_{3r}|T_i,A_i,T_\ell,A_\ell,T_r,A_r)\right\}, \\ &= \mathbb{E}_{T,A}\left\{\mathrm{I}_{i\ell r}\mathbb{E}(D_{1i}|T_i,A_i)\mathbb{E}(D_{2\ell}|T_\ell,A_\ell)\mathbb{E}(D_{3r}|T_r,A_r)\right\}, \\ &= \mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\mathrm{I}_{i\ell r}\right). \end{aligned}$$

The first identity follows because $\mathrm{I}_{i\ell r}$ *is function of* T_i, T_ℓ, T_r*, and the second identity follows because, the observed data are i.i.d. By analogy, we show that*

$$\mathbb{E}(D_{1i}D_{2\ell}D_{3r}) = \mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\right).$$

Therefore, we have that

$$\mu = \frac{\mathbb{E}\left(D_{1i}D_{2\ell}D_{3r}\mathrm{I}_{i\ell r}\right)}{\mathbb{E}\left(D_{1i}D_{2\ell}D_{3r}\right)} = \frac{\mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\mathrm{I}_{i\ell r}\right)}{\mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\right)}, \tag{4.1}$$

Remark 4.1.2 (Unbiasedness)**.**

- *FI estimators. Under the disease model, we have*

$$\begin{aligned}\mathbb{E}\left\{G_{i\ell r,\mathrm{FI}}(\mu_0,\tau_{\rho_0},\tau_\pi)\right\} &= \mathbb{E}\left\{\rho_{1i}(\tau_{\rho_0})\rho_{2r}(\tau_{\rho_0})\rho_{3\ell}(\tau_{\rho_0})(I_{i\ell r}-\mu_0)\right\}\\ &= \mathbb{E}\left\{\rho_{1i}\rho_{2r}\rho_{3\ell}(I_{i\ell r}-\mu_0)\right\}.\end{aligned}$$

  Since the relevant terms in the above expression depend on the test result T and covariates A, the Remark 4.1.1 can be applied. That is to say, $\mathbb{E}\left\{G_{i\ell r,\mathrm{FI}}(\mu_0,\tau_{\rho_0},\tau_\pi)\right\} = 0$ when the disease model holds.

- *MSI estimators. Under the disease model, we can verify that $\mathbb{E}\left\{D_{ki,\mathrm{MSI}}(\tau_{\rho_0})|T_i,A_i\right\} = \rho_{ki}$. In fact, we have that*

$$\begin{aligned}\mathbb{E}\left\{D_{ki,\mathrm{MSI}}(\tau_{\rho_0})|T_i,A_i\right\} &= \mathbb{E}\left\{V_iD_{ki}+(1-V_i)\rho_{ki}(\tau_{\rho_0})|T_i,A_i\right\}\\ &= \mathbb{E}\left[\mathbb{E}\left\{V_iD_{ki}+(1-V_i)\rho_{ki}(\tau_{\rho_0})|T_i,A_i,V_i\right\}|T_i,A_i\right]\\ &= \Pr(V_i=1|T_i,A_i)\mathbb{E}\left(D_{ki}|T_i,A_i\right)\\ &\quad+\Pr(V_i=0|T_i,A_i)\mathbb{E}\left(\rho_{ki}(\tau_{\rho_0})|T_i,A_i\right)\\ &= \Pr(V_i=1|T_i,A_i)\Pr(D_{ki}=1|T_i,A_i)\\ &\quad+\Pr(V_i=0|T_i,A_i)\Pr(D_{ki}=1|T_i,A_i)\\ &= \Pr(D_{ki}=1|T_i,A_i)=\rho_{ki}.\end{aligned}$$

  Therefore,

$$\mathbb{E}\left\{G_{i\ell r,\mathrm{MSI}}(\mu_0,\tau_{\rho_0},\tau_\pi)\right\} = \mathbb{E}\left\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r}-\mu_0)\right\}.$$

  And hence, the MSI–estimating function is unbiased under the disease model.

- *IPW estimators. The expression,*

$$\mathbb{E}\left\{\frac{V_i}{\pi_i(\tau_{\pi_0})}\middle|T_i,A_i\right\} = \frac{\mathbb{E}\left\{V_i|T_i,A_i\right\}}{\pi_i} = 1,$$

  is correct provided that the verification model holds. Therefore, it is not too difficult to prove that

$$\mathbb{E}\left\{\frac{V_iD_{ki}}{\pi_i(\tau_{\pi_0})}\middle|T_i,A_i\right\} = \rho_{ki}.$$

  Thus,

$$\begin{aligned}\mathbb{E}\left\{G_{i\ell r,\mathrm{IPW}}(\mu_0,\tau_\rho,\tau_{\pi_0})\right\} &= \mathbb{E}\left\{\frac{V_iV_\ell V_rD_{1i}D_{2\ell}D_{3r}}{\pi_i(\tau_{\pi_0})\pi_\ell(\tau_{\pi_0})\pi_k(\tau_{\pi_0})}\left(I_{i\ell r}-\mu_0\right)\right\}\\ &= \mathbb{E}\Big\{\left(I_{i\ell r}-\mu_0\right)\mathbb{E}(V_iD_{1i}\pi_i^{-1}(\tau_{\pi_0})|T_i,A_i)\mathbb{E}(V_\ell D_{2\ell}\pi_\ell^{-1}(\tau_{\pi_0})|T_\ell,A_\ell)\\ &\quad\times\mathbb{E}(V_rD_{3r}\pi_r^{-1}(\tau_{\pi_0})|T_r,A_r)\Big\}\\ &= \mathbb{E}\left\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r}-\mu_0)\right\}.\end{aligned}$$

- *SPE estimators. First, note that*

$$\begin{aligned}\mathbb{E}\left\{D_{ki,\mathrm{SPE}}(\tau_{\rho_0},\tau_\pi)|T_i,A_i\right\} &= \mathbb{E}\left\{\frac{V_i}{\pi_i(\tau_\pi)}\middle|T_i,A_i\right\}\left\{\mathbb{E}(D_{ki}|T_i,A_i)-\rho_{ki}(\tau_{\rho_0})\right\}+\rho_{ki}(\tau_{\rho_0})\\ &= \mathbb{E}\left\{\frac{V_i}{\pi_i(\tau_\pi)}\middle|T_i,A_i\right\}\left\{\mathbb{E}(D_{ki}|T_i,A_i)-\rho_{ki}\right\}+\rho_{ki}\\ &= \rho_{ki}.\end{aligned}$$

On the other hand, we also have

$$\begin{aligned}\mathbb{E}\left\{D_{ki,\mathrm{SPE}}(\tau_\rho,\tau_{\pi_0})|T_i,A_i\right\} &= \mathbb{E}\left\{\frac{V_i}{\pi_i(\tau_{\pi_0})}\middle|T_i,A_i\right\}\mathbb{E}(D_{ki}|T_i,A_i)\\ &\quad-\rho_{ki}(\tau_\rho)\mathbb{E}\left\{\frac{V_i}{\pi_i(\tau_{\pi_0})}-1\middle|T_i,A_i\right\}\\ &= \mathbb{E}(D_{ki}|T_i,A_i)-\rho_{ki}(\tau_\rho)\times 0\\ &= \rho_{ki}.\end{aligned}$$

Hence, we conclude that these imply

$$\mathbb{E}\left\{G_{i\ell r,\mathrm{SPE}}(\mu_0,\tau_{\rho_0},\tau_\pi)\right\}=\mathbb{E}\left\{G_{i\ell r,\mathrm{SPE}}(\mu_0,\tau_\rho,\tau_{\pi_0})\right\}=\mathbb{E}\left\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r}-\mu_0)\right\}=0,$$

under the disease or verification model. When both models hold, it is easy to show that $\mathbb{E}\left\{G_{i\ell r,\mathrm{SPE}}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\right\}$ *equals to* 0.

4.1.2 Asymptotic distribution

Recall that $\tau=(\tau_\rho^\top,\tau_\pi^\top)^\top$ is the vector of parameters of the models used to estimate $\rho=(\rho_1,\rho_2)^\top$, or π, or both. According to the formula of bias-corrected estimators of VUS, it is easy to realize that the proposed VUS estimators could be found as solution of appropriate estimating equations. The estimating functions of the VUS for FI, MSI, IPW and SPE estimators are denoted by

$$\begin{aligned}G_{i\ell r,\mathrm{FI}}(\mu,\tau_\rho,\tau_\pi) &= \rho_{1i}(\tau_\rho)\rho_{2\ell}(\tau_\rho)\rho_{3r}(\tau_\rho)\left(I_{i\ell r}-\mu\right),\\ G_{i\ell r,\mathrm{MSI}}(\mu,\tau_\rho,\tau_\pi) &= D_{1i,\mathrm{MSI}}(\tau_\rho)D_{2\ell,\mathrm{MSI}}(\tau_\rho)D_{3r,\mathrm{MSI}}(\tau_\rho)\left(I_{i\ell r}-\mu\right),\\ G_{i\ell r,\mathrm{IPW}}(\mu,\tau_\rho,\tau_\pi) &= \frac{V_iV_\ell V_r}{\pi_i(\tau_\pi)\pi_\ell(\tau_\pi)\pi_r(\tau_\pi)}D_{1i}D_{2\ell}D_{3r}\left(I_{i\ell r}-\mu\right),\\ G_{i\ell r,\mathrm{SPE}}(\mu,\tau_\rho,\tau_\pi) &= D_{1i,\mathrm{SPE}}(\tau_\rho,\tau_\pi)D_{2\ell,\mathrm{SPE}}(\tau_\rho,\tau_\pi)D_{3r,\mathrm{SPE}}(\tau_\rho,\tau_\pi)\left(I_{i\ell r}-\mu\right),\end{aligned}$$

and

$$\begin{aligned}D_{ki,\mathrm{MSI}}(\tau_\rho) &= V_iD_{ki}+(1-V_i)\rho_{ki}(\tau_\rho),\\ D_{ki,\mathrm{SPE}}(\tau_\rho,\tau_\pi) &= V_iD_{ki}\pi_i^{-1}(\tau_\pi)-\rho_{ki}(\tau_\rho)(V_i\pi_i^{-1}(\tau_\pi)-1),\end{aligned}$$

for $k=1,2,3$. For simplicity, we denote these estimating functions with $G_{i\ell r,*}(\mu,\tau_\rho,\tau_\pi)$.

For studying consistency and asymptotic normality of the bias-corrected VUS estimators, we need the following assumptions. Note that the estimating functions $G_{i\ell r,*}(\mu,\tau_\rho,\tau_\pi)$ require estimates of τ_ρ and/or τ_π for the working, so the existence of unique solutions $\hat{\tau}_\rho$ and $\hat{\tau}_\pi$ is necessary. Let $G_i^\tau(\tau_\rho,\tau_\pi)=\left(g_i^{\tau_\rho}(\tau_\rho)^\top,g_i^{\tau_\pi}(\tau_\pi)^\top\right)^\top$, $(\tau_\rho^\top,\tau_\pi^\top)^\top\in\boldsymbol{\tau_\rho}\times\boldsymbol{\tau_\pi}$, where $\boldsymbol{\tau_\rho}$ and $\boldsymbol{\tau_\pi}$ are the parameter spaces of τ_ρ and τ_π. The regularity conditions are:

(R1) the parameter space $\boldsymbol{\tau_\rho}\times\boldsymbol{\tau_\pi}$ has finite dimension and is compact;

(R2) the true value $(\tau_{\rho_0}, \tau_{\pi_0})$ exists and is interior to the parameter space $\boldsymbol{\tau_\rho} \times \boldsymbol{\tau_\pi}$ such that $\mathbb{E}\{G_i^\tau(\tau_\rho, \tau_\pi)\} \neq 0$ if $(\tau_\rho, \tau_\pi) \neq (\tau_{\rho_0}, \tau_{\pi_0})$ and $\mathbb{E}\{G_i^\tau(\tau_{\rho_0}, \tau_{\pi_0})\} = 0$;

(R3) the variance of $G_i^\tau(\tau_{\rho_0}, \tau_{\pi_0})$ exists and is finite;

(R4) $\mathbb{E}\left\{\partial G_i^\tau(\tau_\rho, \tau_\pi)/\partial(\tau_\rho, \tau_\pi)^\top|_{(\tau_\rho,\tau_\pi)=(\tau_{\rho_0},\tau_{\pi_0})}\right\}$ exists and is invertible;

(R5) There exists a neighborhood $\mathcal{N}$ of $(\tau_{\rho_0}, \tau_{\pi_0})$ such that the quantities $\sup_{(\tau_\rho,\tau_\pi)\in\mathcal{N}} \|G_i^\tau(\tau_\rho, \tau_\pi)\|$, $\sup_{(\tau_\rho,\tau_\pi)\in\mathcal{N}} \|\partial G_i^\tau(\tau_\rho, \tau_\pi)/\partial(\tau_\rho, \tau_\pi)^\top\|$ and $\sup_{(\tau_\rho,\tau_\pi)\in\mathcal{N}} \|G_i^\tau(\tau_\rho, \tau_\pi)G_i^\tau(\tau_\rho, \tau_\pi)^\top\|$ have finite expected values, where $\|\mathbf{X}\| \equiv \sum_i \sum_j X_{ij}^2$.

In addition, we assume that the predictors of the disease and verification regression models are the sufficient smooth functions with the existence of the moment condition the first derivatives, so that the following conditions hold.

(C1) The U–process

$$U_{n,*}(\mu, \tau_\rho, \tau_\pi) = \sqrt{n}\,\{G_*(\mu, \tau_\rho, \tau_\pi) - e(\mu, \tau_\rho, \tau_\pi)\}$$

is stochastically equicontinuous, where

$$\begin{aligned} G_*(\mu, \tau_\rho, \tau_\pi) \;=\; & \frac{1}{6n(n-1)(n-2)} \sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{k=1 \\ r\neq\ell, r\neq i}}^{n} \Big\{ G_{i\ell r,*}(\mu, \tau_\rho, \tau_\pi) + G_{ir\ell,*}(\mu, \tau_\rho, \tau_\pi) \\ & + G_{\ell ir,*}(\mu, \tau_\rho, \tau_\pi) + G_{\ell ri,*}(\mu, \tau_\rho, \tau_\pi) + G_{ri\ell,*}(\mu, \tau_\rho, \tau_\pi) + G_{r\ell i,*}(\mu, \tau_\rho, \tau_\pi) \Big\} \end{aligned}$$

and

$$\begin{aligned} e(\mu, \tau_\rho, \tau_\pi) \;=\; & \frac{1}{6}\mathbb{E}\Big\{ G_{i\ell r,*}(\mu, \tau_\rho, \tau_\pi) + G_{ir\ell,*}(\mu, \tau_\rho, \tau_\pi) + G_{\ell ir,*}(\mu, \tau_\rho, \tau_\pi) \\ & + G_{\ell ri,*}(\mu, \tau_\rho, \tau_\pi) + G_{ri\ell,*}(\mu, \tau_\rho, \tau_\pi) + G_{r\ell i,*}(\mu, \tau_\rho, \tau_\pi) \Big\}. \end{aligned}$$

(C2) $e(\mu, \tau_\rho, \tau_\pi)$ is differentiable in $(\mu, \tau_\rho, \tau_\pi)$.

(C3) $G_*(\mu, \tau_\rho, \tau_\pi)$ and $\frac{\partial G_*(\mu,\tau_\rho,\tau_\pi)}{\partial(\tau_\rho,\tau_\pi)^\top}$ converge uniformly to $e(\mu, \tau_\rho, \tau_\pi)$ and $\frac{\partial e(\mu,\tau_\rho,\tau_\pi)}{\partial(\tau_\rho,\tau_\pi)^\top}$.

Theorem 4.1.3 (Consistency)**.** *Suppose that the regularity conditions (R1)–(R5) and (C1)–(C3) hold. If the disease model and/or verification model holds, then* $\hat{\mu}_* \xrightarrow{p} \mu_0$.

Proof. Let $(\tau_{\rho_0}, \tau_{\pi_0})$ be defined as in condition (R2). Note that the parameters τ_ρ and τ_π are estimated by using the classic estimating equations, thus the condition (R1)–(R5) certify that the estimators $\hat{\tau}_\rho$ and $\hat{\tau}_\pi$ are consistent (Newey and McFadden, 1994).

We can show that

$$\mathbb{E}\{G_{i\ell r,*}(\mu_0, \tau_{\rho_0}, \tau_\pi)\} = \mathbb{E}\{G_{i\ell r,*}(\mu_0, \tau_\rho, \tau_{\pi_0})\} = \mathbb{E}\{G_{i\ell r,*}(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\} = 0$$

(see Remark 4.1.2). Therefore, $e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) = 0$ if the disease model or verification model or both hold. By condition (C2) and the implicit function theorem, there exists a neighborhood of $(\tau_{\rho_0}, \tau_{\pi_0})$ in which it is uniquely defined a continuously differentiable function $m(\tau_\rho, \tau_\pi)$, such that $m(\tau_{\rho_0}, \tau_{\pi_0}) = \mu_0$ and $e_*(m(\tau_\rho, \tau_\pi), \tau_\rho, \tau_\pi) = 0$. In cause of the consistency of estimators $\hat{\tau}_\rho$ and $\hat{\tau}_\pi$, we have that $\tilde{\mu}_* = m(\hat{\tau}_\rho, \hat{\tau}_\pi) \xrightarrow{p} \mu_0$. On the other hand, $G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) = 0$ and condition (C3) implies that $e_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) \xrightarrow{p} 0$. Thus, $\hat{\mu}_* \xrightarrow{p} \tilde{\mu}_*$. This implies the consistency of the bias-corrected estimators $\hat{\mu}_*$. □

Theorem 4.1.4 (Asymptotic normality). *Suppose the conditions in Theorem 4.1.3 are satisfied. If the disease model and/or verification model holds, then*

$$\sqrt{n}\,(\hat{\mu}_* - \mu_0) \xrightarrow{d} \mathcal{N}(0, \Lambda_*), \tag{4.2}$$

where the star indicates FI, MSI, IPW, SPE; and Λ_* is a suitable asymptotic variance.

Proof. We have

$$\begin{aligned} 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) \\ 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + \sqrt{n}e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - \sqrt{n}e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi). \end{aligned}$$

Since $e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) = 0$, so we get

$$\begin{aligned} 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + \sqrt{n}e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - \sqrt{n}e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + \sqrt{n}e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) - \sqrt{n}e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) \\ 0 &= \sqrt{n}\left\{G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\right\} + \sqrt{n}\left\{e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\right\} + \sqrt{n}e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) \\ &\quad - \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) + \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) \\ 0 &= \left[\sqrt{n}\left\{G_*(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\right\} - \sqrt{n}\left\{G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) - e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\right\}\right] \\ &\quad + \sqrt{n}\left\{e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\right\} + \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}). \end{aligned}$$

Condition (C1) implies that the first term in right hand side of the third identity equals to $o_p(1)$. Using the Mean–Value Theorem, we get

$$\begin{aligned} 0 &= o_p(1) + \sqrt{n}\left\{e(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) - e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\right\} + \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) \\ 0 &= o_p(1) + \frac{\partial e(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \mu}\sqrt{n}(\hat{\mu}_* - \mu_0) + \frac{\partial e^\top(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \tau_\rho}\sqrt{n}(\hat{\tau}_\rho - \tau_{\rho_0}) \\ &\quad + \frac{\partial e^\top(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \tau_\pi}\sqrt{n}(\hat{\tau}_\pi - \tau_{\pi_0}) + \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}), \end{aligned} \tag{4.3}$$

where $|\bar{\mu} - \mu_0| \le |\hat{\mu}_* - \mu_0|, |\bar{\tau}_\rho - \tau_{\rho_0}| \le |\hat{\tau}_\rho - \tau_{\rho_0}|$ and $|\bar{\tau}_\pi - \tau_{\pi_0}| \le |\hat{\tau}_\pi - \tau_{\pi_0}|$. It is straightforward to show that

$$\begin{aligned} \frac{\partial e(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \mu} &\to -\Pr(D_1 = 1)\Pr(D_2 = 1)\Pr(D_3 = 1) = -\theta_1\theta_2\theta_3, \\ \frac{\partial e^\top(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \tau_\rho} &\to \frac{\partial e^\top(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})}{\partial \tau_\rho}, \\ \frac{\partial e^\top(\bar{\mu}, \bar{\tau}_\rho, \bar{\tau}_\pi)}{\partial \tau_\pi} &\to \frac{\partial e^\top(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})}{\partial \tau_\pi}. \end{aligned}$$

By standard results on the limit distribution of U-statistics (van der Vaart, 2000, Theorem 12.3, Chap. 12),

$$\sqrt{n}\left\{G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) - e(\mu_0, \tau_{\rho_0}, \tau_{\pi_0})\right\} = \sqrt{n}G_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}) \xrightarrow{p} \sqrt{n}\tilde{G}_*(\mu_0, \tau_{\rho_0}, \tau_{\pi_0}),$$

where $\sqrt{n}\tilde{G}_*(\mu, \tau_\rho, \tau_\pi)$ is the projection of $U_{n,*}$ onto the set of all statistics of the form $\sum_{i=1}^{n} B_i(X_i)$ is given by

$$\begin{aligned} \sqrt{n}\tilde{G}_n(\mu, \tau_\rho, \tau_\pi) &= \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\Big\{G_{i\ell r,*}(\mu, \tau_\rho, \tau_\pi) + G_{ir\ell,*}(\mu, \tau_\rho, \tau_\pi) \\ &\quad + G_{\ell ir,*}(\mu, \tau_\rho, \tau_\pi) + G_{\ell ri,*}(\mu, \tau_\rho, \tau_\pi) + G_{ri\ell,*}(\mu, \tau_\rho, \tau_\pi) + G_{r\ell i,*}(\mu, \tau_\rho, \tau_\pi)\big|O_i\Big\} \end{aligned}$$

for $\ell \neq i$ and $r \neq \ell, r \neq i$, where $O_i = (\mathcal{D}_i^\top, V_i, T_i, A_i^\top)^\top$.

