Statistica Sinica Preprint No: SS-2022-0290

Title	Receiver Operating Characteristic Curve for Complex Survey Data
Manuscript ID	SS-2022-0290
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0290
Complete List of Authors	Tamy H. M. Tsujimoto and Jianwen Cai
Corresponding Authors	Jianwen Cai
E-mails	cai@bios.unc.edu

RECEIVER OPERATING CHARACTERISTIC CURVE FOR COMPLEX SURVEY DATA

TAMY H. M. TSUJIMOTO, JIANWEN CAI

University of North Carolina at Chapel Hill

Abstract: The receiver operating characteristic (ROC) curve is frequently used to evaluate the accuracy of medical diagnostic tests. Currently, analysis based on the ROC curve has been performed in large public-use data arising from complex survey samples by ignoring the sampling scheme. This paper proposes a nonparametric estimator for the ROC curve that accounts for complex survey sampling. The asymptotic properties of the estimator are developed using empirical process arguments. Simulation studies showed that our proposed estimator performed well in the practical situations we considered, with better performance for larger sample size and disease proportions. The estimator was illustrated in the National Health and Nutrition Examination Survey (NHANES) to evaluate the discrimination of a traditional risk calculator of undiagnosed diabetes.

Key words and phrases: Complex survey data; ROC curve; Horvitz-Thompson estimator; empirical processes.

1. Introduction

Diagnostic medicine is the process of identifying the disease or condition that a patient has, and ruling out conditions that the patient does not have, through assessment of the patient's signs, symptoms, and results of various diagnostic tests (Zhou et al., 2009). It has evolved over the years as the advances in technology allowed the development of new diagnostic tests for detecting diseases. Examples of diagnostic tests include biochemical serum markers, such as prostate specific antigen (PSA) for prostate cancer, CA-125 for ovarian cancer, creatinine for kidney dysfunction, and cholesterol and blood pressure for cardiovascular disease. Given the importance of diagnostic medicine to population's overall health, and understanding of disease mechanism, statistical methods that assess the accuracy of diagnostic tests in a reliable way are crucial.

The receiver operating characteristic (ROC) curve is the most popular method to assess the performance of a continuous diagnostic test. The curve is defined as the plot of the false positive rate (1-specificity) versus the true-positive rate (sensitivity) across all possible cutoffs of the diagnostic test. The false-positive rate (FPR) is the proportion of non-diseased individuals that test positive for the disease based on the diagnostic test, and the true-positive rate (TPR) is the proportion of diseased individuals that test positive for the disease. The curve is especially useful to compare the performance of different diagnostic tests and obtain optimal cutoffs for the diagnostic test to minimize the misclassification of diseased and non-diseased individuals. Although we focus on medical diagnosis, the ROC curve is widely used in many binary classification problems. A comprehensive discussion on ROC curves can be found in Pepe (2000) and Inácio et al. (2021), for example.

The ROC curve has been widely used in the analysis of data arising from complex surveys. Sample surveys play a critical role in providing essential information in a broad range of areas, serving as an essential resource to guide actions and policies. In the United States, the National Center for Health Statistics (NCHS) is the principal health statistics agency under the Centers for Disease Control and Prevention (CDC), and conducts several population surveys, such as the National Health and Nutrition Examination Survey (NHANES), the National Health Interview Survey (NHIS), and the National Survey of Family Growth (NSFG).

In large scale surveys, the final sample usually does not represent a simple random sample of independent, identically distributed observations from an infinite population. Instead, these studies generally use complex survey designs, including stratification, multistage cluster sampling and unequal selection probabilities to obtain a representative sample in the most effective manner from a finite population. Failure to account for complex survey design may result in biased and inconsistent parameter estimators, underestimated standard errors, and possibly misleading conclusions.

Due to the limited availability of statistical methods, analyses using ROC curves on complex survey data is currently done by ignoring the sampling scheme, even in papers that correctly account for the survey design in other aspects of the analysis. For example, Pandya et al. (2011) assessed the discrimination of traditional cardiovascular disease risk scores in the Third National Health and Nutrition Examination Survey (NHANES III) using unweighted ROC curves. Similarly, DeBoer and Gurka (2014) used an ROC curve to assess the ability of metabolic syndrome Z-score to discriminate impaired glucose tolerance in adolescents, without accounting for the survey design.

