Chapter 5
Estimation of ROC Curve with Multiple Types of Missing Gold Standard

Danping Liu and Xiao-Hua Zhou

Abstract In evaluating the diagnostic accuracy of a test, the gold standard might be missing because of high cost or harmfulness to the patient. The estimation of the diagnostic accuracy could be biased if the missingness is not handled appropriately. In this chapter, we propose a likelihood-based approach to jointly estimate the selection model and disease model when the missing data mechanism is a mixture of ignorable and nonignorable missingness. The receiver operating characteristic (ROC) curve and its area are estimated empirically using imputation and reweighting techniques. The proposed method extends the results of Liu and Zhou (2010, Biometrics, 66, 1119–1128), as the latter assumes a single source of nonignorable missingness. We perform simulation studies to compare the performance of the proposed method and the existing approaches in the literature. This methodology is motivated from and applied to a study in Alzheimer's disease (AD), where two reasons of missingness are modeled separately.

5.1 Introduction

A medical diagnostic test is often evaluated by its sensitivity, specificity or the receiver operating characteristic (ROC) curve. Many methods to estimate the ROC curve require the true disease status to be verified without error, which is called " gold standard." However, the gold standard could be subject to missingness, because of high cost or harmfulness to the patient. Deleting the subjects with missing gold standard results in biased estimates of the ROC curve, known as "verification bias."

D. Liu (✉)
Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, Bethesda, MD 20892, USA
e-mail: danping.liu@nih.gov

X.-H. Zhou
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

Northwest HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98108, USA
e-mail: azhou@uw.edu

Z. Chen et al. (eds.), *Applied Statistics in Biomedicine and Clinical Trials Design*, ICSA Book Series in Statistics, DOI 10.1007/978-3-319-12694-4_5

Under ignorable missingness, or missing at random (MAR) assumption, existing methods to adjust for the verification bias include but are not limited to Begg and Greenes (1983), Begg (1987), Zhou (1996), Zhou (1998), Rodenberg and Zhou (2000), Alonzo and Pepe (2005), and Liu and Zhou (2011). The verification of gold standard may also be associated with some unobserved covariates related to the missing disease status. Hence, the MAR assumption may not hold. The nonignorable (NI)verification bias was first discussed by Baker (1995), and later developed by Zhou (1998), Kosinski and Barnhart (2003), Zhou and Castelluccio (2003), Zhou and Castelluccio (2004). Rotnitzky et al. (2006) proposed a "doubly robust" estimator for the area under ROC curve (AUC), but they specified a NI parameter (the log odds ratio of verification for diseased vs. healthy subject). Liu and Zhou (2010) considered a likelihood-based approach to estimate the NI parameter. Then the empirical AUC estimators were constructed using imputation or reweighting techniques.

Modeling missingness mechanism by a selection model is a key step in many existing methods for ROC analysis. As the NI missingness assumption is not testable from the data without specifying a parametric model, a good understanding of the reason of missing data facilitates the selection model setup. All the above literature assume a single model of missingness, which is either ignorable or NI. However, missing data in practice may come from multiple sources. Different variables may account for each source of missingness, which may be either ignorable or NI. The mixture of ignorable and NI missingness was first discussed by Harel and Schafer (2009). They separately modeled the ignorable and NI missingness mechanism, and proposed a general framework of partially MAR and latently MAR models.

In this chapter, we assume the missing gold standard come from multiple sources, part of which are ignorable and part of which are not. When there are only two types of missingness, our setting of the selection models resembles the partially MAR model in Harel and Schafer (2009). But we also allow for more than two sources of missingness. We propose a two-step procedure to adjust for the verification bias: the first step estimates the verification probability and disease probability by maximizing the likelihood; the second step constructs empirical estimators for the AUC. This extends the results of Liu and Zhou (2010), in which the NI missingness was described by a single selection model. A more plausible missingness model would result in a more accurate estimator for the selection probability, and consequently a more accurate AUC estimator.

The methodology is motivated by the same Alzheimer's disease (AD) data set as in Liu and Zhou (2010). Since the gold standard of AD requires brain autopsy, it is automatically missing for the alive patients. Another reason of missingness may be that the the patients or their family opt not to have brain autopsy. Due to the fact that living people may have better health status and hence are less likely to have AD, the former type of missingness is probably NI, while the latter type can be assumed as ignorable. The data set includes the information of whether a patient is dead or not, so it could be used to improve the previous selection model in Liu and Zhou (2010).

The chapter is organized as follows. Section 5.2 discusses the framework of the selection models for the missingness mechanism, as well as the maximum likelihood estimator. We construct several empirical estimators for AUC in Sect. 5.3.