Under regularity conditions (R1)–(R5), we get

$$\sqrt{n}\,(\hat{\tau}_\rho - \tau_{\rho_0}) = -n^{-1/2}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\rho}(\tau_\rho)\right\}}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\tau_{\rho_0}}\right]^{-1}\sum_{i=1}^{n} g_i^{\tau_\rho}(\tau_{\rho_0}) + o_p(1) \tag{4.4}$$

and

$$\sqrt{n}\,(\hat{\tau}_\pi - \tau_{\pi_0}) = -n^{-1/2}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\pi}(\tau_\pi)\right\}}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\tau_{\pi_0}}\right]^{-1}\sum_{i=1}^{n} g_i^{\tau_\pi}(\tau_{\pi_0}) + o_p(1). \tag{4.5}$$

Applying these results to (4.3) gives

$$\begin{aligned}
\theta_1\theta_2\theta_3\sqrt{n}(\hat{\mu}_* - \mu) = {} & o_p(1) - \frac{1}{\sqrt{n}}\frac{\partial e^\top(\mu_0,\tau_{\rho_0},\tau_{\pi_0})}{\partial \tau_\rho}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\rho}(\tau_\rho)\right\}}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\tau_{\rho_0}}\right]^{-1}\sum_{i=1}^{n} g_i^{\tau_\rho}(\tau_{\rho_0}) \\
& - \frac{1}{\sqrt{n}}\frac{\partial e^\top(\mu_0,\tau_{\rho_0},\tau_{\pi_0})}{\partial \tau_\pi}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\pi}(\tau_\pi)\right\}}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\tau_{\pi_0}}\right]^{-1}\sum_{i=1}^{n} g_i^{\tau_\pi}(\tau_{\pi_0}) \\
& + \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\Big\{G_{i\ell r,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{ir\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{\ell ir,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) \\
& + G_{\ell ri,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{ri\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{r\ell i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\big|O_i\Big\} \\
= {} & o_p(1) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[-\frac{\partial e^\top(\mu_0,\tau_{\rho_0},\tau_{\pi_0})}{\partial \tau_\rho}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\rho}(\tau_\rho)\right\}}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\tau_{\rho_0}}\right]^{-1} g_i^{\tau_\rho}(\tau_{\rho_0})\right. \\
& - \frac{\partial e^\top(\mu_0,\tau_{\rho_0},\tau_{\pi_0})}{\partial \tau_\pi}\left[\left.\frac{\partial \mathbb{E}\left\{g_i^{\tau_\pi}(\tau_\pi)\right\}}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\tau_{\pi_0}}\right]^{-1} g_i^{\tau_\pi}(\tau_{\pi_0}) \\
& + \frac{1}{2}\mathbb{E}\Big\{G_{i\ell r,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{ir\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{\ell ir,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) \\
& \left. + G_{\ell ri,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{ri\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{r\ell i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\big|O_i\Big\}\right] \\
= {} & o_p(1) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Q_{i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) = o_p(1) + \frac{1}{\sqrt{n}}Q_*(\mu_0,\tau_{\rho_0},\tau_{\pi_0}).
\end{aligned}$$

Therefore, the asymptotic distribution of $\sqrt{n}(\hat{\mu}_* - \mu_0)$ can be determined by calculating the asymptotic distribution of $\frac{1}{\sqrt{n}}Q_*(\mu_0,\tau_{\rho_0},\tau_{\pi_0})$. Note that if the observed data are i.i.d, then $Q_{i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})$ also are i.i.d. By regularity condition (R2), we get $\mathbb{E}\{g_i^{\tau_\rho}(\tau_{\rho_0})\} = \mathbb{E}\{g_i^{\tau_\pi}(\tau_{\pi_0})\} = 0$. In addition, we easily show that

$$\begin{aligned}
0 \;=\; & \mathbb{E}\Big[\mathbb{E}\Big\{G_{i\ell r,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{ir\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{\ell ir,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{\ell ri,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) \\
& + G_{ri\ell,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) + G_{r\ell i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\big|O_i\Big\}\Big].
\end{aligned}$$

Therefore, $\mathbb{E}\{Q_{i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\} = 0$, and hence, $\frac{1}{\sqrt{n}}Q_*(\mu_0,\tau_{\rho_0},\tau_{\pi_0}) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}\mathrm{ar}\{Q_{i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\})$ by the Central Limit Theorem. Hence, $\sqrt{n}\,(\hat{\mu}_* - \mu_0) \xrightarrow{d} \mathcal{N}(0,\Lambda_*)$, where

$$\Lambda_* = \frac{\mathbb{V}\mathrm{ar}\{Q_{i,*}(\mu_0,\tau_{\rho_0},\tau_{\pi_0})\}}{\theta_1^2\theta_2^2\theta_3^2}.$$

□

4.1.3 Consistent variance estimator

Under condition (C3), a consistent estimator of Λ_* can be obtained by

$$\hat{\Lambda}_* = \frac{\mathbb{V}\mathrm{ar}\left\{\hat{Q}_{i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\right\}}{\hat{\theta}^2_{1,*}\hat{\theta}^2_{2,*}\hat{\theta}^2_{3,*}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\hat{Q}^2_{i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)}{\hat{\theta}^2_{1,*}\hat{\theta}^2_{2,*}\hat{\theta}^2_{3,*}},$$

where $\hat{\theta}_{k,*}$ are the proposed estimates of θ_k, $k = 1, 2, 3$ (see Section 3.1). Here,

$$\begin{aligned}
&\hat{Q}_{i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) \\
&= -\left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G^\top_{i\ell r,*}(\hat{\mu}_*, \tau_\rho, \hat{\tau}_\pi)}{\partial \tau_\rho}\right|_{\tau_\rho=\hat{\tau}_\rho}\right\}\left\{\sum_{i=1}^{n}\left.\frac{\partial g_i^{\tau_\rho}(\tau_\rho)}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\hat{\tau}_\rho}\right\}^{-1} g_i^{\tau_\rho}(\hat{\tau}_\rho) \\
&\quad - \left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G^\top_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \tau_\pi)}{\partial \tau_\pi}\right|_{\tau_\pi=\hat{\tau}_\pi}\right\}\left\{\sum_{i=1}^{n}\left.\frac{\partial g_i^{\tau_\pi}(\tau_\pi)}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\hat{\tau}_\pi}\right\}^{-1} g_i^{\tau_\pi}(\hat{\tau}_\pi) \\
&\quad + \frac{1}{2(n-1)(n-2)}\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\Big\{G_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{ir\ell,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{\ell ir,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) \\
&\quad + G_{\ell ri,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{ri\ell,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{r\ell i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\Big\}.
\end{aligned}$$

For fixed i, we show that

$$\begin{aligned}
\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\{G_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{ir\ell,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\} &= 2\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n} G_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi), \\
\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\{G_{\ell ir,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{ri\ell,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\} &= 2\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n} G_{\ell ir,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi), \\
\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\{G_{\ell ri,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{r\ell i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\} &= 2\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n} G_{r\ell i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\hat{Q}_{i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) \\
&= -\left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G^\top_{i\ell r,*}(\hat{\mu}_*, \tau_\rho, \hat{\tau}_\pi)}{\partial \tau_\rho}\right|_{\tau_\rho=\hat{\tau}_\rho}\right\}\left\{\sum_{i=1}^{n}\left.\frac{\partial g_i^{\tau_\rho}(\tau_\rho)}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\hat{\tau}_\rho}\right\}^{-1} g_i^{\tau_\rho}(\hat{\tau}_\rho) \\
&\quad - \left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=i\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G^\top_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \tau_\pi)}{\partial \tau_\pi}\right|_{\tau_\pi=\hat{\tau}_\pi}\right\}\left\{\sum_{i=1}^{n}\left.\frac{\partial g_i^{\tau_\pi}(\tau_\pi)}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\hat{\tau}_\pi}\right\}^{-1} g_i^{\tau_\pi}(\hat{\tau}_\pi) \\
&\quad + \frac{1}{(n-1)(n-2)}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{i\ell r,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{\ell ir,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi) + G_{r\ell i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)\Big\}.
\end{aligned}$$

Note that the quantity $\hat{Q}_{i,*}(\hat{\mu}_*, \hat{\tau}_\rho, \hat{\tau}_\pi)$ will not consist of the term of $g_i^{\tau_\rho}(\tau_\rho)$ or $g_i^{\tau_\pi}(\tau_\pi)$ if the estimating function $G_{i\ell r,*}(\mu, \tau_\rho, \tau_\pi)$ corresponds to FI and MSI estimators, or to the IPW approach.

The first derivatives $\left.\frac{\partial g_i^{\tau_\rho}(\tau_\rho)}{\partial \tau_\rho^\top}\right|_{\tau_\rho=\hat{\tau}_\rho}$ and $\left.\frac{\partial g_i^{\tau_\pi}(\tau_\pi)}{\partial \tau_\pi^\top}\right|_{\tau_\pi=\hat{\tau}_\pi}$ are obtained in (3.15) and (3.16) or (3.17). The explicit forms of the first partial derivatives of $G_{i\ell r,*}(\mu,\tau_\rho,\tau_\pi)$ with respect to the nuisance parameter τ_ρ are straightforwardly obtained by using product rule for derivatives and the expression (3.19). For the first partial derivatives of the estimating function $G_{i\ell r,*}(\mu,\tau_\rho,\tau_\pi)$ with respect to the nuisance parameter τ_π, we need to compute the derivatives of π_i^{-1}. In fact, π_i are obtained by a logistic or probit model, i.e.,

$$\pi_i = \frac{e^{U_i^\top \tau_\pi}}{1+e^{U_i^\top \tau_\pi}} \quad \text{or} \quad \pi_i = \Phi\left(U_i^\top \tau_\pi\right).$$

It is easy to show that

$$\frac{\partial}{\partial \tau_\pi}\pi_i^{-1} = -U_i\frac{1-\pi_i}{\pi_i} \quad \text{or} \quad \frac{\partial}{\partial \tau_\pi}\pi_i^{-1} = -U_i\frac{\phi\left(U_i^\top \tau_\pi\right)}{\Phi^2\left(U_i^\top \tau_\pi\right)}.$$

Here, $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal random variable, respectively.

4.2 Simulation studies

4.2.1 Correctly specified models

The disease status $\mathcal{D}$ is generated by a trinomial random vector (D_1, D_2, D_3), such that D_k is a Bernoulli random variable with mean θ_k, $k = 1, 2, 3$. We set $\theta_1 = 0.4, \theta_2 = 0.35$ and $\theta_3 = 0.25$. The pairs T, A are generated from the following conditional models

$$T, A|D_k \sim \mathcal{N}_2\left(\mu_k, \Lambda\right), \qquad k = 1, 2, 3,$$

where $\mu_k = k(\mu_T, \mu_A)^\top$. We consider three values of Λ,

$$\begin{pmatrix} 1.2 & 1 \\ 1 & 1 \end{pmatrix}, \qquad \begin{pmatrix} 1.75 & 0.1 \\ 0.1 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 5.5 & 3 \\ 3 & 2.5 \end{pmatrix}.$$

The true VUS value is equal to 0.9472 for the first value of Λ and $(\mu_T, \mu_A) = (3, 2)$; is equal to 0.7175 for the second value of Λ and $(\mu_T, \mu_A) = (2, 1)$; is equal to 0.4778 for the third value of Λ and $(\mu_T, \mu_A) = (2, 1)$. We simulate the verification status V by using the following model

$$\text{logit}\left\{\Pr(V = 1|T, A)\right\} = \delta_0 + \delta_1 T + \delta_2 A.$$

The parameters $(\delta_0, \delta_1, \delta_2)$ are fixed equal to $(1, -2.87, 4.06)$ when the first value of Λ is considered, and equal to $(1, -2.2, 4)$ otherwise. These choices give rise to a verification rate of about 0.52. Under our data–generating setting, the disease process follows a multinomial logistic model. We consider three sample sizes, i.e., $n = 200$ $n = 500$, and $n = 1000$. Each simulation experiment was based on 1000 replications.

FI, MSI, IPW and SPE estimates of VUS are computed under correct working models for both the disease and the verification processes. Table 4.1 shows Monte Carlo means, Monte Carlo standard deviations (MC.sd), the square roots of the variances estimated via asymptotic results (Asy.sd) and bootstrap standard deviations (Boot.sd) of $\hat{\mu}$.

We observe that the proposed estimators perform well in almost all cases. In fact, the FI and MSI approach always have small bias (magnitude $< 0.3\%$), even when the sample size is 200.

Table 4.1: Simulation results for bias–corrected estimators of VUS w.r.t parametric approaches.

	Sample size	Estimator	Mean	Bias(%)	MC.sd	Asy.sd	Boot.sd
Case I: TRUE = 0.9472	$n = 200$	FI	0.9471	−0.01	0.0251	0.0219	0.0256
		MSI	0.9466	−0.06	0.0252	0.0222	0.0258
		IPW	0.9498	0.27	0.0377	0.0261	0.0271
		SPE	0.9461	−0.11	0.0323	0.0274	0.0315
	$n = 500$	FI	0.9470	−0.02	0.0144	0.0143	0.0149
		MSI	0.9468	−0.04	0.0144	0.0144	0.0150
		IPW	0.9480	0.09	0.0244	0.0192	0.0192
		SPE	0.9467	−0.05	0.0228	0.0181	0.0224
	$n = 1000$	FI	0.9472	0.00	0.0101	0.0107	0.0115
		MSI	0.9473	0.01	0.0101	0.0109	0.0118
		IPW	0.9475	0.03	0.0190	0.0182	0.0185
		SPE	0.9472	0.00	0.0176	0.0172	0.0175
Case II: TRUE = 0.7175	$n = 200$	FI	0.7185	0.14	0.0549	0.0559	0.0566
		MSI	0.7165	−0.14	0.0552	0.0571	0.0577
		IPW	0.7261	1.20	0.0981	0.1197	0.0754
		SPE	0.7155	−0.28	0.1021	0.0981	0.1106
	$n = 500$	FI	0.7183	0.11	0.0357	0.0356	0.0357
		MSI	0.7176	0.01	0.0358	0.0360	0.0361
		IPW	0.7272	1.35	0.0814	0.0549	0.0564
		SPE	0.7184	0.12	0.0813	0.0698	0.0864
	$n = 1000$	FI	0.7178	0.05	0.0259	0.0255	0.0252
		MSI	0.7175	0.00	0.0259	0.0257	0.0256
		IPW	0.7192	0.24	0.0796	0.0682	0.0667
		SPE	0.7178	0.05	0.0723	0.0634	0.0715
Case III: TRUE = 0.4778	$n = 200$	FI	0.4788	0.21	0.0575	0.0558	0.0574
		MSI	0.4775	−0.06	0.0584	0.0576	0.0589
		IPW	0.4760	−0.38	0.1054	0.0767	0.0876
		SPE	0.4815	0.77	0.1121	0.1472	0.1418
	$n = 500$	FI	0.4782	0.08	0.0360	0.0350	0.0354
		MSI	0.4779	0.02	0.0364	0.0358	0.0361
		IPW	0.4804	0.54	0.0792	0.0608	0.0640
		SPE	0.4868	1.88	0.0943	0.1101	0.0995
	$n = 1000$	FI	0.4780	0.04	0.0246	0.0241	0.0243
		MSI	0.4776	−0.04	0.0253	0.0255	0.0260
		IPW	0.4781	0.07	0.0615	0.0587	0.0590
		SPE	0.4785	0.14	0.0810	0.0782	0.0796

Meanwhile, a large biases (greater than 1%) occurs for the IPW and SPE estimators in case II and III, when the sample size is 200 and 500. The Monte Carlo mean of the all bias–corrected estimators (FI, MSI, IPW and SPE) of VUS become closer to the true value as the sample size equals 1000, which is not surprising.

Table 4.1 indicates that the asymptotic variance procedure have a good performance in the sense that the standard deviations obtained by this procedure are comparable with those obtained from Monte Carlo experiments and bootstrap resampling process. Comparing the four proposed estimators of the VUS, FI and MSI are generally more efficient than IPW and SPE.

4.2.2 Model misspecification

In the previous part, the performance of the partially parametric bias–corrected estimators are investigated in the setting of correct model specification. Here, we study the behaviors of the proposed approaches in finite samples under incorrect specification of the disease and/or verification model.

In our simulations, we consider the data generated with respect to the first value of Λ in the previous section. The bias–corrected estimators of VUS are obtained under the following three settings for the working models:

(i) Wrong π: The model for verification process is

$$\text{logit}\left\{\Pr(V=1|T,A)\right\} = \tau_{\pi_1} + \tau_{\pi_2} T^{2/3} + \tau_{\pi_3} \log|A|.$$

(ii) Wrong ρ: The disease model is fitted with T and A^3 as predictor.

(iii) Wrong π and ρ: Both working models are defined in (i) and (ii).

Table 4.2: Simulation results correspond to model misspecification. The true VUS is 0.9472.

	Estimator	Monte Carlo Mean			Relative Bias (%)		
		200	500	1000	200	500	1000
Wrong π	FI	0.9471	0.9470	0.9472	−0.01	−0.02	0.00
	MSI	0.9466	0.9468	0.9473	−0.06	−0.04	0.01
	IPW	0.9699	0.9732	0.9770	2.40	2.75	3.15
	SPE	0.9504	0.9482	0.9474	0.33	0.11	0.02
Wrong ρ	FI	0.9629	0.9609	0.9606	1.66	1.45	1.41
	MSI	0.9614	0.9600	0.9594	1.50	1.31	1.28
	IPW	0.9498	0.9480	0.9475	0.28	0.09	0.03
	SPE	0.9504	0.9479	0.9471	0.33	0.07	−0.01
Wrong π and ρ	FI	0.9629	0.9609	0.9606	1.66	1.45	1.41
	MSI	0.9614	0.9600	0.9594	1.50	1.31	1.28
	IPW	0.9699	0.9732	0.9770	2.40	2.75	3.15
	SPE	0.9602	0.9750	0.9795	1.37	2.93	3.41

Table 4.2 shows the Monte Carlo mean across 1000 realizations as well as the relative bias (%) corresponding to three values of sample size, 200, 500 and 1000. When only the verification model is incorrect, IPW leads to serious biases (magnitude of relative bias: 2.40%, 2.75%, 3.15%) even if the sample size is increasing. This is reasonable, because the estimated verification probabilities are no longer valid. When only the disease model is incorrect, FI and MSI estimators have bias greater than 1.2% because the estimated disease probabilities are no longer valid.