The current literature on ROC curves in the context of complex survey sampling predominantly revolves around developing methodologies for the area under the ROC curve (AUC-ROC). For example, Bisoffi et al. (2000) approximated the standard error of the AUC-ROC for two-phase sampling design using bootstrap and jackknife methods. More recently, Yao et al. (2015) proposed a nonparametric estimator for the AUC that accounts for complex sampling, and employed jackknife method and balanced repeated replication for the variance estimation.

While the AUC-ROC offers a convenient summary of diagnostic measure performance, it lacks the granularity of the ROC curve, which comprehensively depicts the trade-off between sensitivity and specificity across all potential thresholds and, as noted by Pepe et al. (2003), small discrepancies in AUC-ROC values can correspond to significant differences in ROC curves. Furthermore, the AUC-ROC's inability to provide optimal cutoff values for diagnostic measures based on sensitivity and specificity further highlights its limitations compared to the ROC curve.

In light of the existing methods' limitations, we propose a non-parametric ROC curve estimator that accounts for complex survey sampling. The proposed estimator's asymptotic properties are derived and evaluated through simulation studies. Furthermore, the method's practical utility is demonstrated by applying it to the National Health and Nutrition Examination Survey (NHANES).

2. Methods

2.1 Setup

Classical sampling theory concerns the inference for finite population quantities (parameters). In this context, the design-based (also called randomization-based) inference is often employed, where the characteristics of interest are considered fixed quantities associated with the finite population. The source of randomness is resulting from the sampling scheme, with random variables indicating whether the population unit is contained in the sample. When the questions of interest are based on parameters of a statistical model, the model-based (also called prediction-based) inference is often preferred. In this framework, the characteristics of interest are considered to be random variables generated from a statistical model.

In this paper, we handle the model-based and design-based inference jointly, using the super-population framework described in Rubin-Bleuer et al. (2005), and followed by Boistard et al. (2017) and Han and Wellner (2021). Under this approach, the finite population is viewed as a realization from a statistical model (superpopulation model), and a sample is drawn from this finite population according to the sampling design. Inference under this approach requires to explicitly account for two sources of randomness: the model-based randomness, accounting for the difference between the finite population parameter and the superpopulation model parameter, and the design-based randomness, accounting for the difference between the sample estimator and the finite population parameter (Pfeffermann, 2000).

Consider a sequence of finite populations $\mathcal{U}^N$ of size $N = 1, 2, \cdots$, with corresponding set of indices $U_N = \{1, \cdots, N\}$. Each index $i \in U_N$ is associated with a unique vector $(y_i, z_i) \in \mathbb{R}^p \times \mathbb{R}^q_+$ representing, respectively, the characteristics of interest, and the sampling design information available at the time of the design of the survey on all units. We assume that $\{(y_i, z_i)\}_{i=1}^N$ are realizations of random variables (Y, Z), $Y : \Omega \mapsto \mathbb{R}^p$, $Z : \Omega \mapsto \mathbb{R}^q_+$, defined on a common probability space $(\Omega, \mathfrak{F}, \mathbb{P}_m)$, and denote $\boldsymbol{y}^N = (y_1, \cdots, y_N)$, $\boldsymbol{Y}^N = (Y_1, \cdots, Y_N)$, $\boldsymbol{z}^N = (z_1, \cdots, z_N)$, and $\boldsymbol{Z}^N = (Z_1, \cdots, Z_N)$.

Let $\mathfrak{S}_N = \{s : s \subset U_N\}$ be the collection of subsets of U_N selected under a given sampling scheme and let $\sigma(\mathfrak{S}_N)$ be the σ-algebra generated by $\mathfrak{S}_N$. A sampling design associated with a sampling scheme is a function $\mathfrak{p} : \sigma(\mathfrak{S}_N) \times \mathbb{R}_+^{q \times N} \mapsto [0, 1]$ such that

(i) for all s in $\mathfrak{S}_N$, $\boldsymbol{z}^N \mapsto \mathfrak{p}(s, \boldsymbol{z}^N)$ is Borel-measurable on $\mathbb{R}_+^{q \times N}$;

(ii) for $\boldsymbol{z}^N \in \mathbb{R}_+^{q \times N}$, $A \mapsto \mathfrak{p}(A, \boldsymbol{z}^N)$ is a probability measure on $\sigma(\mathfrak{S}_N)$.