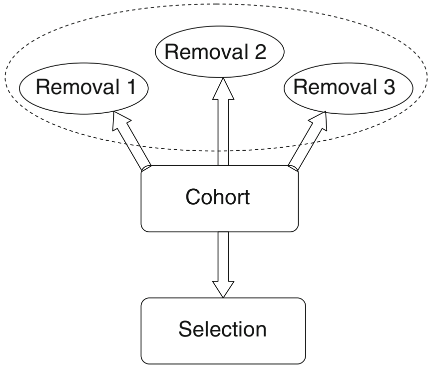

Fig. 5.1 Illustration of the simultaneous selection process—single source of missingness

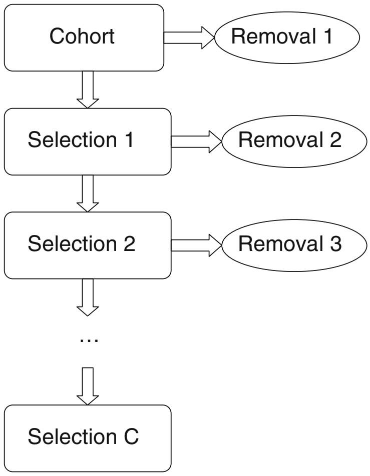

Fig. 5.2 Illustration of the sequential selection process in C steps—multiple sources of missingness

The simulation results are reported in Sect. 5.4, followed by analysis of the AD data set in Sect. 5.5. Finally, the concluding remarks are made in Sect. 5.6.

5.2 Multiple Types of Missingness

We assume that the disease verification process could go through C steps: at each step, a portion of the sample are selected to go through the next step, while the others are removed from gold standard verification. This process is illustrated in Fig. 5.2. As a comparison, the NI selection model in Liu and Zhou (2010) assumes that all the selection steps take place simultaneously, which is illustrated in Fig. 5.1. Therefore, the selection model in Liu and Zhou (2010) actually models the " overall" selection probability. In practical applications, different sources of missingness may indeed occur sequentially. For example, a survey may have NI unit nonresponse and ignorable item nonresponse (Harel and Schafer 2009), where the unit nonresponse apparently happens earlier. When there is no evident temporal order for the sources of missingness, the sequential assumption still provides a convenient way to model the missingness, by factoring out each of the sources.

Denote T_i, D_i, and X_i to be the test result, disease status and the covariates for the ith patient. Denote V_{ci} to be the selection indicator at the cth step ($c = 1, 2, \cdots, C$), with 1 indicating selection and 0 indicating removal. Denote W_{ci} to be the variables that are associated with the cth type of missingness, which may include covariates X_i, test result T_i, and their interactions. For notation simplicity, suppose there are only two types of missingness in D_i ($C = 2$). This could be easily extended to more than two types. The selection model is specified by the following conditional probabilities:

$$\pi_{1i} \equiv \Pr(V_{1i} = 1 | D_i, T_i, X_i) = \text{expit}(W_{1i}^T \beta_1 + \alpha_1 D_i), \tag{5.1}$$

$$\pi_{2i} \equiv \Pr(V_{2i} = 1 | D_i, T_i, X_i, V_{1i} = 1) = \text{expit}(W_{2i}^T \beta_2 + \alpha_2 D_i). \tag{5.2}$$

Note that $\Pr(V_{2i} = 1 | V_{1i} = 0) = 0$, which implies that, subjects removed in the first step cannot re-enter the verification sample. Then a total of three groups of verification status are defined by V_{1i} and V_{2i}: (1) verified sample ($V_{1i} = V_{2i} = 1$); (2) missing at step one ($V_{1i} = V_{2i} = 0$); (3) missing at step two ($V_{1i} = 1$, $V_{2i} = 0$). The NI parameters α_1 and α_2 could be 0, indicating the missingness at the first or the second step is ignorable. We can easily write out the " overall" verification probability:

$$\begin{aligned}\pi_i &\equiv \Pr(V_{1i} = V_{2i} = 1 | D_i, T_i, X_i)\\ &= \pi_{1i}\pi_{2i}\\ &= \text{expit}(W_{1i}^T \beta_1 + \alpha_1 D_i)\text{expit}(W_{2i}^T \beta_2 + \alpha_2 D_i).\end{aligned}$$

In addition, we also need to specify a disease model:

$$\rho_i \equiv \Pr(D_i = 1 | T_i, X_i) = \text{expit}(Z_i^T \gamma), \tag{5.3}$$

where Z_i is the design matrix of variables associated with the disease status.