The simulation results show that the SPE estimators behaves well in all of the first two scenarios in which either the disease or the verification process were correctly modeled, due to its doubly robustness property. When both working models were incorrectly specified, SPE leads to biased results (magnitude of relative bias ranging from 1.37% to 3.41%).

4.3 Others bias-corrected methods for estimating the VUS

4.3.1 Numerical method

Let $F_k(\cdot)$ be denote cumulative distribution functions of T for subject corresponding to $D_k = 1$, with $k = 1, 2, 3$. The expressions of TCFs can be rewritten as $\text{TCF}_1(c_1) = F_1(c_1)$, $\text{TCF}_2(c_1, c_2) = F_2(c_2) - F_2(c_1)$ and $\text{TCF}_3(c_2) = 1 - F_3(c_2)$, with $c_1 < c_2$. Using these notations, Nakas and Yiannoutsos (2004); Nakas (2014) wrote the functional form of the ROC surface as

$$\text{ROC}_s\left(\text{TCF}_1(c_1), \text{TCF}_3(c_2)\right) = F_2\left(F_3^{-1}\left(1 - \text{TCF}_3(c_2)\right)\right) - F_2\left(F_1^{-1}\left(\text{TCF}_1(c_1)\right)\right). \tag{4.6}$$

Visually, it is easy to see that (4.6) is $\text{TCF}_2(c_1, c_2)$. Based on (4.6), the VUS is determined by the following expression

$$\int_0^1 \int_0^1 \text{ROC}\left(p_1, p_3\right) \mathrm{d}p_3 \, \mathrm{d}p_1, \tag{4.7}$$

where $p_1 = \text{TCF}_1(c_1)$ and $p_3 = \text{TCF}_3(c_2)$. An estimate of (4.7) can be obtained by using Trapezoidal rule for a double integral.

For $m = 1, \ldots, M$, we denote with $(c_{1,m}, c_{2,m})$ m-th pairs of the cut points such that $c_{1,m}$ and $c_{2,m}$ are sorted in ascending order. As the computation in Chapter 3, the bias–corrected ROC surfaces are the 3D plot of

$$\left(\widehat{\text{TCF}}_{1,*}(c_{1,m}), \widehat{\text{TCF}}_{2,*}(c_{1,m}, c_{2,m}), \widehat{\text{TCF}}_{3,*}(c_{2,m})\right).$$

Here, the star $*$ stands for FI, MSI, IPW, SPE and KNN. Let $p_{1,m} = \widehat{\text{TCF}}_{1,*}(c_{1,m})$, $p_{3,m} = \widehat{\text{TCF}}_{3,*}(c_{2,m})$ and $p_{2,m,l} = \widehat{\text{ROC}}_s(p_{1,m}, p_{3,l}) = \widehat{\text{TCF}}_{2,*}(c_{1,m}, c_{2,m})$. Note that for the cut points $c_{2,m} < c_{2,m+1}$, so $p_{3,m} > p_{3,m+1}$. The trapezoidal rule yields

$$\widehat{\text{VUS}}_{\text{TR},*} = \sum_{k=1}^{M-1} \left(p_{1,k+1} - p_{1,k}\right) \sum_{l=1}^{M-1} \left(p_{3,l} - p_{3,l+1}\right) \frac{p_{2,k+1,l+1} + p_{2,k+1,l} + p_{2,k,l+1} + p_{2,k,l}}{4}. \tag{4.8}$$

This estimator comes from numerical estimation, and, therefore, there is a limitation. The statistical properties of the estimator are unavailable, such as, for example, the closed–form of variance and the asymptotic distribution. For example, we have to use a bootstrap resampling method for constructing confidence intervals.

4.3.2 Nearest-neighbor imputation

In the previous section, the partially parametric estimators for correcting verification bias in estimation of the VUS are presented. The methods work well when the parametric models (disease and/or verification model) are correctly specified. To reduce the effects of model misspecification, we here employ the nearest-neighbor imputation to make a bias–corrected estimator for VUS.

Recall that $\rho_{ki} = \Pr(D_{ki} = 1 | T_i = 1, A_i = 1)$, $k = 1, 2, 3$, is the probability that the i-th patient has true disease status belongs to class k given the test results.

Given a finite positive integer K and a suitable distance measure (e.g. Euclidean, Manhattan, Lagrange and Mahalanobis), a nearest–neighbor imputation estimate of ρ_{ki}, for a subject with true disease status not verified, could be defined as

$$\hat{\rho}_{ki,K} = \frac{1}{K} \sum_{l=1}^{K} D_{ki(l)},$$

where $\{(T_{i(l)}, A_{i(l)}, D_{ki(l)}) : V_{i(l)} = 1, l = 1, \ldots, K\}$ is a set of K observed data pairs and $(T_{i(l)}, A_{i(l)})$ denotes the l-th nearest neighbor to (T_i, A_i) among all (T, A)'s corresponding to the verified patients, i.e., to those D_{kh}'s with $V_h = 1$. The disease status D_{ki} could therefore be replaced by $\hat{\rho}_{ki,K}$ on non–verified units. Specifically, for any $i = 1, \ldots, n$, we define

$$\hat{D}_{ki,K} = V_i D_{ki} + (1 - V_i)\hat{\rho}_{ki,K}.$$

The proposed nonparametric verification bias–corrected VUS estimator is

$$\hat{\mu}_{\text{KNN}} = \frac{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \mathrm{I}_{i\ell r} \hat{D}_{1i,K} \hat{D}_{2\ell,K} \hat{D}_{3r,K}}{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \hat{D}_{1i,K} \hat{D}_{2\ell,K} \hat{D}_{3r,K}}, \qquad K \in \mathbb{N}. \tag{4.9}$$

Here, $I_{i\ell r}$ is indicator function of T_i, T_ℓ, T_r, which defined in Section 2.4.

Following the discussion in Section 3.2.4, the number of neighbors, K, and the distance measure play a key role in the KNN estimator. The selection of a suitable distance is typically dictated by features of the data (diagnostic test and covariates) and possible subjective evaluation. Similar with case of the ROC surface, a value of K around 3 could be adequate in case of low dimension of (T, A). On the other hand, one can employ the selection rule defined in (3.37) to find out a good choice for K when the dimension of (T, A) is large.

In order to obtain variance estimates, there are some possible approaches, which among the most popular method is bootstrap resampling. To apply this procedure for the KNN estimator, we implement the following steps. From the original observations $(D_{1i}, D_{2i}, D_{3i}, V_i, T_i, A_i)$, with $i = 1, \ldots, n$, consider B bootstrap samples $(D_{1i}^{*b}, D_{2i}^{*b}, D_{3i}^{*b}, V_i^{*b}, T_i^{*b}, A_i^{*b})$ with $b = 1, \ldots, B$. For the b-th sample, compute the bootstrap estimates $\hat{\rho}_{ki,K}^{*b}$, and hence, obtain $\hat{D}_{ki,K}^{*b} = V_i^{*b} D_{ki}^{*b} + (1 - V_i^{*b})\hat{\rho}_{ki,K}^{*b}$. After that, the bootstrap estimates of the standard deviation of $\hat{\mu}_{\text{KNN}}$ is

$$sd(\hat{\mu}_{\text{KNN}}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\mu}_{\text{KNN}}^{*b} - \hat{\mu}_{\text{KNN}}^{*}\right)^2},$$

where $\hat{\mu}_{\text{KNN}}^{*}$ is the mean of the B bootstrap estimates $\hat{\mu}_{\text{KNN}}^{*b}$,

$$\hat{\mu}_{\text{KNN}}^{*b} = \frac{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \mathrm{I}_{i\ell r}^{*b} \hat{D}_{1i,K}^{*b} \hat{D}_{2\ell,K}^{*b} \hat{D}_{3r,K}^{*b}}{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \hat{D}_{1i,K}^{*b} \hat{D}_{2\ell,K}^{*b} \hat{D}_{3r,K}^{*b}}, \qquad K \in \mathbb{N}.$$

Here, $\mathrm{I}_{i\ell r}^{*b}$ is the indicator function of $T_i^{*b}, T_\ell^{*b}, T_r^{*b}$, and defined in Section 2.4.

Based on the estimate of the standard deviation of the KNN estimator, a $(1 - \alpha)$ bootstrap confidence interval could be constructed, where α is a significant level. The simplest is the normal interval, which is defined as $\hat{\mu}_{\text{KNN}} \pm z_{1-\alpha/2} sd(\hat{\mu}_{\text{KNN}})$, where z_α denote the α-th quantile of a standard normal random variable. However, in case of small sample size, use of the normal approximation for the construction of confidence intervals may be inappropriate. In such situations, some other ways can be employed, e.g., pivotal interval, studentized pivotal interval or percentile interval.

4.4 Real data examples

In this section, we implement the proposed estimators for two real datasets, used also in Section 3.4. For the diagnosis of EOC, we use the same selection process for generating the missing mechanism. As for the ROC surface, here we use a multinomial logistic model to estimate the disease probabilities. For the SPE and IPW approaches, results for the correctly specified and misspecified model for the verification process are given, i.e., threshold and logistic regression model. We use the Mahalanobis distance and choose $K = 1, 3$ for the KNN estimator. The results are summarized in Table 4.3.

Table 4.3: Bias–corrected (and Full) estimated VUS for the marker CA125, assessing the classification into three classes of EOC: benign disease, early stage (I and II) and late stage (III and IV).

	VUS Estimate	Asy.sd	Boot.sd	95% C.I. (with Asy.sd)
Full	0.5663			
FI	0.5150	0.0404	0.0417	(0.4357, 0.5942)
MSI	0.5183	0.0415	0.0431	(0.4368, 0.5997)
IPW.logit	0.5500	0.0416	0.0471	(0.4685, 0.6314)
SPE.logit	0.5581	0.0443	0.0463	(0.4712, 0.6450)
IPW.thres	0.5353	0.0393	0.0457	(0.4583, 0.6123)
SPE.thres	0.5470	0.0440	0.0438	(0.4608, 0.6331)
1NN	0.5123	—	0.0471	(0.4199, 0.6046)
3NN	0.5104	—	0.0466	(0.4190, 0.6018)

In analogy, Table 4.4 shows the bias–corrected VUS estimates, along with the Naïve estimate. The table also gives the estimated standard deviations (via asymptotic theory), bootstrap standard deviations and approximated 95% confidence intervals. Despite the limited sample size, the results show that T has some ability to predict response to therapy for late stage EOC patients.

In two tables (Table 4.3 and 4.4), the bootstrap procedure is performed with 250 replications. In case of the KNN estimator, the 95% confidence intervals are constructed by using bootstrap standard deviation estimates.

Table 4.4: Bias–corrected (and Naïve) estimated VUS for the test T predicting the response to therapy of late stage EOC patients.

	VUS Estimate	Asy.sd	Boot.sd	95% C.I. (with Asy.sd)
Naïve	0.3452			
FI	0.3005	0.0512	0.0538	(0.2002, 0.4009)
MSI	0.3197	0.0629	0.0656	(0.1963, 0.4430)
IPW	0.3231	0.0654	0.0755	(0.1949, 0.4512)
SPE	0.3110	0.0675	0.0704	(0.1787, 0.4433)
1NN	0.3230	—	0.0699	(0.1859, 0.4600)
3NN	0.3052	—	0.0709	(0.1662, 0.4441)

4.5 Discussion

In this chapter, we have proposed several methods to correct for verification bias the VUS under missing at random assumption. By using imputation and re-weighting techniques, we have provided the FI, MSI, IPW and SPE estimators. These approaches are working well if the disease model or the verification model is corrected. However, in case of model misspecification, all four estimators yield biased results, and hence, the KNN estimator could be useful. An alternative way to estimate VUS in presence of verification bias is the use of numerical method to calculate the volume under bias–corrected ROC surface.

In real applications, doctors can employ more than one diagnostic test to identify the presence of a disease. For example, they can consider two diagnostic tests, coded as DT1 and DT2. In such situation, we will have two different ROC surfaces, and of course, two different VUSs, $\mu^{(1)}$ and $\mu^{(2)}$ say. An interesting question is "How to choose the best diagnostic test?". From a statistical point of view, a suitable answer comes from testing the significance of the VUS difference $\mu^{(1)} - \mu^{(2)}$. In fact, we consider two following hypotheses

$$\text{(I)} \begin{cases} H_0: & \mu^{(1)} - \mu^{(2)} = 0 \\ H_1: & \mu^{(1)} - \mu^{(2)} > 0 \end{cases} \qquad \text{or} \qquad \text{(II)} \begin{cases} H_0: & \mu^{(1)} - \mu^{(2)} = 0 \\ H_1: & \mu^{(1)} - \mu^{(2)} < 0 \end{cases}.$$

The first hypothesis identifies test DT1 is a best, whereas the second one implies that the diagnostic test DT2 is to be preferred. To perform the above tests, we can develop simple tests procedures.

Table 4.5: Decision rules for normal test.

The hypothesis	t–statistic	Wald statistic
(I)	$1 - \Phi\left(\frac{\sqrt{n}\hat{\Delta}_{\mu,*}}{\sqrt{\hat{\Lambda}'_*}}\right) < \alpha$	$\Pr\left(\chi^2 > \frac{n\hat{\Delta}^2_{\mu,*}}{\hat{\Lambda}'_*}\right) < \alpha$
(II)	$\Phi\left(\frac{\sqrt{n}\hat{\Delta}_{\mu,*}}{\sqrt{\hat{\Lambda}'_*}}\right) < \alpha$	$\Pr\left(\chi^2 < \frac{n\hat{\Delta}^2_{\mu,*}}{\hat{\Lambda}'_*}\right) < \alpha$

Let $T^{(1)}$ and $T^{(2)}$ denote test results of two diagnostic tests DT1 and DT2. The VUS difference is equal to

$$\Delta_\mu = \mu^{(1)} - \mu^{(2)} = \frac{\mathbb{E}\left\{\mathrm{I}^{(D)}_{ijk} D_{1i} D_{2j} D_{3k}\right\}}{\mathbb{E}\left\{D_{1i} D_{2j} D_{3k}\right\}}, \tag{4.10}$$

where $\mathrm{I}^{(D)}_{ijk} = \mathrm{I}^{(1)}_{ijk} - \mathrm{I}^{(2)}_{ijk}$. If we redefine $\mathbf{T} = \left(T^{(1)}, T^{(2)}\right)$ and replace I_{ijk} with $\mathrm{I}^{(D)}_{ijk}$, then estimation of Δ_μ is formally identical to estimation of a single VUS. As a corollary, the estimate $\hat{\Delta}_\mu$ inherits all of the properties of $\hat{\mu}_*$, i.e., $\hat{\Delta}_\mu$ is consistent and asymptotically normally distributed. In particular, we get $\sqrt{n}\left(\hat{\Delta}_{\mu,*} - \Delta_\mu\right) \xrightarrow{d} \mathcal{N}(0, \hat{\Lambda}'_*)$, where $\hat{\Lambda}'_*$ is obtained in the same way as $\hat{\Lambda}_*$. Under this result, the hypothesis testing is based on this asymptotic distribution. In such situation, two popular asymptotic test statistics are t–statistic and Wald statistic, specifically

$$t_{\Delta_\mu=0} = \frac{\sqrt{n}\hat{\Delta}_{\mu,*}}{\sqrt{\hat{\Lambda}'_*}}, \qquad \text{Wald} = \left(t_{\Delta_\mu=0}\right)^2 = \frac{n\hat{\Delta}^2_{\mu,*}}{\hat{\Lambda}'_*}.$$

One can verify that $t_{\Delta_\mu=0} \dot{\sim} \mathcal{N}(0,1)$ and $\text{Wald} = \left(t_{\Delta_\mu=0}\right)^2 \dot{\sim} \chi^2(1)$, under the null hypothesis H_0. Choosing a significance level $\alpha \in (0,1)$, we reject H_0 if one of the expressions in Table 4.5 is satisfied.

Alternatively, one can employ bootstrap hypothesis testing. We implement the following steps: (i) draw a bootstrap sample of n observations $(D_1^*, D_2^*, D_3^*, V^*, T^{(1)*}, T^{(2)*}, A^*)$ with replacement from the original sample; (ii) calculate the bias–corrected VUS of the $T^{(1)}$ and $T^{(2)}$, say, $\hat{\mu}_*^{(1)*}$ and $\hat{\mu}_*^{(2)*}$, and evaluate the difference $\Delta_{\mu,*}^* = \hat{\mu}_*^{(1)*} - \hat{\mu}_*^{(2)*}$; (iii) repeat step (i) and (ii) for B times and obtain B values of $\Delta_{\mu,*}^*$. One sided bootstrap p–value for the hypothesis (I) is then estimated as

$$p - \text{value} \approx \frac{\#\{\Delta_{\mu,*}^{*,b} > 0\}}{B},$$

with $b = 1, \ldots, B$. We reject H_0 if p–value is less then the significance level α. In analogy, we have the definition of one sided bootstrap p–value for the hypothesis (II). The bootstrap procedure may require a large B and perform well in case of uncorrelated diagnostic tests (Nakas and Yiannoutsos, 2004).

Also in real applications, the verification status could somehow depends, in addition to the test T and covariates, also on the true disease status, a possibility not foreseen by the MAR assumption. In such situations, we face a nonignorable missingness mechanism and the all proposed estimators can not be applied. This motivates to find the methods to deal with nonignorable verification bias in estimation of VUS. This work will be presented in next chapter.

Chapter 5

NI verification bias in estimation of the VUS

5.1 Model for NI missing data mechanism

5.1.1 Model settings

To deal with NI missing data mechanism, in what follows we extend parametric models adopted in Liu and Zhou (2010) for the two–class problem to the three–class case. More precisely, with three disease categories, we fix the model for the verification process as follows

$$\pi = \Pr(V = 1|D_1, D_2, T, A) = \frac{\exp\{h(T, A; \tau_\pi) + \lambda_1 D_1 + \lambda_2 D_2\}}{1 + \exp\{h(T, A; \tau_\pi) + \lambda_1 D_1 + \lambda_2 D_2\}}, \tag{5.1}$$

where D_1 and D_2 are defined in the previous section, $h(T, A; \tau_\pi)$ is an arbitrary working function, and τ_π is a set of parameters. Here, $\lambda = (\lambda_1, \lambda_2)^\top$ is the non-ignorable parameter: the missing data mechanism is MAR if $\lambda_1 = \lambda_2 = 0$; NI, otherwise. As for the disease model, we employ the multinomial logistic regression for the whole sample, i.e.,

$$\rho_k = \Pr(D_k = 1|T, A) = \frac{\exp\{f(T, A; \tau_{\rho_k})\}}{1 + \exp\{f(T, A; \tau_{\rho_1})\} + \exp\{f(T, A; \tau_{\rho_2})\}}, \tag{5.2}$$

where $f(T, A; \tau_{\rho_k})$ is an arbitrary working function, and τ_{ρ_k} is a set of parameters, for $k = 1, 2$. The parameters $\lambda, \tau_\pi, \tau_\rho$, with $\tau_\rho = (\tau_{\rho_1}^\top, \tau_{\rho_2}^\top)^\top$, can be estimated jointly by using a likelihood–based approach.