Note that since p does not depend on $\boldsymbol{y}^N$, only non-informative sampling designs are considered. Similarly to Boistard et al. (2017), for each $\omega \in \Omega$ we define a probability measure $A \mapsto \mathbb{P}_d(A, \omega) = \sum_{s \in A} \mathfrak{p}(s, \boldsymbol{Z}^N(\omega))$, and we say that $(\mathfrak{S}_N, \sigma(\mathfrak{S}_N), \mathbb{P}_d)$ is the design probability space. We will work on a product probability space $(\mathfrak{S}_N \times \Omega, \sigma(\mathfrak{S}_N) \times \mathfrak{F}, \mathbb{P}_{d,m})$ that includes both the super-population and the design space with probability measure $\mathbb{P}_{d,m}$ defined as $\mathbb{P}_{d,m}(s \times E) = \int_E \mathbb{P}_d(s, \omega)\, \mathrm{d}P_m(\omega)$, with $(s, E) \in \sigma(\mathfrak{S}_N) \times \mathfrak{F}$. We adopt $\mathbb{E}_d$, $\mathbb{E}_m$ and $\mathbb{E}_{d,m}$ to denote the expectation with respect to the probability space $(\mathfrak{S}_N, \sigma(\mathfrak{S}_N), \mathbb{P}_d)$, $(\Omega, \mathfrak{F}, \mathbb{P}_m)$ and $(\mathfrak{S}_N \times \Omega, \sigma(\mathfrak{S}_N) \times \mathfrak{F}, \mathbb{P}_{d,m})$, respectively. For a sample s drawn according to a sampling design p, the sampling indicators $\xi_i = I(i \in s)$ are random variables defined on $(\mathfrak{S}_N \times \Omega, \sigma(\mathfrak{S}_N) \times \mathfrak{F}, \mathbb{P}_{d,m})$, with first-order inclusion probabilities defined as $\pi_i(\omega) = \mathbb{E}_{d,m}[\xi_i | \boldsymbol{Z}^N(\omega)]$, and second-order inclusion probabilities defined as $\pi_{ij}(\omega) = \mathbb{E}_{d,m}[\xi_i \xi_j | \boldsymbol{Z}^N(\omega)]$.

2.2 ROC curve for complex survey sampling

Let $\{(Y_i, Z_i) = (X_i, D_i, Z_i) \in \mathbb{R} \times \{0, 1\} \times \mathbb{R}_+^q\}_{i=1}^N$ be i.i.d realizations of the diagnostic test measure X, the disease indicator D, and the sampling design information Z. We denote the cumulative distribution function (cdf) of X conditioned on $D = 0$ as G, and similarly, the cdf of X conditioned on $D = 1$ as F. We assume that F and G have continuous probability density functions (pdf) f and g, respectively. The ROC curve is defined as the plot of $\{(1 - G(c), 1 - F(c)) : c \in \mathbb{R}\}$, or equivalently, as the plot of $\{(s, R(s)) : s \in [0, 1]\}$, where $R(s) = 1 - F \circ G^{-1}(1 - s)$, with $G^{-1}(s) = \inf\{x \in \mathbb{R} : G(x) \geq s\}$, and $F \circ G^{-1}(.) \equiv F(G^{-1}(.))$. The area under the ROC curve (AUC-ROC) is $A = \int_0^1 R(s)\ ds$. The corresponding finite-population quantities are $R_N(s) = 1-F_N \circ G_N^{-1}(1-s)$ and $A_N = \int_0^1 R_N(s)ds$, where $G_N(x) = N_0^{-1} \sum_{i=1}^N I(X_i \leq x, D_i = 0)$, $F_N(x) = N_1^{-1} \sum_{i=1}^N I(X_i \leq x, D_i = 1)$, and $N_d = \sum_{i=1}^N I(D_i = d)$, $d = 0, 1$.

Consider a sample $\mathfrak{s}$, consisting of n $(0 \leq n \leq N)$ units drawn from the finite population using a sampling design p. A survey-weighted estimator for the ROC curve can be obtained by substituting F and G by their Hájek type estimators:

$$R_n(s) = 1 - F_n \circ G_n^{-1}(1 - s), \tag{2.1}$$

where

$$G_n(x) = \frac{1}{\hat{N}_0} \sum_{i=1}^N \frac{\xi_i}{\pi_i} I(X_i \leq x, D_i = 0) \quad \text{and} \quad F_n(x) = \frac{1}{\hat{N}_1} \sum_{i=1}^N \frac{\xi_i}{\pi_i} I(X_i \leq x, D_i = 1), \tag{2.2}$$

with $\hat{N}_d = \sum_{i=1}^N \xi_i \pi_i^{-1} I(D_i = d)$, $d = 0, 1$.