Define

$$\begin{aligned}\pi_{1i}(d) &\equiv \Pr(V_{1i} = 1 | D_i = d, T_i, X_i)\\ \pi_{2i}(d) &\equiv \Pr(V_{2i} = 1 | D_i = d, T_i, X_i, V_{1i} = 1).\end{aligned}$$

For a subject with disease verification, we observe $V_{1i} = V_{2i} = 1$, D_i, T_i and X_i, and the contribution to the likelihood is

$$l_i = \rho_i^{D_i}(1-\rho_i)^{1-D_i}\pi_{1i}\pi_{2i}.$$

For a subject missing at step one, we observe $V_{1i} = V_{2i} = 0$, T_i and X_i, and its contribution to the likelihood is

$$l_i = \rho_i(1-\pi_{1i}(1))(1-\pi_{2i}(1)) + (1-\rho_i)(1-\pi_{1i}(0))(1-\pi_{2i}(0)).$$

For a subject missing at step two, we observe $V_{1i} = 1$, $V_{2i} = 0$, T_i and X_i. The likelihood contribution is

$$l_i = \rho_i \pi_{1i}(1)(1 - \pi_{2i}(1)) + (1 - \rho_i)\pi_{1i}(0)(1 - \pi_{2i}(0)).$$

Hence, the log likelihood is $L = \sum_i \log l_i$. Note that if $\alpha_c = 0$, $\pi_{ci}(1) = \pi_{ci}(0) = \pi_{ci}$, and the parameter β_c is separated with other parameters in the likelihood function. The estimated verification and disease probabilities, denoted by $\hat{\pi}_i = \hat{\pi}_{1i}\hat{\pi}_{2i}$ and $\hat{\rho}_i$, are then obtained by substituting the estimated parameters.

5.3 ROC Curve and Its Area

With the gold standard observed, the true and false positive rates at threshold s can be estimated as

$$TPR(s) = \frac{\sum_i I(T_i > s)D_i}{\sum_i D_i}$$

$$FPR(s) = \frac{\sum_i I(T_i > s)(1 - D_i)}{\sum_i (1 - D_i)}$$

The AUC is the probability of correctly ordering a case and a control's test result, which is estimated by the Wilcoxon statistic:

$$\hat{\nu} = \left\{ \sum_{i \neq j} I_{ij} D_i (1 - D_j) \right\} \bigg/ \left\{ \sum_{i \neq j} D_i (1 - D_j) \right\}.$$

Similar to Alonzo and Pepe (2005), Liu and Zhou (2010), we replace the unobserved D_i with some estimated version.

The full imputation (FI) estimator replaces every D_i with the estimated disease probability $\hat{\rho}_i$ regardless of its missingness. Hence, the TPR(s), FPR(s), and AUC are given as follows:

$$TPR(s) = \frac{\sum_i I(T_i > s)\hat{\rho}_i}{\sum_i \hat{\rho}_i}, FPR(s) = \frac{\sum_i I(T_i > s)(1 - \hat{\rho}_i)}{\sum_i (1 - \hat{\rho}_i)},$$

$$\hat{\nu}_{FI} = \left\{ \sum_{i \neq j} I_{ij} \hat{\rho}_i (1 - \hat{\rho}_j) \right\} \bigg/ \left\{ \sum_{i \neq j} \hat{\rho}_i (1 - \hat{\rho}_j) \right\}.$$

Denote $\rho_i^{(1)} \equiv \Pr(D_i = 1 | V_{1i} = 0, V_{2i} = 0, T_i, X_i)$ and $\rho_i^{(2)} \equiv \Pr(D_i = 1 | V_{1i} = 1, V_{2i} = 0, T_i, X_i)$ to be the disease probability given the verification indicator. Note that by Bayes rule,

$$\rho_i^{(1)} = \frac{\rho_i(1 - \pi_{1i}(1))(1 - \pi_{2i}(1))}{\rho_i(1 - \pi_{1i}(1))(1 - \pi_{2i}(1)) + (1 - \rho_i)(1 - \pi_{1i}(0))(1 - \pi_{2i}(0))}$$

$$\rho_i^{(2)} = \frac{\rho_i \pi_{1i}(1)(1-\pi_{2i}(1))}{\rho_i \pi_{1i}(1)(1-\pi_{2i}(1)) + (1-\rho_i)\pi_{1i}(0)(1-\pi_{2i}(0))}.$$