It is worth noting that, under (5.1), an application of Bayes' rule gives that

$$\begin{aligned}
\frac{\Pr(D_1 = 1|V = 1, T, A)}{\Pr(D_1 = 1|V = 0, T, A)} &= \frac{\Pr(V = 0|T, A)}{\Pr(V = 1|T, A)} \exp\{h(T, A; \tau_\pi) + \lambda_1\}, \\
\frac{\Pr(D_2 = 1|V = 1, T, A)}{\Pr(D_2 = 1|V = 0, T, A)} &= \frac{\Pr(V = 0|T, A)}{\Pr(V = 1|T, A)} \exp\{h(T, A; \tau_\pi) + \lambda_2\}, \\
\frac{\Pr(D_3 = 1|V = 1, T, A)}{\Pr(D_3 = 1|V = 0, T, A)} &= \frac{\Pr(V = 0|T, A)}{\Pr(V = 1|T, A)} \exp\{h(T, A; \tau_\pi)\}.
\end{aligned}$$

Therefore,

$$\frac{\Pr(D_1 = 1|V = 1, T, A)}{\Pr(D_1 = 1|V = 0, T, A)} \Big/ \frac{\Pr(D_3 = 1|V = 1, T, A)}{\Pr(D_3 = 1|V = 0, T, A)} = \exp(\lambda_1), \tag{5.3}$$

$$\frac{\Pr(D_2 = 1|V = 1, T, A)}{\Pr(D_2 = 1|V = 0, T, A)} \Big/ \frac{\Pr(D_3 = 1|V = 1, T, A)}{\Pr(D_3 = 1|V = 0, T, A)} = \exp(\lambda_2), \tag{5.4}$$

so that, according to (5.3) and (5.4), λ_1 and λ_2 can also be interpreted as log-odds ratios of belonging to class 1 (instead of class 3) and to class 2 (instead of class 3), respectively, for a verified subject compared to an unverified subject with the same test result and covariates.

5.1.2 Parameter estimation

As in Liu and Zhou (2010), in our model, for simplicity, we take $h(T, A; \tau_\pi) = \tau_{\pi_1} + \tau_{\pi_2} T + A^\top \tau_{\pi_3}$ and $f(T, A; \tau_{\rho_k}) = \tau_{\rho_{1k}} + \tau_{\rho_{2k}} T + A^\top \tau_{\rho_{3k}}$, which is a natural choice in practice. For fixed T and A, the observed distribution is fully determined by the three probabilities $\Pr(D_1 = 1, D_2 = 0, V = 1|T, A)$, $\Pr(D_1 = 0, D_2 = 1, V = 1|T, A)$ and $\Pr(D_1 = 0, D_2 = 0, V = 1|T, A)$. It is easy to show that

$$\begin{aligned}\Pr(D_1 = 1, D_2 = 0, V = 1|T, A) &= \Pr(D_1 = 1, D_2 = 0|T, A)\Pr(V = 1|D_1 = 1, D_2 = 0, T, A)\\ &= \Pr(D_1 = 1|T, A)\Pr(V = 1|D_1 = 1, D_2 = 0, T, A)\\ &= \rho_1 \pi_{10}.\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\Pr(D_1 = 0, D_2 = 1, V = 1|T, A) &= \rho_2 \pi_{01},\\ \Pr(D_1 = 0, D_2 = 0, V = 1|T, A) &= (1 - \rho_1 - \rho_2)\pi_{00},\end{aligned}$$

with $\pi_{01} = \Pr(V = 1|D_1 = 0, D_2 = 1, T, A)$ and $\pi_{00} = \Pr(V = 1|D_1 = 0, D_2 = 0, T, A)$. Then,

$$\Pr(V = 1|T, A) = \rho_1 \pi_{10} + \rho_2 \pi_{01} + (1 - \rho_1 - \rho_2)\pi_{00},$$

and $\Pr(V = 0|T, A) = 1 - \Pr(V = 1|T, A) = 1 - \rho_1 \pi_{10} + \rho_2 \pi_{01} + (1 - \rho_1 - \rho_2)\pi_{00}$. It follows that the log-likelihood function can be written as:

$$\begin{aligned}&\log L(\lambda, \tau_\pi, \tau_\rho)\\ &= \sum_{i=1}^{n} \Big\{ D_{1i} V_i \log(\rho_{1i}\pi_{10i}) + D_{2i} V_i \log(\rho_{2i}\pi_{01i}) + (1 - D_{1i} - D_{2i}) V_i \log((1 - \rho_{1i} - \rho_{2i})\pi_{00i})\\ &\quad + (1 - V_i)\log(1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}) \Big\}. \qquad (5.5)\end{aligned}$$

The estimates $\hat{\lambda}$, $\hat{\tau}_\pi$, and $\hat{\tau}_\rho$ can be obtained by maximizing $\log L(\lambda, \tau_\pi, \tau_\rho)$ or by solving the score equations

$$\begin{aligned}0 &= \sum_{i=1}^{n} \left\{ D_{1i} V_i (1 - \pi_{10i}) - \frac{(1 - V_i)\rho_{1i}\pi_{10i}(1 - \pi_{10i})}{1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}} \right\},\\ 0 &= \sum_{i=1}^{n} \left\{ D_{2i} V_i (1 - \pi_{01i}) - \frac{(1 - V_i)\rho_{2i}\pi_{01i}(1 - \pi_{01i})}{1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}} \right\},\\ 0 &= \sum_{i=1}^{n} U_i \Big\{ D_{1i} V_i (1 - \pi_{10i}) + D_{2i} V_i (1 - \pi_{01i}) + (1 - D_{1i} - D_{2i}) V_i (1 - \pi_{00i})\\ &\quad - (1 - V_i)\frac{\rho_{1i}\pi_{10i}(1 - \pi_{10i}) + \rho_{2i}\pi_{01i}(1 - \pi_{01i}) + (1 - \rho_{1i} - \rho_{2i})\pi_{00i}(1 - \pi_{00i})}{1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}} \Big\},\\ 0 &= \sum_{i=1}^{n} U_i \left\{ V_i (D_{1i} - \rho_{1i}) - (1 - V_i)\frac{(\pi_{10i} - \pi_{00i})\rho_{1i}(1 - \rho_{1i}) - (\pi_{01i} - \pi_{00i})\rho_{1i}\rho_{2i}}{1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}} \right\},\\ 0 &= \sum_{i=1}^{n} U_i \left\{ V_i (D_{2i} - \rho_{2i}) - (1 - V_i)\frac{(\pi_{01i} - \pi_{00i})\rho_{2i}(1 - \rho_{2i}) - (\pi_{10i} - \pi_{00i})\rho_{1i}\rho_{2i}}{1 - \rho_{1i}\pi_{10i} - \rho_{2i}\pi_{01i} - (1 - \rho_{1i} - \rho_{2i})\pi_{00i}} \right\},\end{aligned}$$

where $U_i = (1, T_i, A_i^\top)^\top$. The above equations are obtained by using the following results

$$\begin{aligned}
\frac{\partial}{\partial \lambda_1}\pi_{10i} &= \pi_{10i}(1-\pi_{10i}),\\
\frac{\partial}{\partial \lambda_2}\pi_{01i} &= \pi_{01i}(1-\pi_{01i}),\\
\frac{\partial}{\partial \tau_\pi}\pi_{d_1d_2i} &= U_i(1-\pi_{d_1d_2i})\pi_{d_1d_2i}
\end{aligned}$$

(here (d_1, d_2) is a pair in the set $\{(1,0),(0,1),(0,0)\}$), and

$$\begin{aligned}
\tfrac{\partial}{\partial \tau_{\rho_1}}\rho_{1i} &= U_i\rho_{1i}(1-\rho_{1i}); & \tfrac{\partial}{\partial \tau_{\rho_2}}\rho_{1i} &= -U_i\rho_{1i}\rho_{2i};\\
\tfrac{\partial}{\partial \tau_{\rho_2}}\rho_{2i} &= U_i\rho_{2i}(1-\rho_{2i}); & \tfrac{\partial}{\partial \tau_{\rho_1}}\rho_{2i} &= -U_i\rho_{1i}\rho_{2i}.
\end{aligned}$$

5.1.3 Identifiability

In this section we verify that the working model based on (5.1), with $h(T, A; \tau_\pi) = \tau_{\pi_1} + \tau_{\pi_2}T + A^\top\tau_{\pi_3}$, and (5.2), with $f(T, A; \tau_{\rho_k}) = \tau_{\rho_{1k}} + \tau_{\rho_{2k}}T + A^\top\tau_{\rho_{3k}}$, is identifiable. Since the log–likelihood (5.5) is fully determined by the three probabilities $\Pr(D_1 = 1, D_2 = 0, V = 1|T, A)$, $\Pr(D_1 = 0, D_2 = 1, V = 1|T, A)$ and $\Pr(D_1 = 0, D_2 = 0, V = 1|T, A)$, we have to show that such probabilities are uniquely determined by the parameters for all possible T and A. For the sake of simplicity, in the remainder of this section the auxiliary covariates A is omitted (actually, we can always view A as fixed while varying T).

Let $\xi = (\lambda_1, \lambda_2, \tau_{\pi_1}, \tau_{\pi_2}, \tau_{\rho_{11}}, \tau_{\rho_{21}}, \tau_{\rho_{12}}, \tau_{\rho_{22}})^\top$ be the set of parameters. For given $T = t$, we can write

$$\begin{aligned}
\log(\rho_1\pi_{10}) &= (\tau_{\rho_{11}} + \tau_{\rho_{21}}t) - \log\{1 + \exp(\tau_{\rho_{11}} + \tau_{\rho_{21}}t) + \exp(\tau_{\rho_{12}} + \tau_{\rho_{22}}t)\} + (\tau_{\pi_1} + \tau_{\pi_2}t) + \lambda_1\\
&\quad - \log\{1 + \exp(\tau_{\pi_1} + \tau_{\pi_2}t)\exp(\lambda_1)\},\\
\log(\rho_2\pi_{01}) &= (\tau_{\rho_{12}} + \tau_{\rho_{22}}t) - \log\{1 + \exp(\tau_{\rho_{11}} + \tau_{\rho_{21}}t) + \exp(\tau_{\rho_{12}} + \tau_{\rho_{22}}t)\} + (\tau_{\pi_1} + \tau_{\pi_2}t) + \lambda_2\\
&\quad - \log\{1 + \exp(\tau_{\pi_1} + \tau_{\pi_2}t)\exp(\lambda_2)\},\\
\log(\rho_3\pi_{00}) &= -\log\{1 + \exp(\tau_{\rho_{11}} + \tau_{\rho_{21}}t) + \exp(\tau_{\rho_{12}} + \tau_{\rho_{22}}t)\} + (\tau_{\pi_1} + \tau_{\pi_2}t)\\
&\quad - \log\{1 + \exp(\tau_{\pi_1} + \tau_{\pi_2}t)\}.
\end{aligned}$$

Let $x(t) = \tau_{\pi_1} + \tau_{\pi_2}t$, $y(t) = \tau_{\rho_{11}} + \tau_{\rho_{21}}t$ and $z(t) = \tau_{\rho_{12}} + \tau_{\rho_{22}}t$, for each $t \in \mathbb{R}$. The above expressions, which refer to the quantities characterizing the log–likelihood function (5.5), can be rewritten as

$$\begin{aligned}
\log(\rho_3\pi_{00}) &= -\log\{1 + \exp(y(t)) + \exp(z(t))\} + x(t) - \log\{1 + \exp(x(t))\},\\
\log(\rho_1\pi_{10}) &= y(t) - \log\{1 + \exp(y(t)) + \exp(z(t))\} + x(t) + \lambda_1 - \log\{1 + \exp(x(t))\exp(\lambda_1)\}\\
&= \log(\rho_3\pi_{00}) + \log\{1 + \exp(x(t))\} + y(t) + \lambda_1 - \log\{1 + \exp(x(t))\exp(\lambda_1)\}\\
&= \log(\rho_3\pi_{00}) + y(t) + \log\{1 + \exp(x(t))\} - \log\{\exp(-\lambda_1) + \exp(x(t))\}\\
&= \log(\rho_3\pi_{00}) + y(t) + \log\left\{\frac{1 + \exp(x(t))}{\exp(-\lambda_1) + \exp(x(t))}\right\},\\
\log(\rho_2\pi_{01}) &= z(t) - \log\{1 + \exp(y(t)) + \exp(z(t))\} + x(t) + \lambda_2 - \log\{1 + \exp(x(t))\exp(\lambda_2)\}\\
&= \log(\rho_3\pi_{00}) + z(t) + \log\left\{\frac{1 + \exp(x(t))}{\exp(-\lambda_2) + \exp(x(t))}\right\}.
\end{aligned}$$

Now, assume that there are two distinct points ξ and ξ^* ($\xi \neq \xi^*$) in the parameter space, such that the following equations (with obvious notation) hold:

$$\rho_1\pi_{10} = \rho_1^*\pi_{10}^*, \tag{5.6}$$
$$\rho_2\pi_{01} = \rho_2^*\pi_{01}^*, \tag{5.7}$$
$$\rho_3\pi_{00} = \rho_3^*\pi_{00}^*, \tag{5.8}$$

for all $t \in \mathbb{R}$. By using (5.8), the equations (5.6) and (5.7) are equivalent to

$$y(t) - y^*(t) = \log\left\{\frac{1+\exp(x^*(t))}{\exp(-\lambda_1^*)+\exp(x^*(t))}\right\} - \log\left\{\frac{1+\exp(x(t))}{\exp(-\lambda_1)+\exp(x(t))}\right\}, \tag{5.9}$$
$$z(t) - z^*(t) = \log\left\{\frac{1+\exp(x^*(t))}{\exp(-\lambda_2^*)+\exp(x^*(t))}\right\} - \log\left\{\frac{1+\exp(x(t))}{\exp(-\lambda_2)+\exp(x(t))}\right\}, \tag{5.10}$$

respectively. In (5.9) and (5.10) the left sides are straight lines. Thus, in order to (5.9) and (5.10) hold for all t, the right sides must be constants. If these constants were 0 (because $\lambda_1 = \lambda_1^* = \lambda_2 = \lambda_2^* = 0$), then (5.8) would no longer hold for $\xi \neq \xi^*$ and all t. Alternatively, the right sides of (5.9) and (5.10) are non-zero constants if $\tau_{\pi_2} = \tau_{\pi_2}^* = 0$. Then, as a consequence, (5.8) still is valid, for $\xi \neq \xi^*$ and all t, eventually if $\tau_{\rho_{21}} = \tau_{\rho_{21}}^* = 0$ and $\tau_{\rho_{22}} = \tau_{\rho_{22}}^* = 0$. This allows us to state that: if $\Pr(D_k|T) \neq \Pr(D_k)$, with $k = 1, 2$, then the considered model (with the particular choice for the functions h and f) is identifiable, i.e., the joint probabilities $\Pr(D_1 = 1, D_2 = 0, V = 1|T = t)$, $\Pr(D_1 = 0, D_2 = 1, V = 1|T = t)$ and $\Pr(D_1 = 0, D_2 = 0, V = 1|T = t)$ are determined by a unique set of parameters. Of course, this claim can be easily extended to handle the presence of a covariate vector, A.

5.2 The proposal

5.2.1 VUS estimators

Let $\rho_{k(v)} = \Pr(D_k = 1|V = v, T, A)$, for $k = 1, 2$ and $v = 0, 1$. It is easy to see, for instance, that

$$\rho_{1(v)} = \frac{\Pr(V = v, D_1 = 1|D_2 = 0, T, A)}{\Pr(V = v|T, A)} = \frac{\Pr(V = v|D_1 = 1, D_2 = 0, T, A)\Pr(D_1 = 1|T, A)}{\Pr(V = v|T, A)}.$$

Hence, we can get, in particular,

$$\rho_{1(0)} = \frac{(1-\pi_{10})\rho_1}{(1-\pi_{10})\rho_1 + (1-\pi_{01})\rho_2 + (1-\pi_{00})\rho_3},$$
$$\rho_{2(0)} = \frac{(1-\pi_{01})\rho_2}{(1-\pi_{10})\rho_1 + (1-\pi_{01})\rho_2 + (1-\pi_{00})\rho_3},$$
$$\rho_{3(0)} = \frac{(1-\pi_{00})\rho_3}{(1-\pi_{10})\rho_1 + (1-\pi_{01})\rho_2 + (1-\pi_{00})\rho_3}.$$

Clearly, we also may consider quantities as

$$\rho_{1(1)} = \frac{\pi_{10}\rho_1}{\pi_{10}\rho_1 + \pi_{01}\rho_2 + \pi_{00}\rho_3}.$$

Then, we observe that

$$\begin{aligned}\mathbb{E}(D_{1i}D_{2\ell}D_{3r}\mathrm{I}_{i\ell r}) &= \mathbb{E}_{T,A}\left\{\mathrm{I}_{i\ell r}\mathbb{E}(D_{1i}D_{2\ell}D_{3r}|T_i, A_i, T_\ell, A_\ell, T_r, A_r)\right\},\\ &= \mathbb{E}_{T,A}\left\{\mathrm{I}_{i\ell r}\mathbb{E}(D_{1i}|T_i, A_i)\mathbb{E}(D_{2\ell}|T_\ell, A_\ell)\mathbb{E}(D_{3r}|T_r, A_r)\right\},\\ &= \mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\mathrm{I}_{i\ell r}\right).\end{aligned}$$

Similarly, we have

$$\mathbb{E}(D_{1i}D_{2\ell}D_{3r}) = \mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\right),$$

so that (2.10) can be rewritten as

$$\mu = \frac{\mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\mathrm{I}_{i\ell r}\right)}{\mathbb{E}_{T,A}\left(\rho_{1i}\rho_{2\ell}\rho_{3r}\right)}. \tag{5.11}$$

Equation (5.11) suggests how to build estimators of VUS when some disease labels are missing in the sample: we can use suitable estimates $\hat{\rho}_{ki}$ to replace the D_{ki}'s in (2.12). Therefore, a FI estimator of VUS is simply

$$\hat{\mu}_{\mathrm{FI}} = \frac{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\mathrm{I}_{i\ell r}\hat{\rho}_{1i}\hat{\rho}_{2\ell}\hat{\rho}_{3r}}{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\hat{\rho}_{1i}\hat{\rho}_{2\ell}\hat{\rho}_{3r}}, \tag{5.12}$$

where $\hat{\rho}_{ki}$ ($k = 1, 2, 3$ and $i = 1, \ldots, n$) are the estimated disease probabilities obtained from the disease model (5.2).

Since $\mathbb{E}[V_i\rho_{k(1)i} + (1 - V_i)\rho_{k(0)i}|T, A] = \rho_{ki}$, an alternative FI estimator of VUS could be obtained by replacing D_{ki}'s in (2.12) with the estimates $\tilde{D}_{ki,\mathrm{FI}} = V_i\hat{\rho}_{k(1)i} + (1 - V_i)\hat{\rho}_{k(0)i}$. Unlike FI approach, MSI estimator only replace the disease status D_{ki} by the estimate $\hat{\rho}_{k(0)i}$ for unverified subjects. Define $D_{ki,\mathrm{MSI}} = V_iD_{ki} + (1 - V_i)\rho_{k(0)i}$ and let $\tilde{D}_{ki,\mathrm{MSI}}$ be the estimated version with $\rho_{k(0)i}$ replaced by $\hat{\rho}_{k(0)i}$, and

$$\begin{aligned}
\hat{\rho}_{1(0)i} &= \frac{(1-\hat{\pi}_{10i})\hat{\rho}_{1i}}{(1-\hat{\pi}_{10i})\hat{\rho}_{ki} + (1-\hat{\pi}_{01i})\hat{\rho}_{2i} + (1-\hat{\pi}_{00i})\hat{\rho}_{3i}},\\
\hat{\rho}_{2(0)i} &= \frac{(1-\hat{\pi}_{01i})\hat{\rho}_{2i}}{(1-\hat{\pi}_{10i})\hat{\rho}_{1i} + (1-\hat{\pi}_{01i})\hat{\rho}_{2i} + (1-\hat{\pi}_{00i})\hat{\rho}_{3i}},\\
\hat{\rho}_{3(0)i} &= \frac{(1-\hat{\pi}_{00i})\hat{\rho}_{3i}}{(1-\hat{\pi}_{10i})\hat{\rho}_{1i} + (1-\hat{\pi}_{01i})\hat{\rho}_{2i} + (1-\hat{\pi}_{00i})\hat{\rho}_{3i}}.
\end{aligned}$$

Here, $\hat{\pi}_{10i} = \widehat{\Pr}(V_i = 1|D_{1i} = 1, D_{2i} = 0, T_i, A_i)$, $\hat{\pi}_{01i} = \widehat{\Pr}(V_i = 1|D_{1i} = 0, D_{2i} = 1, T_i, A_i)$ and $\hat{\pi}_{00i} = \widehat{\Pr}(V_i = 1|D_{1i} = 0, D_{2i} = 0, T_i, A_i)$. Such estimates are derived from the verification model (5.1). Then, the MSI estimator of VUS is

$$\hat{\mu}_{\mathrm{MSI}} = \frac{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\mathrm{I}_{i\ell r}\tilde{D}_{1i,\mathrm{MSI}}\tilde{D}_{2\ell,\mathrm{MSI}}\tilde{D}_{3r,\mathrm{MSI}}}{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\tilde{D}_{1i,\mathrm{MSI}}\tilde{D}_{2\ell,\mathrm{MSI}}\tilde{D}_{3r,\mathrm{MSI}}}. \tag{5.13}$$

In the IPW approach, instead, each observation in the subset of verified units is weighted by the inverse of the probability that the unit was selected for verification. Thus, the IPW estimator of VUS is

$$\hat{\mu}_{\mathrm{IPW}} = \frac{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\mathrm{I}_{i\ell r}V_iV_\ell V_rD_{1i}D_{2\ell}D_{3r}\hat{\pi}_i^{-1}\hat{\pi}_\ell^{-1}\hat{\pi}_r^{-1}}{\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}V_iV_\ell V_rD_{1i}D_{2\ell}D_{3r}\hat{\pi}_i^{-1}\hat{\pi}_\ell^{-1}\hat{\pi}_r^{-1}}. \tag{5.14}$$

Clearly, the estimates $\hat{\pi}_i$ also arise from the selection model (5.1).