Both probabilities could be estimated by replacing ρ_i, $\pi_{1i}(d)$, $\pi_{2i}(d)$ with their maximum likelihood estimators. The second approach, mean score imputation (MSI) only replaces the missing D_i's with $\hat{\rho}_i^{(1)}$ or $\hat{\rho}_i^{(2)}$, depending on the source of missingness for subject i. Let $D_{MSI,i} = I(V_{1i} = V_{2i} = 1)D_i + I(V_{1i} = V_{2i} = 0)\rho_i^{(1)} + I(V_{1i} = 1, V_{2i} = 0)\rho_i^{(2)}$, and $\hat{D}_{MSI,i}$ be the estimated version with $\rho_i^{(\cdot)}$ replaced by $\hat{\rho}_i^{(\cdot)}$. The estimated $TPR(s)$, $FPR(s)$, and AUC are

$$TPR(s) = \frac{\sum_i I(T_i > s)\hat{D}_{MSI,i}}{\sum_i \hat{D}_{\mathrm{MSI},i}}, FPR(s) = \frac{\sum_i I(T_i > s)(1-\hat{D}_{MSI,i})}{\sum_i (1-\hat{D}_{\mathrm{MSI},i})},$$

$$\hat{\nu}_{MSI} = \left\{\sum_{i\neq j} I_{ij}\hat{D}_{MSI,i}(1-\hat{D}_{MSI,j})\right\} \Big/ \left\{\sum_{i\neq j} \hat{D}_{MSI,i}(1-\hat{D}_{MSI,j})\right\}.$$

The third method is inverse probability weighting (IPW). We only make use of the verified subset ($V_{1i}V_{2i} = 1$), but weight each subject with inverse of the selection probability. The corresponding TPR, FPR, and AUC estimators are

$$TPR(s) = \frac{\sum_i I(T_i > s)\ V_i D_i \big/ \hat{\pi}_i}{\sum_i\ V_i D_i \big/ \hat{\pi}_i}, FPR(s) = \frac{\sum_i I(T_i > s)\ V_i\ (1-D_i)\big/ \hat{\pi}_i}{\sum_i\ V_i\ (1-D_i)\big/ \hat{\pi}_i},$$

$$\hat{\nu}_{IPW} = \left\{\sum_{i\neq j} I_{ij}\frac{I(V_{1i}V_{2i}=1)D_i(1-D_j)}{\hat{\pi}_i\hat{\pi}_j}\right\} \Big/ \left\{\sum_{i\neq j} \frac{I(V_{1i}V_{2i}=1)D_i(1-D_j)}{\hat{\pi}_i\hat{\pi}_j}\right\}.$$

The forms of the AUC estimators are analogous to those in Liu and Zhou (2010). The difference is in the likelihood function of the model parameters. Hence, the asymptotic variance of the AUC estimators can be proved using similar arguments as in the Theorem 3 of Liu and Zhou (2010). We briefly state the results here. Denote θ to be the parameters in the selection and disease models. The estimating function for the complate data is $U_{ij}^*(\nu,\theta) \equiv D_i(1-D_j)(I_{ij}-\nu)$. The estimating functions for FI, MSI, and IPW estimators are

$$U_{ij}^{FI}(\nu,\theta) \equiv \rho_i(1-\rho_j)(I_{ij}-\nu), \tag{5.4}$$

$$U_{ij}^{MSI}(\nu,\theta) \equiv D_{MSI,i}(1-D_{MSI,i})(I_{ij}-\nu), \tag{5.5}$$

$$U_{ij}^{IPW}(\nu,\theta) \equiv \frac{I(M_i = M_j = 0)D_i(1-D_j)}{\pi_i\pi_j}. \tag{5.6}$$

We denote these estimating functions by $U_{ij}(\nu,\theta)$ for the notation simplicity. Let

$$Q_i(\nu,\theta) \equiv E_j\left[U_{ij}(\nu,\theta) + U_{ji}(\nu,\theta)\right] + \left[E\frac{\partial}{\partial\theta}U_{ij}(\nu,\theta)\right] I(\theta)^{-1}\dot{l}_i(\theta),$$

where E_j is the expectation with respect to (V_j, D_j, T_j, X_j), $\dot{l}_i(\theta)$ is the ith subject's contribution to the score function, and $I(\theta) \equiv -E\frac{\partial}{\partial\theta}\dot{l}_i(\theta)$ is the information matrix for θ. Let

$$\hat{Q}_i \equiv n^{-1}\left[\sum_{j=1}^{n} U_{ij}(\hat{\nu},\hat{\theta}) + U_{ji}(\hat{\nu},\hat{\theta})\right] - n^{-1}\left[\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\frac{\partial}{\partial\theta}U_{ij}(\hat{\nu},\hat{\theta})\right] \times \left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\dot{l}_i(\hat{\theta})\right]^{-1}\dot{l}_i(\hat{\theta}),$$

We have $\sqrt{n}(\hat{\nu}-\nu) \xrightarrow{d} N(0,\Omega)$, where $\Omega = \frac{var(Q_i(\nu,\theta))}{[\Pr(D_i=0)\Pr(D_i=1)]^2}$. The variance of the AUC estimator contains two sources of variabilities, one from using the U-statistic as an estimator of AUC, the other from estimating the disease and verification models. We note that the variance estimator is different from that of Liu and Zhou (2010), since the likelihood function and the estimated $\hat{\theta}$ are both different.

5.4 Simulation

In this section, we compare the finite sample performance of the proposed estimators with (1) the doubly robust (DR) estimator in Rotnitzky et al. (2006), and (2) the FI, MSI, and IPW estimators in Liu and Zhou (2010) under NI missingness assumption, denoted by NI method. Both DR and NI methods assume the one-step verification process.

We generate two covariates X_1 and X_2 from standard normal distribution and binary distribution, respectively, and the test result from uniform distribution $U(-1,1)$. The disease status is generated from a Bernoulli(ρ) distribution with

$$\rho \equiv \Pr(D=1|\,T,X_1,X_2) = \text{expit}(X_1 + 0.5X_2 + 2T).$$

Two types of missingness ($C=2$) are simulated under the following cases A and B.

Case A: The first step verification V_1 is NI and the second step verification V_2 is ignorable:

$$\Pr(V_{1i}=1|\,D_i,T_i,X_i) = \text{expit}(1 + 0.8X_1 + 0.7X_2 + T + 1.2D),$$
$$\Pr(V_{2i}=1|\,D_i,T_i,X_i,V_{1i}=1) = \text{expit}(2 + 0.5X_1 + 0.2X_2 + 0.8T).$$

Case B: Both steps of verification, V_1 and V_2, are ignorable:

$$\Pr(V_{1i}=1|\,D_i,T_i,X_i) = \text{expit}(1.6 + 0.8X_1 + 0.7X_2 + T),$$
$$\Pr(V_{2i}=1|\,D_i,T_i,X_i,V_{1i}=1) = \text{expit}(2 + 0.5X_1 + 0.2X_2 + 0.8T).$$

The sample size was taken to be 5000. In both cases, we modeled the first step verification with a NI selection model, and the second step with an ignorable model. The simulation was repeated for 500 times. The results are shown in Table 5.1. We

Table 5.1 Comparison of the proposed method with the NI and DR methods for estimating AUC

			Bias	SD	SE	RMSE	Coverage (%)
Case A	Proposed	FI	−0.16	8.72	8.78	8.80	94.6
		MSI	−0.36	8.51	8.49	8.92	93.4
		IPW	−0.23	9.20	9.16	9.35	95.0
	NI	FI	−0.45	9.10	9.02	9.69	93.6
		MSI	−0.60	8.66	8.55	9.73	91.6
		IPW	−0.82	9.77	9.64	11.50	90.0
	DR	$\alpha = 1.2$	0.20	7.79	7.71	7.91	94.6
		$\alpha = 0$	−1.05	8.60	8.58	11.60	85.6
		$\alpha = -0.3$	−1.72	8.93	8.91	15.63	71.4
Case B	Proposed	FI	−0.19	8.84	8.91	8.93	94.0
		MSI	−0.06	8.14	8.21	8.15	95.2
		IPW	−0.27	9.22	9.32	9.42	94.8
	NI	FI	−0.24	9.13	9.18	9.29	95.0
		MSI	−0.11	8.34	8.34	8.37	95.0
		IPW	−0.47	9.63	9.59	10.26	92.6
	DR	$\alpha = 1.2$	0.48	7.62	7.52	8.55	90.6
		$\alpha = 0$	0.01	8.08	8.01	8.08	95.4
		$\alpha = -0.3$	−0.42	8.29	8.21	8.85	94.4

SD standard deviation, *SE* standard error, *RMSE* root mean square error, *FI* full imputation, *MSI* mean score imputation, *IPW* inverse probability weighting, *NI* nonignorable, *DR* doubly robust, *AUC* area under ROC curve

report the bias (in percentage of the true AUC), 1000 times the empirical standard deviation (SD) of the estimates, 1000 times the average standard error (SE) estimates, the root mean square error (RMSE) and the 95% confidence interval (CI) coverage.

For both cases A and B, the bias for the proposed method is generally the smallest. The NI method treats the two types of missingness as a whole, and uses one single selection model to describe the verification process. In case A, the bias for NI method is still relatively small compared to the variance. In case B, as the verification process is truly ignorable, the disease model could still be estimated consistently regardless of the misspecified verification model. Therefore, the performance of FI and MSI estimators is good, while the IPW estimator is a bit biased. Although the NI method is not biased seriously, it is less efficient than the proposed method, especially for the IPW estimator. This is because a better understanding of the missingness mechanism adds information to estimating the selection probability. The bias for DR method is small only with approximately correct NI parameter specification ($\alpha = 1.2$ for case A and $\alpha = 0$ for case B), and substantial if the specified parameter is far from the truth. In case B, it is likely that the DR estimator is not very sensitive to α, which explains the good coverage rates even with incorrect α. Although the DR estimator has the smallest variance, it is hard in practice to specify the correct NI parameter.