The last estimator is the pseudo doubly robust (PDR) estimator. We define

$$D_{ki,\mathrm{PDR}} = \frac{V_i D_{ki}}{\pi_i} - \frac{\rho_{k(0)i}(V_i - \pi_i)}{\pi_i}.$$

An estimated version, $\tilde{D}_{ki,\mathrm{PDR}}$, is obtained by entering the estimates $\hat{\pi}_i$ and $\hat{\rho}_{k(0)i}$ in the expression above. Then, the PDR estimator of VUS is

$$\hat{\mu}_{\mathrm{PDR}} = \frac{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \mathrm{I}_{i\ell r} \tilde{D}_{1i,\mathrm{PDR}} \tilde{D}_{2\ell,\mathrm{PDR}} \tilde{D}_{3r,\mathrm{PDR}}}{\sum_{i=1}^{n} \sum_{\ell=1,\ell\neq i}^{n} \sum_{\substack{r=1 \\ r\neq\ell, r\neq i}}^{n} \tilde{D}_{1i,\mathrm{PDR}} \tilde{D}_{2\ell,\mathrm{PDR}} \tilde{D}_{3r,\mathrm{PDR}}}. \tag{5.15}$$

The PDR estimator has the same nature as the SPE estimator discussed in Chapter 4 under MAR assumption. However, under NI missing data mechanism it no longer has the doubly robust property. In fact, correct specification of both the verification model and the disease model is required for the PDR estimator to be consistent.

Note that all VUS estimators basically require maximum likelihood estimates of the parameters λ, τ_π and τ_ρ of the working models (5.1) and (5.2).

5.2.2 Asymptotic behavior

Let $\xi = (\lambda^\top, \tau_\pi^\top, \tau_\rho^\top)^\top$ be the nuisance parameter. Observe that the proposed VUS estimators can be found as solutions of appropriate estimating equations (solved along with the score equations). The estimating functions for FI, MSI, IPW and PDR estimators have generic term (corresponding to a generic triplet of sample units), respectively,

$$\begin{aligned}
G_{i\ell r,\mathrm{FI}}(\mu,\xi) &= \rho_{1i}(\tau_\rho)\rho_{2\ell}(\tau_\rho)\rho_{3r}(\tau_\rho)\,(I_{i\ell r} - \mu)\,, \\
G_{i\ell r,\mathrm{MSI}}(\mu,\xi) &= D_{1i,\mathrm{MSI}}(\xi)D_{2\ell,\mathrm{MSI}}(\xi)D_{3r,\mathrm{MSI}}(\xi)\,(I_{i\ell r} - \mu)\,, \\
G_{i\ell r,\mathrm{IPW}}(\mu,\xi) &= \frac{V_i V_\ell V_r D_{1i} D_{2\ell} D_{3r}}{\pi_i(\xi)\pi_\ell(\xi)\pi_k(\xi)}\,(I_{i\ell r} - \mu)\,, \\
G_{i\ell r,\mathrm{PDR}}(\mu,\xi) &= D_{1i,\mathrm{PDR}}(\xi)D_{2\ell,\mathrm{PDR}}(\xi)D_{3r,\mathrm{PDR}}(\xi)\,(I_{i\ell r} - \mu)\,.
\end{aligned}$$

In the following, we will use the general notation $G_{i\ell r,*}(\mu,\xi)$, where the star stands for FI, MSI, IPW and PDR. We define the observed data as the set $\{O_i = (\mathcal{D}_i^\top, V_i, T_i, A_i^\top)^\top,\ \ i = 1,\ldots,n\}$.

Remark 5.2.1. *Here, we show that the estimating functions $G_{i\ell r,*}$ are unbiased under the working disease and verification models. Recall that $\xi = (\lambda^\top, \tau_\pi^\top, \tau_\rho^\top)^\top$.*

- *FI estimator. We have*

$$\begin{aligned}
\mathbb{E}\left\{G_{i\ell r,\mathrm{FI}}(\mu_0,\xi_0)\right\} &= \mathbb{E}\left\{\rho_{1i}(\tau_{0\rho})\rho_{2\ell}(\tau_{0\rho})\rho_{3r}(\tau_{0\rho})(I_{i\ell r} - \mu)\right\} \\
&= \mathbb{E}\left\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r} - \mu_0)\right\}.
\end{aligned}$$

Hence, $\mathbb{E}\left\{G_{i\ell r,\mathrm{FI}}(\mu_0,\xi_0)\right\} = 0$ from (4.1).

- *MSI estimator. Consider* $\mathbb{E}\{D_{ki,\mathrm{MSI}}(\xi_0)|T_i, A_i\}$. *We have*

$$\begin{aligned}
\mathbb{E}\{D_{ki,\mathrm{MSI}}(\xi_0)|T_i, A_i\} &= \mathbb{E}\left\{V_i D_{ki} + (1-V_i)\rho_{k(0)i}(\xi_0)|T_i, A_i\right\} \\
&= \mathbb{E}\left[\mathbb{E}\left\{V_i D_{ki} + (1-V_i)\rho_{k(0)i}(\xi_0)|T_i, A_i, V_i\right\}|T_i, A_i\right] \\
&= \Pr(V_i = 1|T_i, A_i)\mathbb{E}\left(D_{ki}|V_i = 1, T_i, A_i\right) \\
&\quad + \Pr(V_i = 0|T_i, A_i)\mathbb{E}\left(\rho_{k(0)i}(\xi_0)|V_i = 0, T_i, A_i\right) \\
&= \Pr(V_i = 1|T_i, A_i)\Pr(D_{ki} = 1|V_i = 1, T_i, A_i) \\
&\quad + \Pr(V_i = 0|T_i, A_i)\Pr(D_{ki} = 1|V_i = 0, T_i, A_i) \\
&= \Pr(D_{ki} = 1|T_i, A_i) = \rho_{ki}.
\end{aligned}$$

  Therefore,

$$\begin{aligned}
\mathbb{E}\{G_{i\ell r,\mathrm{MSI}}(\mu_0, \xi_0)\} &= \mathbb{E}\{D_{1i,\mathrm{MSI}}(\xi_0)D_{2\ell,\mathrm{MSI}}(\xi_0)D_{3r,\mathrm{MSI}}(\xi_0)(I_{i\ell r} - \mu_0)\} \\
&= \mathbb{E}\Big[(I_{i\ell r} - \mu_0)\mathbb{E}\{D_{1i,\mathrm{MSI}}(\xi_0)|T_i, A_i\}\mathbb{E}\{D_{2\ell,\mathrm{MSI}}(\xi_0)|T_\ell, A_\ell\} \\
&\quad \times \mathbb{E}\{D_{3r,\mathrm{MSI}}(\xi_0)|T_r, A_r\}\Big] \\
&= \mathbb{E}\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r} - \mu_0)\}.
\end{aligned}$$

- *IPW estimator. In this case,*

$$\begin{aligned}
\mathbb{E}\left(V_i D_{ki}\pi_i^{-1}(\xi_0)|T_i, A_i\right) &= \pi_i^{-1}(\xi_0)\mathbb{E}\left(V_i D_{ki}|T_i, A_i\right) \\
&= \pi_i^{-1}(\xi_0)\mathbb{E}\left\{D_{ki}\mathbb{E}\left(V_i|D_{1i}, D_{2i}, T_i, A_i\right)\big|T_i, A_i\right\} \\
&= \pi_i^{-1}\mathbb{E}\left(\pi_i D_{ki}|T_i, A_i\right) = \rho_{ki}.
\end{aligned}$$

  Thus,

$$\begin{aligned}
\mathbb{E}\{G_{i\ell r,\mathrm{IPW}}(\mu_0, \xi_0)\} &= \mathbb{E}\left\{\frac{V_i V_\ell V_r D_{1i} D_{2\ell} D_{3r}}{\pi_i(\xi_0)\pi_\ell(\xi_0)\pi_k(\xi_0)}(I_{i\ell r} - \mu_0)\right\} \\
&= \mathbb{E}\Big\{(I_{i\ell r} - \mu_0)\mathbb{E}(V_i D_{1i}\pi_i^{-1}(\xi_0)|T_i, A_i)\mathbb{E}(V_\ell D_{2\ell}\pi_\ell^{-1}(\xi_0)|T_\ell, A_\ell) \\
&\quad \times \mathbb{E}(V_r D_{3r}\pi_r^{-1}(\xi_0)|T_r, A_r)\Big\} \\
&= \mathbb{E}\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r} - \mu_0)\}.
\end{aligned}$$

- *PDR estimator.*

$$\begin{aligned}
\mathbb{E}\{D_{ki,\mathrm{PDR}}(\xi_0)|T_i, A_i\} &= \mathbb{E}\left[\mathbb{E}\left\{\frac{V_i D_{ki}}{\pi_i(\xi_0)} - \rho_{k(0)i}(\xi_0)\left(\frac{V_i}{\pi_i(\xi_0)} - 1\right)\middle|D_{1i}, D_{2i}, T_i, A_i\right\}\middle|T_i, A_i\right] \\
&= \mathbb{E}\Bigg\{D_{ki}\mathbb{E}\left(\frac{V_i}{\pi_i(\xi_0)}\middle|D_{1i}, D_{2i}, T_i, A_i\right) \\
&\quad - \rho_{k(0)i}(\xi_0)\mathbb{E}\left(\frac{V_i}{\pi_i(\xi_0)} - 1\middle|D_{1i}, D_{2i}, T_i, A_i\right)\middle|T_i, A_i\Bigg\} \\
&= \mathbb{E}(D_{ki}|T_i, A_i) = \rho_{ki}.
\end{aligned}$$

  Hence,

$$\begin{aligned}
\mathbb{E}\{G_{i\ell r,\mathrm{PDR}}(\mu_0, \xi_0)\} &= \mathbb{E}\{D_{1i,\mathrm{PDR}}(\xi_0)D_{2\ell,\mathrm{PDR}}(\xi_0)D_{3r,\mathrm{PDR}}(\xi_0)(I_{i\ell r} - \mu_0)\} \\
&= \mathbb{E}\Big[(I_{i\ell r} - \mu_0)\mathbb{E}\{D_{1i,\mathrm{PDR}}(\xi_0)|T_i, A_i\}\mathbb{E}\{D_{2\ell,\mathrm{PDR}}(\xi_0)|T_\ell, A_\ell\} \\
&\quad \times \mathbb{E}\{D_{3r,\mathrm{PDR}}(\xi_0)|T_r, A_r\}\Big] \\
&= \mathbb{E}\{\rho_{1i}\rho_{2\ell}\rho_{3r}(I_{i\ell r} - \mu_0)\}.
\end{aligned}$$

Recall that the nuisance parameters ξ is estimated by maximizing the log–likelihood function (5.5). Let $\mathcal{S}_i(\xi)$ is the i–th subject's contribution to the score function, and $\mathcal{I}(\xi) = -\mathbb{E}\frac{\partial}{\partial\xi^\top}\mathcal{S}_i(\xi)$ the Fisher information matrix for ξ. To give general theoretical results, we assume standard regularity conditions, which ensure consistency and asymptotic normality of the maximum likelihood estimator $\hat{\xi}$.

(R1) the parameter space $\boldsymbol{\xi} \equiv \boldsymbol{\lambda} \times \boldsymbol{\tau_\pi} \times \boldsymbol{\tau_\rho}$ has finite dimension and is compact;

(R2) the true value $\xi_0 = (\lambda_0^\top, \tau_{0\pi}^\top, \tau_{0\rho}^\top)^\top$ exists and is interior to the parameter space $\boldsymbol{\xi}$ such that $\mathbb{E}\{\mathcal{S}_i(\xi)\} \neq 0$ if $\xi \neq \xi_0$ and $\mathbb{E}\{\mathcal{S}_i(\xi_0)\} = 0$;

(R3) the variance of $\mathcal{S}_i(\xi_0)$ exists and is finite;

(R4) $\mathbb{E}\left\{\partial\mathcal{S}_i(\xi)/\partial\xi^\top|_{\xi=\xi_0}\right\}$ exists and is invertible;

(R5) there exists a neighborhood $\mathcal{N}$ of ξ_0 such that the expected values of $\sup_{(\xi)\in\mathcal{N}}\|\mathcal{S}_i(\xi)\|$, $\sup_{(\xi)\in\mathcal{N}}\|\partial\mathcal{S}_i(\xi)/\partial\xi^\top\|$ and $\sup_{(\xi)\in\mathcal{N}}\|\mathcal{S}_i(\xi)\mathcal{S}_i(\xi)^\top\|$ are finite, where $\|\mathbf{X}\| \equiv \sum_i\sum_j X_{ij}^2$.

Let μ_0 be the true VUS value. We also assume that:

(C1) The U–process

$$U_{n,*}(\mu,\xi) = \sqrt{n}\{G_*(\mu,\xi) - e_*(\mu,\xi)\}$$

is stochastically equicontinuous, where

$$\begin{aligned} G_*(\mu,\xi) &= \frac{1}{6n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\ell=1,\ell\neq i}^{n}\sum_{\substack{r=1\\ r\neq\ell,r\neq i}}^{n}\Big\{G_{i\ell r,*}(\mu,\xi) + G_{ir\ell,*}(\mu,\xi) \\ &\quad + G_{\ell ir,*}(\mu,\xi) + G_{\ell ri,*}(\mu,\xi) + G_{ri\ell,*}(\mu,\xi) + G_{r\ell i,*}(\mu,\xi)\Big\} \end{aligned}$$

and

$$\begin{aligned} e_*(\mu,\xi) &= \frac{1}{6}\mathbb{E}\Big\{G_{i\ell r,*}(\mu,\xi) + G_{ir\ell,*}(\mu,\xi) + G_{\ell ir,*}(\mu,\xi) + G_{\ell ri,*}(\mu,\xi) \\ &\quad + G_{ri\ell,*}(\mu,\xi) + G_{r\ell i,*}(\mu,\xi)\Big\}; \end{aligned}$$

(C2) $e_*(\mu,\xi)$ is differentiable in (μ,ξ), and $\frac{\partial e_*(\mu,\xi_0)}{\partial\mu}\Big|_{\mu=\mu_0} \neq 0$;

(C3) $G_*(\mu,\xi)$ and $\frac{\partial G_*(\mu,\xi)}{\partial\xi}$ converges uniformly (in probability) to $e_*(\mu,\xi)$ and $\frac{\partial e_*(\mu,\xi)}{\partial\xi}$, respectively.

Then, we prove consistency and asymptotic normality of the proposed estimators.

Theorem 5.2.2 (Consistency)**.** *Suppose that conditions (C1)–(C3) hold, along with standard regularity conditions for the likelihood function (as those given by Newey and McFadden (1994)). Under the verification model (5.1) and the disease model (5.2),* $\hat{\mu}_* \xrightarrow{p} \mu_0$.

Proof. We shown that $\mathbb{E}\{G_{i\ell r,*}(\mu_0,\xi_0)\} = 0$. Then $e_*(\mu_0,\xi_0) = 0$, and, by condition (C2) and an application of implicit function theorem, there exists a neighborhood of ξ_0 in which a continuously differentiable function, $m(\xi)$, is uniquely defined such that $m(\xi_0) = \mu_0$ and $e_*(m(\xi),\xi) = 0$. Since the maximum likelihood estimator $\hat{\xi}$ is consistent, i.e., $\hat{\xi} \xrightarrow{p} \xi_0$, we have that $\tilde{\mu}_* = m(\hat{\xi}) \xrightarrow{p} \mu_0$. On the other hand, $G_*(\hat{\mu}_*,\hat{\xi}) = 0$ and condition (C3) implies that $e_*(\hat{\mu}_*,\hat{\xi}) \xrightarrow{p} 0$. Thus, $\hat{\mu}_* \xrightarrow{p} \tilde{\mu}_*$. □

Next we establish the asymptotic normality of the estimators $\hat{\mu}_*$.

Theorem 5.2.3 (Asymptotic normality)**.** *Suppose the conditions in Theorem 5.2.2 are satisfied. If the verification model (5.1) and the disease model (5.2) hold, then*

$$\sqrt{n}\,(\hat{\mu}_* - \mu_0) \xrightarrow{d} \mathcal{N}(0, \Lambda_*),$$

where the star indicates FI, MSI, IPW, PDR, and Λ_* is a suitable value.