The SE of all three proposed methods are close to the SD, indicating that the variance estimators capture the true variability. As for the comparison of FI, MSI, and IPW estimators, imputation based estimators (FI and MSI) are more efficient than the IPW estimators, and hence are recommended in practice.

5.5 NACC Data Example

The National Alzheimer's Coordinating Center (NACC) was established in 1999 to facilitate the collaborative research among the 34 past and present Alzheimer's Disease Centers (ADCs) in the USA. We extracted the NACC Minimum Data Set containing over 70,000 patients who made visit to ADCs between January 1984 and November 2005. The mini-mental state examination (MMSE) is a brief 30-point questionnaire test used to screen for cognitive impairment. Our interested scientific question is how well the MMSE score classifies patients with and without AD.

The data set analyzed by Liu and Zhou (2010) consists of 53,063 patients in total, only 11 % of which received gold standard verification. The verification process has two natural steps: in step one, all the alive patients automatically missed the disease status; in step two, a subsample of the dead patients were chosen to undergo the brain autopsy and to verify their AD status. Hence, we denote $V_{1i} = 1$ if a subject was dead, and denote $V_{2i} = 1$ if a dead subject finally received the disease verification. Assume that the first step of missingness is NI and the second step is ignorable. We use the verification model (5.1) and (5.2) and the disease model (5.3), where T is the MMSE test, D is the true AD status, and X are the patient covariates. The covariates that might be associated with the verification or the disease include age at the MMSE test, gender, race, marital status, clinical diagnosis of AD, other disease conditions (i.e., stroke, Parkinson's disease, depression). The proposed method treats the case nonfatality as a source of missingness and models its probability separated from other missingness. As a comparison, the NI method pools two types of missingness together and directly models $Pr(\, V_{1i}V_{2i} = 1|\, D_i, T_i, X_i)$.

In Tables 5.2 and 5.3, we compare the NI method and the proposed method in estimating the verification and disease models. For the two-step verification model, the covariate's effect on the first-stage missingness are quite different from that on the second-stage missingness. For example, stroke may increase the chances of death, but does not significantly affect the verification probability for a dead patient; patients with lower MMSE score are more likely to be dead, but among those who died, higher MMSE score is associated with greater probability of verification. Therefore, if we pool the two sources of missingness together and use the one-step NI model instead, the estimated covariate's effect is probably an " average" effect of the two stages. The disease models generally agree with each other for NI and the proposed methods. The comparison of NI estimates and our proposed estimates are shown in Table 5.4. The proposed method gives higher AUC estimates than the NI method. The FI, MSI, and IPW estimators are 0.760 (95 % CI: 0.747, 0.773), 0.759 (95 % CI: 0.745, 0.773), and 0.738 (0.721, 0.755), respectively. Furthermore, Fig. 5.3 shows

Table 5.2 The parameter estimation (log odds ratios) for the verification model using the proposed and NI methods

	Proposed		NI
	Step 1	Step 2	
Intercept	**−2.945 (0.055)**	**−1.465 (0.079)**	**−4.527 (0.089)**
Age (per 10 years increasing)	**0.247 (0.013)**	**−0.174 (0.019)**	**0.086 (0.017)**
Gender (M vs. F)	**0.617 (0.029)**	**0.214 (0.037)**	**0.587 (0.037)**
Race (white vs. others)	**0.802 (0.038)**	**1.350 (0.071)**	**1.696 (0.070)**
Marital status (married vs. others)	**−0.094 (0.027)**	**−0.191 (0.039)**	**−0.195 (0.035)**
Stroke (yes vs. no)	**0.390 (0.034)**	0.033 (0.047)	**0.305 (0.043)**
Parkinson's disease (yes vs. no)	**0.703 (0.051)**	**0.264 (0.064)**	**0.641 (0.058)**
Depression (yes vs. no)	**−0.438 (0.034)**	**0.119 (0.050)**	**−0.202 (0.044)**
Clinical AD (yes vs. no)	**0.195 (0.058)**	**−0.211 (0.039)**	0.079 (0.083)
T: MMSE (per 15 points decreasing)	**0.839 (0.032)**	**−0.444 (0.035)**	**0.203 (0.040)**
D: the gold standard (AD vs non-AD)	**1.016 (0.127)**	—	**0.718 (0.178)**

NI nonignorable, *AD* Alzheimer's disease, *MMSE* mini-mental state examination

Table 5.3 The parameter estimation (log odds ratios) for the disease model using the proposed and NI methods

	Proposed	NI
Intercept	**−1.370 (0.195)**	**−1.101 (0.252)**
Age (per 10 years increasing)	**0.192 (0.033)**	**0.134 (0.034)**
Gender (M vs. F)	**−0.415 (0.074)**	**−0.468 (0.075)**
Race (white vs.others)	0.055 (0.159)	−0.025 (0.175)
Marital status (married vs. others)	0.124 (0.078)	0.129 (0.080)
Stroke (yes vs. no)	−0.042 (0.094)	−0.100 (0.095)
Parkinson's disease (yes vs. no)	**0.265 (0.115)**	0.234 (0.122)
Depression (yes vs. no)	0.063 (0.098)	0.110 (0.099)
Clinical AD (yes vs. no)	**1.891 (0.069)**	**1.881 (0.070)**
T: MMSE (per 15 points decreasing)	**1.063 (0.075)**	**0.784 (0.071)**

NI nonignorable, *AD* Alzheimer's disease, *MMSE* mini-mental state examination

the estimated ROC curve using FI approach under the proposed and the NI method. Under the two-stage verification assumption, the ROC curve is slightly higher than that assuming one-stage verification.

In this example, the proposed selection model does not change the 95 % CI width substantially, but it does change the point estimates of the AUC. Even though the FI and MSI estimators do not directly use the selection probability, these imputation-based estimators could still be affected. This is because the selection and disease probabilities are not distinct in the likelihood function, and we have to specify both

Table 5.4 The AUC estimates using NI method and our proposed method

		AUC	95 % CI
NI	FI	0.735	(0.722, 0.748)
	MSI	0.736	(0.724, 0.747)
	IPW	0.716	(0.698, 0.734)
Proposed	FI	0.760	(0.747, 0.773)
	MSI	0.759	(0.745, 0.773)
	IPW	0.738	(0.721, 0.755)

FI full imputation, *MSI* mean score imputation, *IPW* inverse probability weighting, *NI* nonignorable, *AUC* area under *ROC* curve

models correctly to get the unbiased estimators. The NACC example implies that an unrealistic selection model could obviously lead to biased results. In this data set, about 89 % of the patients missed the AD status, so it does not suffice to use a single selection model to account for all the missing data.

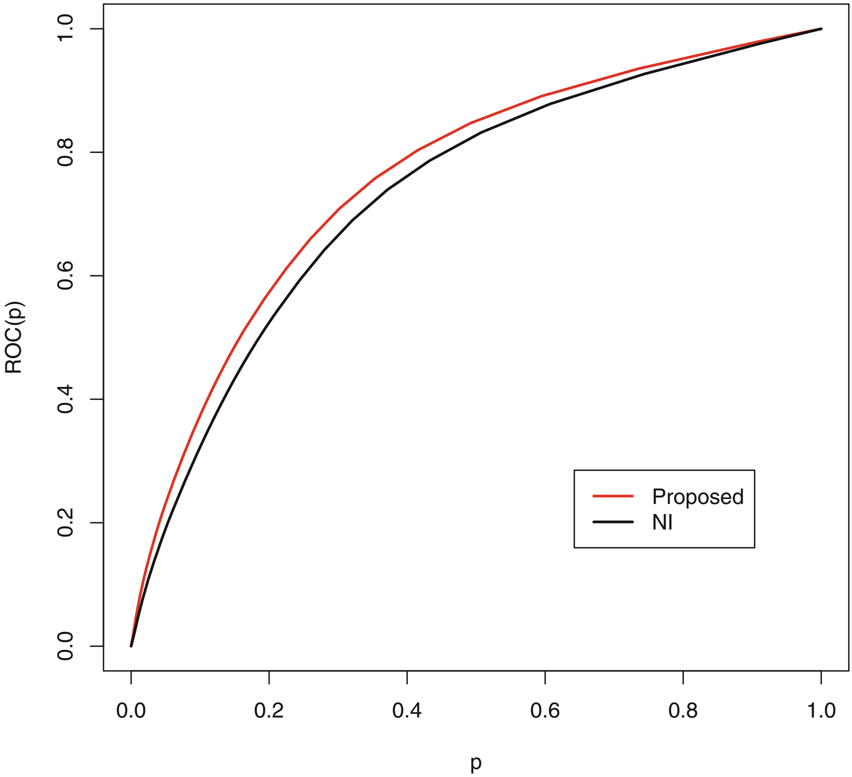

Fig. 5.3 Full imputation (*FI*) estimation of the receiver operating characteristic (*ROC*) curve under the proposed two-stage verification model and the one-stage nonignorable (*NI*) model

5.6 Discussion

In this chapter, we discussed multiple types of missing gold standard in estimating the ROC curve area to extend the results of Liu and Zhou (2010). We assume that different types of ignorable or NI missingness occur sequentially, which are reflected by separate selection models. The overall missingness mechanism might be a mixture of ignorable and NI missingness. The selection and disease probabilities are obtained by maximizing the likelihood. Then the empirical estimators are constructed using imputation or reweighting techniques. The simulation study shows the proposed estimator works well in terms of consistency and CI coverage.