Proof. We have

$$\begin{aligned} 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\xi}) \\ 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\xi}) + \sqrt{n}e_*(\hat{\mu}_*, \hat{\xi}) - \sqrt{n}e_*(\hat{\mu}_*, \hat{\xi}). \end{aligned}$$

Since $e_*(\mu_0, \xi_0) = 0$, we get

$$\begin{aligned} 0 &= \sqrt{n}G_*(\hat{\mu}_*, \hat{\xi}) + \sqrt{n}e_*(\hat{\mu}_*, \hat{\xi}) - \sqrt{n}e_*(\hat{\mu}_*, \hat{\xi}) + \sqrt{n}e_*(\mu_0, \xi_0) - \sqrt{n}e_*(\mu_0, \xi_0) \\ &= \sqrt{n}\left\{G_*(\hat{\mu}_*, \hat{\xi}) - e_*(\hat{\mu}_*, \hat{\xi})\right\} + \sqrt{n}\left\{e_*(\hat{\mu}_*, \hat{\xi}) - e_*(\mu_0, \xi_0)\right\} + \sqrt{n}e_*(\mu_0, \xi_0) \\ &\quad - \sqrt{n}G_*(\mu_0, \xi_0) + \sqrt{n}G_*(\mu_0, \xi_0) \\ &= \left[\sqrt{n}\left\{G_*(\hat{\mu}_*, \hat{\xi}) - e_*(\hat{\mu}_*, \hat{\xi})\right\} - \sqrt{n}\left\{G_*(\mu_0, \xi_0) - e_*(\mu_0, \xi_0)\right\}\right] \\ &\quad + \sqrt{n}\left\{e_*(\hat{\mu}_*, \hat{\xi}) - e_*(\mu_0, \xi_0)\right\} + \sqrt{n}G_*(\mu_0, \xi_0). \end{aligned}$$

Condition (C1) implies that the first term in the right hand side of the last identity equals to $o_p(1)$. Using the Taylor expansion, we have

$$\begin{aligned} 0 &= o_p(1) + \sqrt{n}\left\{e_*(\hat{\mu}_*, \hat{\xi}) - e_*(\mu_0, \xi_0)\right\} + \sqrt{n}G_*(\mu_0, \xi_0) \\ &= o_p(1) + \left.\frac{\partial e_*(\mu, \xi_0)}{\partial \mu}\right|_{\mu=\mu_0} \sqrt{n}(\hat{\mu}_* - \mu_0) + \left.\frac{\partial e_*^\top(\mu_0, \xi)}{\partial \xi}\right|_{\xi=\xi_0} \sqrt{n}(\hat{\xi} - \xi_0) \\ &\quad + \sqrt{n}G_*(\mu_0, \xi_0). \end{aligned} \tag{5.16}$$

It is straightforward to show that

$$\left.\frac{\partial e_*(\mu, \xi_0)}{\partial \mu}\right|_{\mu=\mu_0} = -\Pr(D_1 = 1)\Pr(D_2 = 1)\Pr(D_3 = 1) = -\theta_1\theta_2\theta_3.$$

By standard results on the limit distribution of U-statistics (van der Vaart, 2000, Theorem 12.3, Chap. 12),

$$\sqrt{n}U_{n,*}(\mu_0, \xi_0) = \sqrt{n}\left\{G_*(\mu_0, \xi_0) - e_*(\mu_0, \xi_0)\right\} = \sqrt{n}G_*(\mu_0, \xi_0) \xrightarrow{p} \sqrt{n}\tilde{G}_*(\mu_0, \xi_0),$$

where $\sqrt{n}\tilde{G}_*(\mu, \xi)$ is the projection of $U_{n,*}$ onto the set of all statistics of the form $\sum_{i=1}^n B_i(X_i)$,

$$\begin{aligned} \sqrt{n}\tilde{G}_n(\mu, \xi) &= \frac{1}{2\sqrt{n}} \sum_{i=1}^n \mathbb{E}\Big\{G_{i\ell r,*}(\mu, \xi) + G_{ir\ell,*}(\mu, \xi) + G_{\ell ir,*}(\mu, \xi) \\ &\quad + G_{\ell ri,*}(\mu, \xi) + G_{ri\ell,*}(\mu, \xi) + G_{r\ell i,*}(\mu, \xi)\big|O_i\Big\} \end{aligned}$$

for $\ell \neq i$ and $r \neq \ell, r \neq i$. For the maximum likelihood estimator $\hat{\xi}$, we can write

$$\sqrt{n}\left(\hat{\xi} - \xi_0\right) = \frac{1}{\sqrt{n}}\left[-\left.\frac{\partial \mathbb{E}\left\{\mathcal{S}_i(\xi)\right\}}{\partial \xi^\top}\right|_{\xi=\xi_0}\right]^{-1} \sum_{i=1}^n \mathcal{S}_i(\xi_0) + o_p(1) = \frac{1}{\sqrt{n}}\mathcal{I}(\xi)^{-1} \sum_{i=1}^n \mathcal{S}_i(\xi_0) + o_p(1).$$

Hence, from (5.16),

$$\begin{aligned}
\theta_1\theta_2\theta_3\sqrt{n}(\hat{\mu}_* - \mu_0) &= o_p(1) + \frac{1}{\sqrt{n}}\frac{\partial e_*^\top(\mu_0,\xi)}{\partial \xi}\bigg|_{\xi=\xi_0} \mathcal{I}(\xi)^{-1}\sum_{i=1}^n \mathcal{S}_i(\xi_0) \\
&\quad + \frac{1}{2\sqrt{n}}\sum_{i=1}^n \mathbb{E}\Big\{ G_{i\ell r,*}(\mu_0,\xi_0) + G_{ir\ell,*}(\mu_0,\xi_0) + G_{\ell ir,*}(\mu_0,\xi_0) \\
&\quad + G_{\ell ri,*}(\mu_0,\xi_0) + G_{ri\ell,*}(\mu_0,\xi_0) + G_{r\ell i,*}(\mu_0,\xi_0)\big|O_i\Big\} \\
&= o_p(1) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\frac{\partial e_*^\top(\mu_0,\xi)}{\partial \xi}\bigg|_{\xi=\xi_0} \mathcal{I}(\xi)^{-1}\mathcal{S}_i(\xi_0)\right. \\
&\quad + \frac{1}{2}\mathbb{E}\Big\{ G_{i\ell r,*}(\mu_0,\xi_0) + G_{ir\ell,*}(\mu_0,\xi_0) + G_{\ell ir,*}(\mu_0,\xi_0) \\
&\quad \left. + G_{\ell ri,*}(\mu_0,\xi_0) + G_{ri\ell,*}(\mu_0,\xi_0) + G_{r\ell i,*}(\mu_0,\xi_0)\big|O_i\Big\}\right] \qquad (5.17) \\
&= o_p(1) + \frac{1}{\sqrt{n}}\sum_{i=1}^n Q_{i,*}(\mu_0,\xi_0) = o_p(1) + \frac{1}{\sqrt{n}}Q_*(\mu_0,\xi_0).
\end{aligned}$$

Note that the observed data O_i are i.i.d, then $Q_{i,*}(\mu_0,\xi_0)$ are also i.i.d.. In addition, we easily show that

$$\begin{aligned}
0 &= \mathbb{E}\Bigg[\mathbb{E}\Big\{ G_{i\ell r,*}(\mu_0,\xi_0) + G_{ir\ell,*}(\mu_0,\xi_0) + G_{\ell ir,*}(\mu_0,\xi_0) + G_{\ell ri,*}(\mu_0,\xi_0) \\
&\quad + G_{ri\ell,*}(\mu_0,\xi_0) + G_{r\ell i,*}(\mu_0,\xi_0)\big|O_i\Big\}\Bigg].
\end{aligned}$$

Therefore, $\mathbb{E}\{Q_{i,*}(\mu_0,\xi_0)\} = 0$, and $\frac{1}{\sqrt{n}}Q_*(\mu_0,\xi_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}\text{ar}\{Q_{i,*}(\mu_0,\xi_0)\})$ by the Central Limit Theorem. It follows that

$$\sqrt{n}\,(\hat{\mu}_* - \mu_0) \xrightarrow{d} \mathcal{N}\,(0, \Lambda_*)\,,$$

where

$$\Lambda_* = \frac{\mathbb{V}\text{ar}\,\{Q_{i,*}(\mu_0,\xi_0)\}}{\theta_1^2\theta_2^2\theta_3^2}.$$

□

It is worth noting that the assumed regularity conditions for the likelihood and condition (C1)–(C3) hold in our working model, which is based on (5.1), with $h(T,A;\tau_\pi) = \tau_{\pi_1} + \tau_{\pi_2}T + A^\top\tau_{\pi_3}$, and (5.2), with $f(T,A;\tau_{\rho_k}) = \tau_{\rho_{1k}} + \tau_{\rho_{2k}}T + A^\top\tau_{\rho_{3k}}$.

5.2.3 Variance estimation

Under condition (C3), a consistent estimator of Λ_* can be obtained as

$$\hat{\Lambda}_* = \frac{\mathbb{V}\text{ar}\left\{\hat{Q}_{i,*}(\hat{\mu}_*,\hat{\xi})\right\}}{\hat{\theta}_{1,*}^2\hat{\theta}_{2,*}^2\hat{\theta}_{3,*}^2} = \frac{\frac{1}{n-1}\sum_{i=1}^n \hat{Q}_{i,*}^2(\hat{\mu}_*,\hat{\xi})}{\hat{\theta}_{1,*}^2\hat{\theta}_{2,*}^2\hat{\theta}_{3,*}^2}, \qquad (5.18)$$

where $\hat{\theta}_{k,*}$ are the estimates of the disease prevalence, θ_k for $k = 1,2,3$. Specifically, $\hat{\theta}_{k,\text{FI}} = \frac{1}{n}\sum_{i=1}^n \hat{\rho}_{ki}$, $\hat{\theta}_{k,\text{MSI}} = \frac{1}{n}\sum_{i=1}^n \tilde{D}_{ki,\text{MSI}}$, $\hat{\theta}_{k,\text{IPW}} = \sum_{i=1}^n V_i D_{ki}\hat{\pi}_i^{-1} / \sum_{i=1}^n V_i\hat{\pi}_i^{-1}$ and $\hat{\theta}_{k,\text{PDR}} = \frac{1}{n}\sum_{i=1}^n \tilde{D}_{ki,\text{PDR}}$.

According to (5.17), we have that

$$\begin{aligned}\hat{Q}_{i,*}(\hat{\mu}_*,\hat{\xi}) &= \left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G_{i\ell r,*}^{\top}(\hat{\mu}_*,\xi)}{\partial\xi}\right|_{\xi=\hat{\xi}}\right\}\left\{-\sum_{i=1}^{n}\left.\frac{\partial\mathcal{S}_i(\xi)}{\partial\xi^{\top}}\right|_{\xi=\hat{\xi}}\right\}^{-1}\mathcal{S}_i(\hat{\xi})\\ &+\frac{1}{2(n-1)(n-2)}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{i\ell r,*}(\hat{\mu}_*,\hat{\xi})+G_{ir\ell,*}(\hat{\mu}_*,\hat{\xi})+G_{\ell ir,*}(\hat{\mu}_*,\hat{\xi})\\ &+G_{\ell ri,*}(\hat{\mu}_*,\hat{\xi})+G_{ri\ell,*}(\hat{\mu}_*,\hat{\xi})+G_{r\ell i,*}(\hat{\mu}_*,\hat{\xi})\Big\}.\end{aligned}$$

In addition, for fixed i, we also have that

$$\begin{aligned}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{i\ell r,*}(\hat{\mu}_*,\hat{\xi})+G_{ir\ell,*}(\hat{\mu}_*,\hat{\xi})\Big\} &= 2\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}G_{i\ell r,*}(\hat{\mu}_*,\hat{\xi}),\\ \sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{\ell ir,*}(\hat{\mu}_*,\hat{\xi})+G_{ri\ell,*}(\hat{\mu}_*,\hat{\xi})\Big\} &= 2\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}G_{\ell ir,*}(\hat{\mu}_*,\hat{\xi}),\\ \sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{\ell ri,*}(\hat{\mu}_*,\hat{\xi})+G_{r\ell i,*}(\hat{\mu}_*,\hat{\xi})\Big\} &= 2\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}G_{r\ell i,*}(\hat{\mu}_*,\hat{\xi}).\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{Q}_{i,*}(\hat{\mu}_*,\hat{\xi}) &= \left\{\frac{1}{(n-1)(n-2)}\sum_{i=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq\ell, r\neq i}}^{n}\left.\frac{\partial G_{i\ell r,*}^{\top}(\hat{\mu}_*,\xi)}{\partial\xi}\right|_{\xi=\hat{\xi}}\right\}\left\{-\sum_{i=1}^{n}\left.\frac{\partial\mathcal{S}_i(\xi)}{\partial\xi^{\top}}\right|_{\xi=\hat{\xi}}\right\}^{-1}\mathcal{S}_i(\hat{\xi})\\ &+\frac{1}{(n-1)(n-2)}\sum_{\substack{\ell=1\\ \ell\neq i}}^{n}\sum_{\substack{r=1\\ r\neq i, r\neq\ell}}^{n}\Big\{G_{i\ell r,*}(\hat{\mu}_*,\hat{\xi})+G_{\ell ir,*}(\hat{\mu}_*,\hat{\xi})+G_{r\ell i,*}(\hat{\mu}_*,\hat{\xi})\Big\}. \quad (5.19)\end{aligned}$$

The quantity $\sum_{i=1}^{n}\left.\frac{\partial\mathcal{S}_i(\xi)}{\partial\xi^{\top}}\right|_{\xi=\hat{\xi}}$ could be obtained as the Hessian matrix of the log likelihood function at $\hat{\xi}$. In order to compute $\left.\frac{\partial G_{i\ell r,*}(\hat{\mu}_*,\xi)}{\partial\xi}\right|_{\xi=\hat{\xi}}$, we have to get the derivatives of $\frac{\partial}{\partial\xi}\rho_{ki}(\tau_{0\rho_k})$, $\frac{\partial}{\partial\xi}\rho_{k(0)i}(\xi)$, $\frac{\partial}{\partial\xi}\pi_i^{-1}(\lambda,\tau_\pi)$, $\frac{\partial}{\partial\xi}\pi_{10i}(\lambda,\tau_\pi)$, $\frac{\partial}{\partial\xi}\pi_{01i}(\lambda,\tau_\pi)$ and $\frac{\partial}{\partial\xi}\pi_{00i}(\lambda,\tau_\pi)$.

In Section 5.1.2, we obtain

$$\begin{aligned}\tfrac{\partial}{\partial\lambda_1}\pi_{10i}(\lambda,\tau_\pi) &= \pi_{10i}(1-\pi_{10i}); & \tfrac{\partial}{\partial\lambda_2}\pi_{10i}(\lambda,\tau_\pi) &= 0;\\ \tfrac{\partial}{\partial\lambda_1}\pi_{01i}(\lambda,\tau_\pi) &= 0; & \tfrac{\partial}{\partial\lambda_2}\pi_{01i}(\lambda,\tau_\pi) &= \pi_{01i}(1-\pi_{01i});\\ \tfrac{\partial}{\partial\lambda_1}\pi_{00i}(\lambda,\tau_\pi) &= 0; & \tfrac{\partial}{\partial\lambda_2}\pi_{00i}(\lambda,\tau_\pi) &= 0.\end{aligned}$$

and

$$\frac{\partial}{\partial\tau_\pi}\pi_{d_1d_2i} = U_i(1-\pi_{d_1d_2i})\pi_{d_1d_2i},$$

where (d_1,d_2) belongs to the set $\{(1,0),(0,1),(0,0)\}$. Also, we have

$$\begin{aligned}\tfrac{\partial}{\partial\tau_{\rho_1}}\rho_{1i}(\tau_\rho) &= U_i\rho_{1i}(1-\rho_{1i}); & \tfrac{\partial}{\partial\tau_{\rho_2}}\rho_{1i}(\tau_\rho) &= -U_i\rho_{1i}\rho_{2i};\\ \tfrac{\partial}{\partial\tau_{\rho_2}}\rho_{2i}(\tau_\rho) &= U_i\rho_{2i}(1-\rho_{2i}); & \tfrac{\partial}{\partial\tau_{\rho_1}}\rho_{2i}(\tau_\rho) &= -U_i\rho_{1i}\rho_{2i}.\end{aligned}$$

Moreover,

$$\frac{\partial}{\partial\lambda_s}\pi_i^{-1}(\lambda,\tau_\pi) = -D_{si}\frac{1-\pi_i}{\pi_i}; \qquad \frac{\partial}{\partial\tau_\pi}\pi_i^{-1}(\lambda,\tau_\pi) = -U_i\frac{1-\pi_i}{\pi_i},$$

with $s = 1, 2$. Then, recall that

$$\begin{aligned}
\rho_{1(0)i} &= \frac{(1-\pi_{10i})\rho_{1i}}{(1-\pi_{10i})\rho_{1i} + (1-\pi_{01i})\rho_{2i} + (1-\pi_{00i})\rho_{3i}},\\
\rho_{2(0)i} &= \frac{(1-\pi_{01i})\rho_{2i}}{(1-\pi_{10i})\rho_{1i} + (1-\pi_{01i})\rho_{2i} + (1-\pi_{00i})\rho_{3i}},\\
\rho_{3(0)i} &= \frac{(1-\pi_{00i})\rho_{3i}}{(1-\pi_{10i})\rho_{1i} + (1-\pi_{01i})\rho_{2i} + (1-\pi_{00i})\rho_{3i}}.
\end{aligned}$$

After some algebra, we get

$$\begin{aligned}
\frac{\partial}{\partial\lambda_1}\rho_{1(0)i}(\xi) &= \frac{1}{z^2}\left[-\pi_{10i}(1-\pi_{10i})\rho_{1i}\left\{(1-\pi_{01i})\rho_{2i} + (1-\pi_{00i})\rho_{3i}\right\}\right],\\
\frac{\partial}{\partial\lambda_2}\rho_{1(0)i}(\xi) &= \frac{1}{z^2}\rho_{1i}\rho_{2i}\pi_{01i}(1-\pi_{01i})(1-\pi_{10i}),\\
\frac{\partial}{\partial\tau_\pi}\rho_{1(0)i}(\xi) &= -\frac{U_i}{z^2}\rho_{1i}(1-\pi_{10i})\left\{\rho_{2i}(1-\pi_{01i})(\pi_{10i}-\pi_{01i}) + \rho_{3i}(1-\pi_{00i})(\pi_{10i}-\pi_{00i})\right\},\\
\frac{\partial}{\partial\tau_{\rho_1}}\rho_{1(0)i}(\xi) &= \frac{U_i}{z^2}\rho_{1i}(1-\pi_{10i})\left\{\rho_{2i}(1-\pi_{01i}) + \rho_{3i}(1-\pi_{00i})\right\},\\
\frac{\partial}{\partial\tau_{\rho_2}}\rho_{1(0)i}(\xi) &= -\frac{U_i}{z^2}\rho_{1i}\rho_{2i}(1-\pi_{10i})(1-\pi_{01i}).
\end{aligned}$$

Finally, we set $z = (1-\pi_{10i})\rho_{1i} + (1-\pi_{01i})\rho_{2i} + (1-\pi_{00i})\rho_{3i}$, and get

$$\begin{aligned}
\frac{\partial}{\partial\lambda_1}\rho_{2(0)i}(\xi) &= \frac{1}{z^2}\rho_{1i}\rho_{2i}\pi_{10i}(1-\pi_{10i})(1-\pi_{01i}),\\
\frac{\partial}{\partial\lambda_2}\rho_{2(0)i}(\xi) &= \frac{1}{z^2}\left[-\pi_{01i}(1-\pi_{01i})\rho_{2i}\left\{(1-\pi_{10i})\rho_{1i} + (1-\pi_{00i})\rho_{3i}\right\}\right],\\
\frac{\partial}{\partial\tau_\pi}\rho_{2(0)i}(\xi) &= -\frac{U_i}{z^2}\rho_{2i}(1-\pi_{01i})\left\{\rho_{1i}(1-\pi_{10i})(\pi_{01i}-\pi_{10i}) + \rho_{3i}(1-\pi_{00i})(\pi_{01i}-\pi_{00i})\right\},\\
\frac{\partial}{\partial\tau_{\rho_1}}\rho_{2(0)i}(\xi) &= -\frac{U_i}{z^2}\rho_{1i}\rho_{2i}(1-\pi_{10i})(1-\pi_{01i}),\\
\frac{\partial}{\partial\tau_{\rho_2}}\rho_{2(0)i}(\xi) &= \frac{U_i}{z^2}\rho_{2i}(1-\pi_{01i})\left\{\rho_{1i}(1-\pi_{10i}) + \rho_{3i}(1-\pi_{00i})\right\}.
\end{aligned}$$

The derivative $\frac{\partial}{\partial\xi}\rho_{3(0)i}(\xi)$ can be computed by using the fact that $\rho_{3(0)i} = 1 - \rho_{1(0)i} - \rho_{2(0)i}$.

5.3 Simulation studies

In this section, we provide empirical evidence, through simulation experiments, on the behavior of the proposed VUS estimators in finite samples. The number of replications in each simulation experiment is set to be 1000.

In the study, we consider two scenarios which correspond to quite different values of the true VUS. For both scenarios, we fix three sample sizes: 250, 500 and 1500.

In the first scenario, for each unit, we generate the test result T_i and a covariate A_i from a bivariate normal distribution,

$$(T_i, A_i) \sim \mathcal{N}_2\left(\begin{pmatrix}3.7\\1.85\end{pmatrix}, \begin{pmatrix}3.71 & 1.36\\1.36 & 3.13\end{pmatrix}\right).$$

The disease status $\mathcal{D}_i$ is generated according to model (5.2) with $f(T, A; \tau_{\rho_1}) = 4.6 - 3.3T - 6.4A$ and $f(T, A; \tau_{\rho_2}) = 4 - 1.7T - 3.2A$. Then, the verification label V_i is obtained according to model (5.1) with $h(T, A; \tau_\pi) = 1 + 1.2T - 1.5A$ and $\lambda_1 = -2, \lambda_2 = -1$. Under such data generating process, $\theta_1 = 0.4$, $\theta_2 = 0.35$, $\theta_3 = 0.25$, and the verification rate is roughly 0.57. The true VUS value is 0.791. In the second scenario, we generate the test result and the covariate from independent normal distributions. Specifically, $T_i \sim \mathcal{N}(0.65, 1)$ and $A_i \sim \mathcal{N}(-0.3, 0.64)$. The disease status $\mathcal{D}_i$ is generated according to model (5.2) with $f(T, A; \tau_{\rho_1}) = 4.6 - 3.3T - 6.4A$ and $f(T, A; \tau_{\rho_2}) = 4 - 1.7T - 3.2A$. Then, V_i is obtained according to model (5.1) with $h(T, A; \tau_\pi) = 1 + 1.2T - 1.5A$ and $\lambda_1 = -2.5, \lambda_2 = -1$. Under this setting, $\theta_1 = 0.55$, $\theta_2 = 0.32$, $\theta_3 = 0.13$, and the verification rate is roughly 0.58. The true VUS value is 0.387.