Theoretically, the proposed estimator is generally not robust to model misspecifications, because the likelihood function involves the joint distribution of the disease and verification indicator, and their parameter estimation cannot be separated. That being said, our experience is that mild model misspecification does not create too much bias in the AUC estimators, which is seen in the simulation studies of our previous work (Liu and Zhou 2010). For example, if the true model has a probit link while we specify the logit link, we would expect little bias in the AUC estimators as logit function approximates probit function quite closely. We also found that MSI estimator has slightly better performance than FI and IPW estimators under mild model misspecification. With more severe misspecification, all estimators could have large bias.

As the NI missingness is not nonparametrically testable from the data, we recommended to build up plausible models based on scientific knowledge. In the stages of study design and data collection, careful thoughts about potential missing data are necessary. Then additional information on the reason of missingness can be collected. However, it is quite difficult to gather all the relevant information on the missingness, especially if the missing proportion is high. The missingness may come from quite different sources that could not be explained by a single ignorable or NI selection model. Thus, the heterogeneity of the missingness should be taken into consideration. Stratifying the missingness into several major sources is helpful to remove the heterogeneity, and hence leads to better estimation of the interested parameters. Therefore, the key message of this chapter is that, in practice, if the missingness is known to come from difference sources, it is better to model them separately. When designing new studies, investigators should try their best to collect the information on the reasons of missing data, which could greatly facilitate the model specification. A referee mentioned that machine learning techniques, such as tree-based methods or neural network algorithms are potentially useful to improve the disease and verification models, which is a very interesting extension on the proposed method. However, the difficulty is that, under NI missingness, the disease and verification models need to be estimated jointly, and the model training should be done for both models too, which may be computationally challenging. We leave it as future exploration.

The verification indicator can be also viewed as having more than two categories, indicating different reasons of missingness. Hence, an alternative approach could be directly modeling the verification by a multinomial logistic regression. But the parameters are hard to interpret, and could not explicitly distinguish ignorable versus NI missingness. Our proposed selection models are easy to interpret and implement.

Acknowledgments The authors would like to thank the referees for their insightful comments, which greatly improved the quality of this chapter. This work was supported in part by NIH/NIA grant U01AG016976. Dr. Danping Liu's research is supported by the Intramural Research Program of the National Institute of Health (NIH), *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD). This chapter does not necessarily represent the findings and conclusions of VA HSR&D. Dr. Xiao-Hua Zhou is presently a core investigator and biostatistics unit director at HSR&D Center of Excellence, Department of Veterans Affairs Puget Sound Health Care System, Seattle, Washington.

References

Alonzo TA, Pepe MS (2005) Assessing accuracy of a continuous screening test in the presence of verification bias. Appl Stat 54:173–190

Baker SG (1995) Evaluating multiple diagnostic tests with partial verification. Biometrics 51: 330–337

Begg CB (1987) Biases in assessment of diagnostic tests. Stat Med 6:411–423

Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to verification bias. Biometrics 39:207–215

Harel O, Schafer JL (2009) Partial and latent ignorability in missing-data problems. Biometrika 96:37–50

Kosinski AS, Barnhart HX (2003) Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics 59:163–171

Liu D, Zhou XH (2010) A model for adjusting for nonignorable verification bias in estimation of ROC curve and its area with likelihood-based approach. Biometrics 66:1119–1128

Liu D, Zhou XH (2011) Semiparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias. Biometrics 67:906–916

Rodenberg CA, Zhou XH (2000) ROC curve estimation when covariates affect the verification process. Biometrics 56:1256–1262

Rotnitzky A, Faraggi D, Schisterman E (2006) Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. J Am Stat Assoc 101:1276–1288

Zhou XH (1996) A nonparametric ML estimate of an ROC curve area corrected for verification bias. Biometrics 52:310–316

Zhou XH (1998) Comparing correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. Biometrics 54:349–366

Zhou XH, Castelluccio P (2003) Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. J Stat Plan Inference 115:193–213

Zhou XH, Castelluccio P (2004) Adjusting for non-ignorable verification bias in clinical studies for Alzheimer's disease. Stat Med 23:221–230

Zhou XH, Rodenberg CA (1998) Estimating an ROC curve in the presence of non-ignorable verification bias. Commun Stat 27:635–657