Table 5.1: Monte Carlo means (MCmean), relative bias (Bias), Monte Carlo standard deviations (MCds) and estimated standard deviations (Esd) for the proposed VUS estimators, and the SPE estimator under MAR assumption. CP denotes Monte Carlo coverages for the 95% confidence intervals, obtained through the normal approximation approach applied to each estimator.

	Sample size	Estimator	Mean	Bias(%)	MCsd	SE	CP (%)
Scenario I: TRUE = 0.791	$n = 250$	FI	0.772	−2.4	0.056	0.050	89.9
		MSI	0.770	−2.7	0.057	0.051	90.6
		IPW	0.770	−2.6	0.070	0.061	88.1
		PDR	0.766	−3.2	0.085	0.075	90.8
		SPE (MAR)	0.771	−2.5	0.073	0.138	93.2
	$n = 500$	FI	0.783	−1.0	0.035	0.032	93.3
		MSI	0.782	−1.1	0.036	0.033	93.4
		IPW	0.782	−1.2	0.047	0.042	92.2
		PDR	0.782	−1.2	0.053	0.058	94.0
		SPE (MAR)	0.771	−2.6	0.047	0.040	93.0
	$n = 1500$	FI	0.790	−0.2	0.016	0.016	95.0
		MSI	0.789	−0.2	0.016	0.016	95.2
		IPW	0.788	−0.3	0.025	0.024	94.4
		PDR	0.789	−0.3	0.025	0.024	95.2
		SPE (MAR)	0.771	−2.5	0.027	0.025	89.4
Scenario II: TRUE = 0.387	$n = 250$	FI	0.368	−5.0	0.064	0.057	87.4
		MSI	0.367	−5.2	0.065	0.059	87.9
		IPW	0.377	−2.6	0.084	0.074	87.6
		PDR	0.369	−4.6	0.086	0.075	89.5
		SPE (MAR)	0.346	−10.6	0.063	0.058	84.5
	$n = 500$	FI	0.379	−2.0	0.045	0.041	90.9
		MSI	0.379	−2.1	0.046	0.042	91.3
		IPW	0.380	−1.8	0.060	0.056	91.2
		PDR	0.381	−1.6	0.060	0.053	92.0
		SPE (MAR)	0.345	−10.8	0.044	0.042	76.5
	$n = 1500$	FI	0.388	0.2	0.023	0.022	94.2
		MSI	0.388	0.2	0.023	0.023	94.3
		IPW	0.388	0.3	0.034	0.032	94.9
		PDR	0.389	0.4	0.033	0.029	93.2
		SPE (MAR)	0.346	−10.7	0.026	0.025	76.5

Table 5.1 contains Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations for the proposed VUS estimators (FI, MSI, IPW, PDR) in the two considered

scenarios, at the chosen sample sizes. The table also reports the empirical coverages of the 95% confidence intervals for the VUS, obtained through the normal approximation approach applied to each estimator. To make a comparison, Table 5.1 also gives the results for the SPE discussed in Section 4.1, whose realizations are obtained, in all experiments, under the MAR assumption, i.e., by setting $\lambda_1 = \lambda_2 = 0$ in model (5.1). The comparison allows us to evaluate the possible impact of an incorrect hypothesis MAR on the most robust estimator among those, FI, MSI, IPW and SPE, which are built to work under ignorable missing data mechanism.

Overall, simulation results are consistent with our theoretical findings and show the usefulness of the proposed estimators, which also arises from the comparison with the SPE estimator used improperly. The results also show a good behavior of the estimated standard deviations, which are generally close to the corresponding Monte Carlo values. In general, FI and MSI estimators seem to be more efficient than IPW and PDR estimators. However, for all estimators, acceptable bias levels and sufficiently accurate associated confidence intervals seem to require a large sample size (at least 500, and, prudently, even higher).

This issue of poor accuracy has already been noted by several authors, including Liu and Zhou (2010), in the context of two-class classification problems. In our experience, the trouble appears to arise because of a bad behavior of the maximum likelihood estimates in the verification and disease models. If the sample size is not large enough, the data do not contain enough information to effectively estimate the parameters λ, τ_π, τ_{ρ_1} and τ_{ρ_2}. It seems particularly difficult to get good estimates of nonignorable parameters.

Table 5.2: Monte Carlo means (MCmean) for the maximum likelihood estimators of the elements of nuisance parameters λ, τ_π, τ_{ρ_1} and τ_{ρ_2}.

	Scenario I				Scenario II			
	True	Monte Carlo Mean			True	Monte Carlo Mean		
		250	500	1500		250	500	1500
λ_1	-2.00	-1.01	-1.76	-1.95	-2.50	-2.09	-2.30	-2.50
λ_2	-1.00	-0.45	-0.87	-0.98	-1.00	-0.99	-0.96	-0.97
τ_{π_1}	2.00	1.25	1.80	1.95	1.00	1.17	1.00	1.00
τ_{π_2}	0.50	0.65	0.55	0.51	1.20	1.39	1.28	1.22
τ_{π_3}	-1.20	-1.24	-1.22	-1.21	-1.50	-1.25	-1.40	-1.51
$\tau_{\rho_{11}}$	15.00	15.53	15.28	15.10	4.60	4.44	4.58	4.66
$\tau_{\rho_{21}}$	-3.30	-3.41	-3.36	-3.32	-3.30	-3.29	-3.33	-3.34
$\tau_{\rho_{31}}$	-0.70	-0.89	-0.78	-0.72	-6.40	-6.94	-6.70	-6.48
$\tau_{\rho_{12}}$	9.50	10.03	9.71	9.57	4.00	4.12	4.11	4.05
$\tau_{\rho_{22}}$	-1.70	-1.79	-1.73	-1.71	-1.70	-1.77	-1.76	-1.73
$\tau_{\rho_{32}}$	-0.30	-0.40	-0.34	-0.31	-3.20	-3.62	-3.42	-3.25

Table 5.2, giving the Monte Carlo means for the maximum likelihood estimators of the elements of λ, τ_π, τ_{ρ_1} and τ_{ρ_2}, for the three considered sample sizes, allows us to look at the bias of the estimators. More importantly, Figure 5.1 and Figure 5.2 (which refer to scenario I and II, respectively) graphically depict values of the estimates of λ_1, λ_2, and τ_{π_1} obtained in the thousand replications, for each sample size. The plots clearly show the great variability of the maximum likelihood estimates at lower sample sizes, with many values dramatically different from the corresponding target values. With larger sample size, this phenomenon almost completely vanishes, the maximum likelihood estimators behave pretty well, with a positive impact on the behavior of

the VUS estimators.

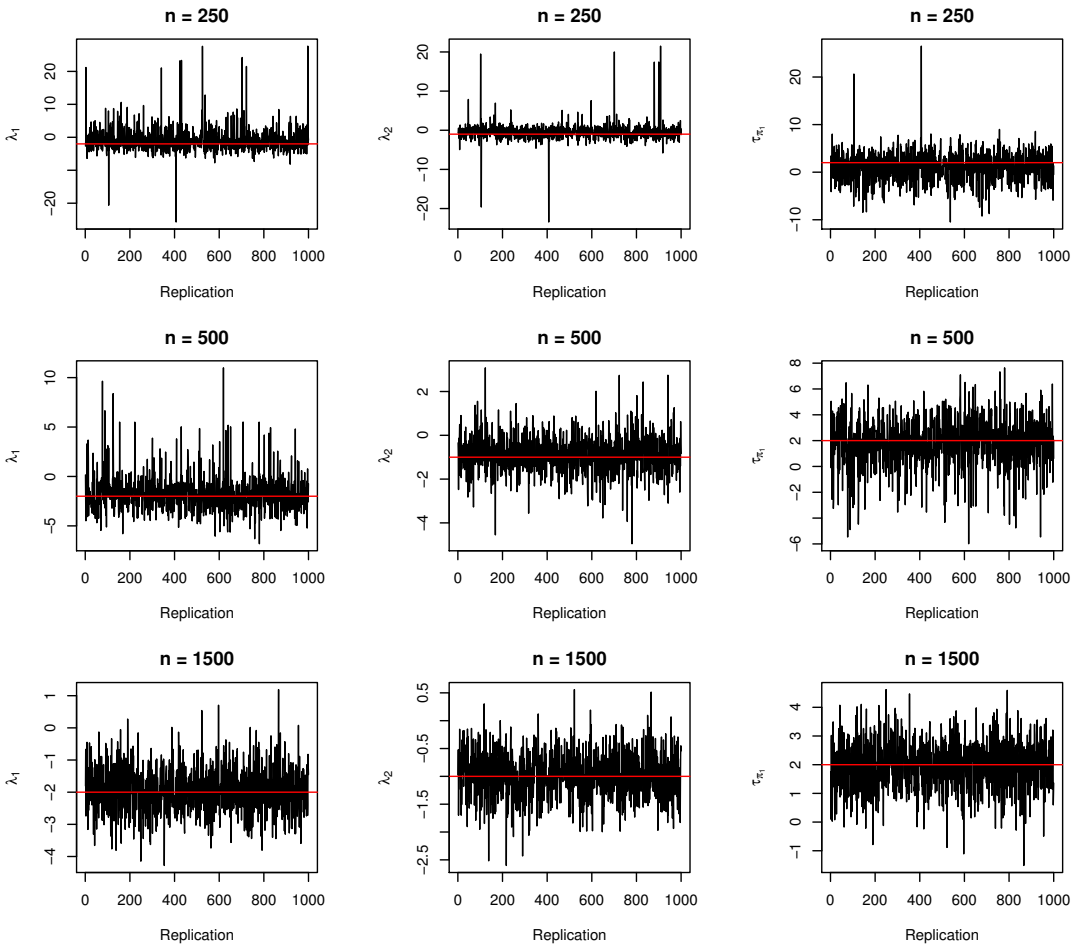


Figure 5.1: The plot of the MLE of $(\lambda_1, \lambda_2, \tau_{\pi_1})$ with respect to scenario I.

5.4 Discussion

In this chapter, we have proposed four bias–corrected estimators of VUS under NI missing data mechanism. The estimators are obtained by a likelihood–based approach, which uses the verification model (5.1) together with the disease model (5.2). The identifiability of the joint model is proved, and hence, the nuisance parameters can be estimated by maximizing the log–likelihood function or solving the score equations. Consistency and asymptotic normality of the proposed FI, MSI, IPW and PDR estimators are established, and variance estimation is discussed.

The proposed VUS estimators are pretty easy to implement and require the use of some numerical routine to maximize the log–likelihood function (or to solve the score equations). Our simulation results show their usefulness, whilst confirming the evidence emerging in the two–class case, according to which a reasonable large sample size is necessary to make sufficiently accurate inference. In practice, among FI, MSI, IPW and PDR estimators, we would reccommend FI and MSI estimators thanks to their greater efficiency.

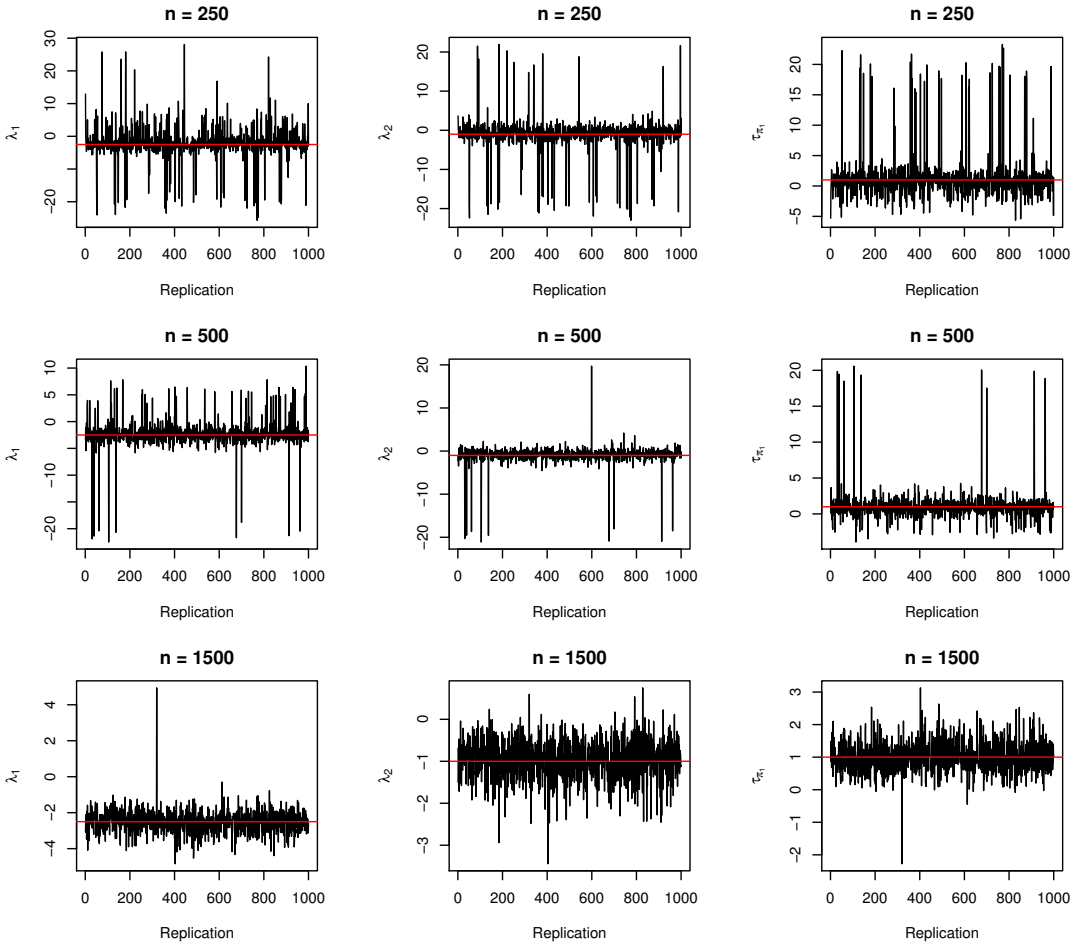


Figure 5.2: The plot of the MLE of $(\lambda_1, \lambda_2, \tau_{\pi_1})$ with respect to scenario II.

The poor accuracy problem seems to be related to an intrinsic difficulty of the maximum likelihood method in providing accurate estimates of the parameters of the disease and verification models, in particular of the nonignorable parameters. Overcoming this drawback is a stimulating challenge and deserves further investigation.

Chapter 6

R package: bcROCsurface

6.1 Introduction

In R, some packages exist for ROC surface analysis under full verification. For example, *DiagTest3Grp* gives some tools for estimating VUS (Luo and Xiong, 2012), *ROCS* deals with the high-throughput class-skewed data (Yu, 2012) and *HUM* provides tools for visualizing the ROC surface (Novoselova et al., 2014). No package is available at the moment for correcting for verification bias estimators of the ROC surface and VUS.

The R package *bcROCsurface* aims to fill in this gap by providing a significant number of new functions for bias-corrected ROC surface analysis. More precisely, it implements five bias-corrected estimators for ROC surface and VUS of a continuous diagnostic test, namely, full imputation (FI), mean score imputation (MSI), inverse probability weighting (IPW), semiparametric efficient (SPE) and K nearest-neighbor (KNN) estimators. These methods perform provided that the missing mechanism is MAR. Beside that, the package also works under full verification.

6.2 Package description

bcROCsurface imports various R packages (e.g., rgl, nnet, boot) and is built on Rcpp (Eddelbuettel, 2013). The package is freely available to download from https://CRAN.R-project.org/package=bcROCsurface, and provides a consistent set of functions for bias–corrected inference on VUS, for constructing and plotting 3D-ROC surfaces as well as ellipsoidal confidence regions of true class fractions at a given cut-point.

Practical use of the package foresees three steps: data preparation, modeling and inference.

Data preparation: In this step, the condition of monotone ordering of the three disease classes under study (Nakas and Yiannoutsos, 2004) is checked. The condition is mandatory to perform the subsequent analyses. In words, the condition assumes that subjects from class 3 have higher test results than subjects in class 2 and the latter have higher test results than subjects in class 1. Function `preDATA()` performs such checks, warning users in case monotone ordering is not satisfied. When satisfied, the function also generates a binary matrix with three columns, corresponding to the coding of the three classes of the disease status, used as input of the main functions.

Modeling: Correction for verification bias requires estimation of a disease and a verification model. The function `psglm()` obtains the verification probabilities specifying a general linear model for the verification process. In practice, the user can select among a logistic, a probit or a threshold regression model. Functions `rhoMLogit()` and `rhoKNN()` estimate the disease probabili-

ties based on a multinomial logistic regression. In particular, `rhoMLogit()` calls the `nnet` package for multinomial logistic modeling, whereas `rhoKNN()` uses K nearest-neighbor regression.

Inference: Two main functions are provided: `ROCs` for construction and plotting ROC surfaces, and `vus()` for estimating VUS values as well as obtaining confidence intervals. Estimation methods can be flexibly selected by the argument `method`, among 6 options. i.e., `method = "full"` if the full data is available, `fi`, `msi`, `ipw`, `spe` and `knn` in presence of verification bias. To plot ROC surfaces and ellipsoid confidence regions, the function `ROCs` employs the plotting functions of `rgl` package. `vus()` employs some core functions, written in the C++ language and integrated in R through the Rcpp and RcppArmadillo package. Confidence intervals of VUS are built based on the asymptotic distribution or the bootstrap resampling process (supported by the parallel computing). In addition, this function also performs the statistical test, H_0: VUS $= 1/6$ versus H_A: VUS $> 1/6$.

Besides the functions above described, the package also offers other functions for calculating asymptotic variances or determine the choice for K with respect to KNN methods.

In addition, we have also developed the Shiny web application (https://cran.r-project.org/web/packages=shiny) to provide the possibility to deploy `bcROCsurface` package over the web. The web interface can be found at https://khanhtoduc.shinyapps.io/bcROCsurface_shiny/.

6.3 Implementation

6.3.1 In R

To illustrate the use of the package, here an example is given. A full guide for use of the `bcROCsurface` package can be found in the vignette document.

In the following example, `ROCs()` is employed to build the bias-corrected ROC surface by SPE estimator. Data come from the study on epithelial ovarian cancer (EOC). This dataset is available in the package. As we mentioned above, in the beginning, the application of `preDATA()` is needed to ensure that the package can be employed. In second step, the functions `rhoMLogit()` and `psglm()` are called. The SPE estimated ROC surface presented in Figure 3.4(b) is the result of implementation of `ROCs()`. In addition, this figure shows a ellipsoidal confidence region (with green color) of the true class fractions at the cut point $(-0.56, 2.31)$. Here is `R` code:

```
> library(bcROCsurface)
# load and attach the dataset
> data(EOC)
# Preparing the missing disease status
> dise <- preDATA(EOC$D, EOC$CA125)
> dise.gpr <- dise$D
> dise.mat <- dise$Dvec
# Estimate the disease probabilities
> rho.out <- rhoMLogit(dise.gpr ~ CA125 + CA153 + Age, data = EOC)
# Estimate the verification probabilites
> pi.out <- psglm(V ~ CA125 + CA153 + Age, data = EOC)
# Build bias-corrected ROC surface
> ROCs("spe", T = EOC$CA125, Dvec = dise.mat, V = EOC$V, rhoEst = rho.out, piEst = pi.out,
       ellipsoid = TRUE, cpst = c(-0.56, 2.31))
```

6.3.2 The web interface

The layout of the `bcROCsurface` web interface is clean and straightforward (Figure 6.1). It provides the possibility to load the datasets for the analysis and to access all functions of `bcROCsurface`

package. Here, user loads a data file (typically, csv, txt or dat file), selects the suitable option for "Separator" and "Quote" to read data correctly, then choose the input variables, i.e. diagnostic test, disease status. If the true disease status is not missing, user follows step 1 and 2 to get the results. Otherwise, user clicks on the square box and selects the verification status, then follows step 1, 2 and 3 to implement the bias-corrected ROC surface analysis.

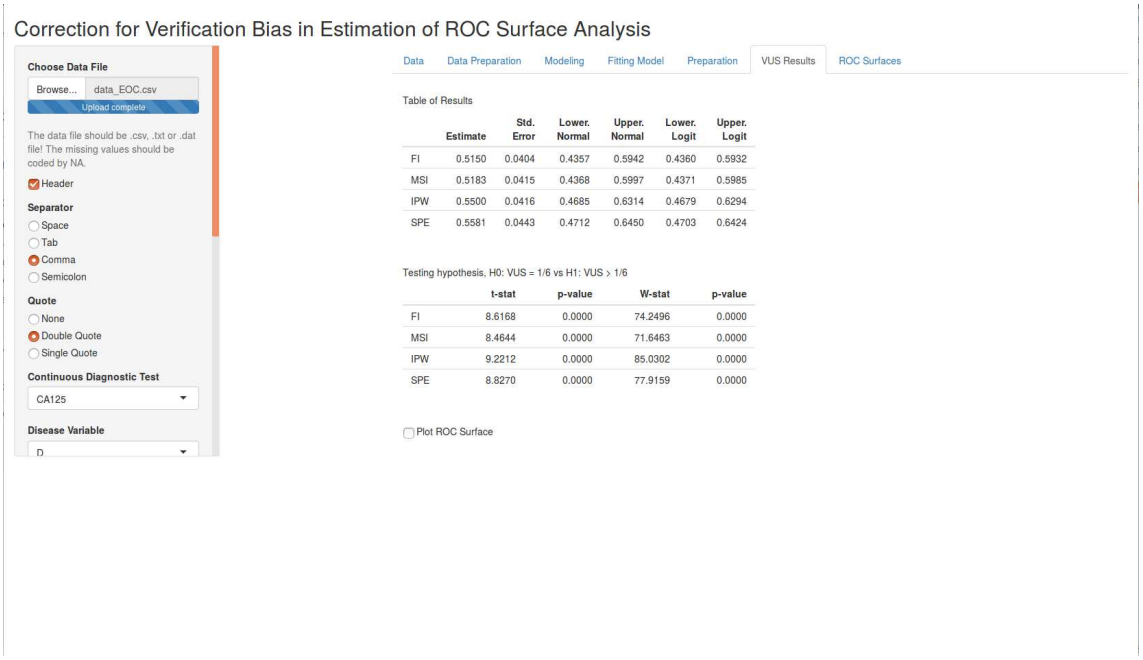


Figure 6.1: Bias-corrected VUS in Shiny application.

Chapter 7

Conclusions

The ROC surfaces and in particular the VUSs, are widely used to examine the effectiveness of diagnostic tests. In practice, however, the estimation of ROC surfaces and their volume underneath may be badly biased (verification bias) due to the missingness of the true disease status of the subjects. Therefore, the correction for verification bias is a problem of great importance. Here, we considered bias-corrected estimation of the ROC surface and VUS for a continuous diagnostic test under MAR and MNAR assumption.

In our approaches, we use the disease probabilities and/or verification probabilities to define the bias–corrected estimators for the ROC surface and VUS. Under MAR assumption, these probabilites are separately estimated via some parametric models (e.g., multinomial logistic; probit or logistic) or in a nonparametric framework. On the contrary, these probabilities are estimated together from likelihood function, under MNAR assumption.

There are still many open questions that deserve further investigation. Here, we mention some of them, which are closely related to the work presented in the thesis.

1. The consistency and asymptotic normality of the KNN estimator for VUS developed in Chapter 4 were not present. Therefore, studying the statistical properties of this estimator is the focus of our current work.

2. The proposed approaches require a specific monotone ordering for the three disease classes with respect to the test result. In other words, our methods are not applicable when an umbrella ordering is of interest. In absence of missing data, Nakas and Alonzo (2007) proposed a nonparametric framework for construction of an umbrella ROC graph and estimation of umbrella volume. Thus, further work is needed to extend our methods to the umbrella ordering.

3. Beside the ROC surface and VUS, there are several summary measures of diagnostic test performance and among them is the three-class Youden index (Nakas et al., 2010, 2013). This measure is frequently used in practice and not only shows the accuracy of a diagnostic test, but also provides a criterion for choosing an optimal cut point (c_1^*, c_2^*), the cut point for which the Youden index is maximized. In order to compute the three-class Youden index, the estimates of TCFs (the true class fractions) are needed, and hence, verification bias could impact also estimation of this index. The construction of bias–corrected estimators for the three-class Youden index should be investigated in future studies.

4. The bias–corrected estimators for the ROC surface and VUS in Chapter 3, 4 and 5 just con-

cern a single continuous diagnostic test. It is worth noting, however, that multiple diagnostic tests might be available in practice. In such situations, a combination of diagnostic test results could increase the accuracy. There are some papers discussed about how to choose the best optimal linear combination of biomarkers to maximize the AUC (Su and Liu, 1993; Pepe and Thompson, 2000; Pepe et al., 2006; Liu et al., 2011; Huang et al., 2011; Kang et al., 2016), the VUS (Zhang and Li, 2011; Kang et al., 2013) and the hypervolume under the ROC manifold (HUM, Hsu and Chen, 2016). However, all existing methods only work in absence of verification bias. Extensions of the available methods for AUC and VUS in presence of verification bias are future challenges.

5. In real applications, various covariates or explanatory variables (e.g. age, gender, race, marital status, etc.) are frequently present, and they may affect the accuracy of diagnostic tests and also their linear combination. Consequently the ROC curve, AUC, the ROC surface and VUS will change as a function of these covariate values. Several approaches for modeling the effect of the covariates on the ROC curve and AUC were proposed in the literature (Pepe, 1998; Schisterman et al., 2004; Zhou et al., 2009; Liu and Zhou, 2011; Fluss et al., 2012). Thus, further work is needed to propose some methods for the ROC surface and VUS.

6. Sometimes, in practice, the comparison of two or more diagnostic tests is necessary. This work could be done by comparing the values of AUC (in case of two classes) and of VUS (in case of three classes). However, Hand (2009, 2010); Hand and Anagnostopoulos (2013) noticed that the AUC has a well-known deficiency when it is used to compare crossed ROC curves, in the sense that the AUC could lead to a mistaken belief about the performance of the diagnostic test as it is actually used with some specific cut points. In addition, the AUC does not take into account the balance of different kinds of misdiagnoses effectively. To overcome these disadvantages, the H measure was proposed as a coherent alternative to the AUC, see Hand (2009, 2010); Hand and Anagnostopoulos (2014). It is worth noting that the VUS is a generalization of the AUC, and hence, the disadvantages of the AUC also appear in VUS. Therefore, a new definition of the H measure for the case of three disease status should be investigated in future work.

In Chapters 3, 4 and 5 of this thesis, the behavior of the proposed estimators were analyzed by means of simulations. It should be recalled that studies of this kind, although certainly useful, can only be partial, since they cannot cover all the endless scenarios that reality can manifest. The results reported by us provide some guidance, but the behavior of the techniques proposed and analyzed can be very different in contexts other than those considered in studies. For example:

(i) actually, small sample sizes can have a negative impact on all estimators, especially the nonparametric ones.

(ii) for fixed sample sizes, a large number of covariates in the involved models can produce inaccurate estimates;

(iii) poor results of the partially parametric approaches and the nonparametric estimator can be observed when the distributions of diagnostic test and/or covariates are skewed (for example, gamma distribution) or mixture distributions;

(iv) for small sample sizes, all estimators can yield unsatisfactory performance when the verification rate is small or unbalanced with respect to the covariates distribution.

Bibliography

Adimari, G. and Chiogna, M. (2015). Nearest-neighbor estimation for ROC analysis under verification bias. *The International Journal of Biostatistics*, 11(1):109–124.

Adimari, G. and Chiogna, M. (2016). Nonparametric verification bias–corrected inference for the area under the ROC curve of a continuous–scale diagnostic test. *Statistics and Its Interface*, In Press.

Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):173–190.

Alonzo, T. A., Pepe, M. S., and Lumley, T. (2003). Estimating disease prevalence in two-phase studies. *Biostatistics*, 4(2):313–326.

Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics*, pages 330–337.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of mathematical psychology*, 12(4):387–415.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*, pages 207–215.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87.

Chi, Y. Y. and Zhou, X. H. (2008). Receiver operating characteristic surfaces in the presence of verification bias. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 57(1):1–23.

Daganzo, C. (1979). *Multinomial probit: the theory and its application to demand forecasting*. Elsevier.

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

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer.

Fluss, R., Reiser, B., and Faraggi, D. (2012). Adjusting ROC curve for covariates in the presence of verification bias. *Journal of Statistical Planning and Inference*, 142(1):1–11.

Fluss, R., Reiser, B., Faraggi, D., and Rotnitzky, A. (2009). Estimation of the ROC curve under verification bias. *Biometrical Journal*, 51(3):475–490.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine learning*, 77(1):103–123.

Hand, D. J. (2010). Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. *Statistics in medicine*, 29(14):1502–1510.

Hand, D. J. and Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance? *Pattern Recognition Letters*, 34(5):492–495.

Hand, D. J. and Anagnostopoulos, C. (2014). A better beta for the h measure of classification performance. *Pattern Recognition Letters*, 40:41–46.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

He, H., Lyness, J. M., and McDermott, M. P. (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in medicine*, 28(3):361–376.

He, H. and McDermott, M. P. (2012). A robust method using propensity score stratification for correcting verification bias for binary tests. *Biostatistics*, 13(1):32–47.

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

Hsu, M.-J. and Chen, Y.-H. (2016). Optimal linear combination of biomarkers for multi-category diagnosis. *Statistics in medicine*, 35(2):202–213.

Huang, X., Qin, G., and Fang, Y. (2011). Optimal combinations of diagnostic tests based on AUC. *Biometrics*, 67(2):568–576.

Kang, L., Liu, A., and Tian, L. (2016). Linear combination methods to improve diagnostic/prognostic accuracy on future observations. *Statistical Methods in Medical Research*, 25(4):1359–1380.

Kang, L. and Tian, L. (2013). Estimation of the volume under the ROC surface with three ordinal diagnostic categories. *Computational Statistics & Data Analysis*, 62:39–51.

Kang, L., Xiong, C., Crane, P., and Tian, L. (2013). Linear combinations of biomarkers to improve diagnostic accuracy with three ordinal diagnostic categories. *Statistics in medicine*, 32(4):631–643.

Kosinski, A. S. and Barnhart, H. X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*, 59(1):163–171.

Leichtle, A. B., Ceglarek, U., Weinert, P., Nakas, C. T., Nuoffer, J.-M., Kase, J., Conrad, T., Witzigmann, H., Thiery, J., and Fiedler, G. M. (2013). Pancreatic carcinoma, pancreatitis, and healthy controls: metabolite models in a three-class diagnostic dilemma. *Metabolomics*, 9(3):677–687.

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

Li, J., Zhou, X. H., and Fine, J. P. (2012). A regression approach to ROC surface, with applications to Alzheimer's disease. *Science China Mathematics*, 55(8):1583–1595.

Liu, C., Liu, A., and Halabi, S. (2011). A min–max combination of biomarkers to improve diagnostic accuracy. *Statistics in medicine*, 30(16):2005–2014.

Liu, D. and Zhou, X. H. (2010). A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood–based approach. *Biometrics*, 66(4):1119–1128.

Liu, D. and Zhou, X. H. (2011). Semiparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias. *Biometrics*, 67(3):906–916.

Luo, J. and Xiong, C. (2012). DiagTest3Grp: an R package for analyzing diagnostic tests with three ordinal groups. *Journal of statistical software*, 51(3):1.

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

Nakas, C. T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems. *REVSTAT-Statistical Journal*, 12(1):43–65.

Nakas, C. T. and Alonzo, T. A. (2007). ROC graphs for assessing the ability of a diagnostic marker to detect three disease classes with an umbrella ordering. *Biometrics*, 63(2):603–609.

Nakas, C. T., Alonzo, T. A., and Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Statistics in medicine*, 29(28):2946–2955.

Nakas, C. T., Dalrymple-Alford, J. C., Anderson, T., and Alonzo, T. A. (2013). Generalization of Youden index for multiple-class classification problems applied to the assessment of externally validated cognition in Parkinson disease screening. *Statistics in Medicine*, 32(6):995–1003.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, 23(22):3437–3449.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Ning, J. and Cheng, P. E. (2012). A comparison study of nonparametric imputation methods. *Statistics and Computing*, 22(1):273–285.

Novoselova, N., Della Beffa, C., Wang, J., Li, J., Pessler, F., and Klawonn, F. (2014). HUM calculator and HUM package for R: easy-to-use software tools for multicategory receiver operating characteristic analysis. *Bioinformatics*, 30(11):1635–1636.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, pages 124–135.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press.

Pepe, M. S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1):221–229.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2):123–140.

Ransohoff, D. F. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, 299(17):926–930.

Reilly, M. and Pepe, M. (1995). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, 82(2):299–314.

Ressom, H. W., Varghese, R. S., Goldman, L., Loffredo, C. A., Abdel-Hamid, M., Kyselova, Z., Mechref, Y., NOVOTNY, M., and Goldman, R. (2008). Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers. *Pacific Symposium on Biocomputing*, 13:216–227.

Roberts, G., Rao, J. N. K., and Kumar, S. (1987). Logistic regression analysis of sample survey data. *Biometrika*, 74(1):1–12.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101(475):1276–1288.

Schisterman, E. F., Faraggi, D., and Reiser, B. (2004). Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*, 23(21):3319–3331.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability. *Journal of Mathematical Psychology*, 40(3):253–269.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press.

Xiong, C., van Belle, G., Miller, J. P., and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups. *Statistics in Medicine*, 25(7):1251–1273.

Yu, T. (2012). ROCS: Receiver operating characteristic surface for class-skewed high-throughput data. *PloS ONE*, 7(7):e40598.

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

Zhou, X. H. and Castelluccio, P. (2003). Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. *Journal of Statistical Planning and Inference*, 115(1):193–213.

Zhou, X. H. and Castelluccio, P. (2004). Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*, 23(2):231–230.

Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine*. John Wiley & Sons.

Zhou, X. H. and Rodenberg, C. A. (1998). Estimating an ROC curve in the presence of nonignorable verification bias. *Communications in Statistics*, 27(3):273–285.

Khanh To Duc

CURRICULUM VITAE

Personal Details

Date of Birth: April 03, 1990
Place of Birth: Thanh Hoa, Vietnam
Nationality: Vietnamese

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 345 667 0711
e-mail: toduc@stat.unipd.it

Current Position

Since January 2014; (expected completion: April 2017)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Statistical evaluation of diagnostic tests under verification bias
Supervisor: Prof. Monica Chiogna
Co-supervisor: Prof. Gianfranco Adimari.

Research interests

- ROC surface analysis
- Applied statistics
- R package
- Missing data
- Verification bias

Education

October 2008 – July 2012
Bachelor degree (*laurea quadriennale***) in mathematics and computer sciences**.
University of Science, Vietnam National University - Ho Chi Minh city, Faculty of Mathematics and Computer Science
Title of dissertation: "The application of nonparametric regression model for the backward heat problems "
Supervisor: Prof. Trong Dang Duc
Final mark: 10/10.

Awards and Scholarship

2014 - 2016
CARIPARO PhD Scholarhip for foreign students, University of Padova.

2011
Awards for Scientific Research Student, rank 4th, University of Science, VNU - HCM.

2010 - 2012
Fellowship of honor program, University of Science, VNU - HCM.

Computer skills

- All Microsoft™ OS, Unix/Linux;
- C/C++, FORTRAN 90;
- Matlab, R, SPSS;
- LaTeX, Microsoft Office™.

Language skills

Vietnamese: native; English: good.

Publications

Articles in journals
To Duc, K., Chiogna, M. and Adimari, G. (2016). Bias-corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests. *Electronic Journal of Statistics* **10 2**, 3063–3113.

Trong D. D., Khanh T. D., Tuan N. H., Minh N. D. (2016). A two dimensional backward heat problem with statistical discrete data. *Submitted*. arXiv:1606.05463.

To Duc, K., Chiogna, M. and Adimari, G. (2016). Nonparametric estimation of ROC surfaces in presence of verification bias. *Submitted*. arXiv:1604.04656.

To Duc, K., Chiogna, M. and Adimari, G. (2016). Estimation for the volume under ROC surface in presence of nonignorable verification bias. *Submitted*. arXiv:1604.08805.

To Duc, K. (2016). bcROCsurface: an R package for correcting verification bias in estimation of the ROC surface and its volume for continuous diagnostic tests. *Submitted*.

Trong D. D., Khanh T. D., Tuan N. H., Minh N. D. (2015). Nonparametric regression in a statistical modified Helmholtz equation using Fourier spectral regularization. *Statistic: A Journal of Theoretical and Applied Statistics* **49 2**, 267–290.

R packages
bcROCsurface: Bias-corrected methods for estimating ROC surface and VUS for continuous diagnostic tests. CRAN.

Conference presentations

To Duc, K., Chiogna, M. and Adimari, G. (2016). Nonparametric estimation of ROC surfaces in presence of verification bias. (oral) *The 22nd International Conference on Computational Statistics*, Oviedo, Spain, 23-26 August.

To Duc, K., Chiogna, M. and Adimari, G. (2016). Bias-corrected estimation methods for receiver operating characteristic surface. (oral) *The 18th International Conference on Biometrics and Biostatistics*, Amsterdam, Netherlands, 4-5 August.

Khanh T. D., Trong D. D., Tuan N. H., Minh N. D. (2013). Nonparametric regression in a statistical modified Helmholtz equation using Fourier spectral regularization. (oral) *The 8th Congress Mathematical Vietnam*, Nha Trang, Viet Nam, 10-14 August.

Thang, P. L., Khanh T. D., Hoang, N. D. (2013). Some applications of the Plackett-Burman design for investigating optimal medium components with soya bean to grow Bacillus subtilis for protein production. (oral) *Statistics and its Interactions with Other Disciplines*, Ho Chi Minh city, Viet Nam, June.

Khanh T. D., Trong D. D., Tuan N. H., Minh N. D. (2013). A two-dimensional backward heat problems with random data. (poster). *Statistics and its Interactions with Other Disciplines*, Ho Chi Minh city, Viet Nam, June.

Teaching experience

October 2012 – December 2012
Mathematical statistics
Degree in mathematics
lab, 25 of hours
Instructor: Dr. Ha Hoang Van

February 2013 – May 2013
Analysis A2
Degree in mathematics
exercises, 25 of hours
Instructor: Dr. Tuan Nguyen Huy

October 2013 – December 2013
Measure theory and probability
Degree in mathematics
exercises, 25 of hours
Instructor: Prof. Trong Dang Duc

October 2013 – December 2013
Mathematical statistics
Degree in mathematics
lab, 25 of hours
Instructor: Dr. Ha Hoang Van

References

Prof. Monica Chiogna
Department of Statistical Sciences
University of Padova
via Cesare Battisti, 241-243
35121 Padova, Italy.
e-mail: monica@stat.unipd.it

Prof. Gianfranco Adimari
Department of Statistical Sciences
University of Padova
via Cesare Battisti, 241-243
35121 Padova, Italy.
e-mail: adimari@stat.unipd.it

Prof. Trong Dang Duc
Department of Mathematics and Computer Science
University of Science, Vietnam National University Ho Chi Minh city (VNU-HCM)
227 Nguyen Van Cu St., Ward 4, District 5
Ho Chi Minh City, Vietnam.
e-mail: ddtrong@hcmus.edu.vn