The correspondent estimator for the area A under $R(s)$ is

$$A_n = \int_0^1 R_n(s)\ ds. \tag{2.3}$$

The estimators $F_n(.)$ and $G_n(.)$ can be seen as ratios of Horvitz-Thompson empirical measures defined in the Supplementary Material with respect to the class of function $\mathcal{F} = \{f_{s,l}(y) \equiv f_{s,l}(x,d) = I(x \le s, d = l) : s \in \mathbb{R}, l \in \{0,1\}\}$. Note that this class of functions is P-Donsker (Kosorok, 2008). The finite-dimensional convergence of $\mathbb{G}_N^{\pi}(f_{s,l})$ can be shown similarly as done in Boistard et al. (2017) using Crámer-Wold device. By Corollary 3.13 from Han and Wellner (2021), $\sqrt{n}(\mathbb{P}_N^{\pi} - \mathbb{P}_N) \rightsquigarrow \mathbb{G}^{\pi}$ in $\ell^{\infty}(\mathcal{F})$, where $\mathbb{G}^{\pi}$ is a tight Gaussian process with covariance function

$$\mathrm{Cov}(\mathbb{G}^{\pi}(f_{s,d}), \mathbb{G}^{\pi}(f_{u,d'})) = \lambda\left(\mu_{\pi_1} P(f_{s,d} f_{u,d'}) + \mu_{\pi_2}(P f_{s,d})(P f_{u,d'})\right) \quad f_{s,d}, f_{u,d'} \in \mathcal{F},$$

with $P(f_{s,l}) = \int_{\mathcal{Y}} f_{s,l}(y) P(dy) = P(X \le s, D = l)$ and $P(f_{s,d} f_{u,d'}) = \int_{\mathcal{Y}} f_{s,d}(y) f_{u,d'}(y) P(dy) = P(X \le s \wedge u, D = d)$ for $d' = d$, and zero if $d' \neq d$.

The proposed estimator R_n for the ROC curve depends on the pair (G_n, F_n) through the map $\psi(A, B) = B(A^{-1})$, where A^{-1} is the inverse map of A. Combining the results from Han and Wellner (2021) and Functional Delta Method (Vaart and Wellner, 1996) arguments presented in the Appendix, the following result will follow:

Theorem 1 (FINITE POPULATION INFERENCE)**.** *Consider the estimators F_n, G_n, R_n and A_n as defined in (2.1), (2.2) and (2.3). Suppose that conditions (A1.1)-(A2.2) in the appendix hold and $n, N \to \infty$.*

(a) *(Survey-weighted empirical distributions).*

$$\sqrt{N}\begin{bmatrix} G_n - G_N \\ \\ F_n - F_N \end{bmatrix} \rightsquigarrow \begin{bmatrix} \mathbb{G}_0^{\pi} \\ \\ \mathbb{G}_1^{\pi} \end{bmatrix} = \begin{bmatrix} \{(1-p)^{-1}\mu_{\pi_1}\}^{1/2} B_1(G) \\ \\ \{p^{-1}\mu_{\pi_1}\}^{1/2} B_2(F) \end{bmatrix},$$

where $B_1(.)$ and $B_2(.)$ denote two independent Brownian bridges and $p = P(D = 1)$.

(b) *(Survey-weighted ROC curve). Suppose that F and G have continuous positive densities f and g, respectively, on $[G^{-1}(a) - \epsilon, G^{-1}(b) + \epsilon]$, $\epsilon > 0$ and that $f(G^{-1})/g(G^{-1})$ is bounded on any subinterval (a, b), $0 < a < b < 1$. Then, for $0 < s < 1$*

$$\sqrt{n}\left(F_n \circ G_n^{-1}(s) - F_N \circ G_N^{-1}(s)\right) \rightsquigarrow \sqrt{\lambda\mu_{\pi_1}}\Big\{p^{-1/2}B_2(F \circ G^{-1}(s)) + \\ + (1-p)^{-1/2}\frac{f(G^{-1}(s))}{g(G^{-1}(s))}B_1(s)\Big\}$$

where $B_1(.)$ and $B_2(.)$ denote two independent Brownian bridges. This result implies that $\sqrt{n}(R_n(s) - R_N(s)) \rightsquigarrow \mathbb{W}(G^{-1}(1-s))$, where $\mathbb{W}(u)$ is a Gaussian process with mean zero and covariance function $\mathbb{E}_{d,m}\{\mathbb{W}(u)\mathbb{W}(t)\} = \sigma^2(u,t)$ given by

$$\sigma^2(u,t) = \lambda\mu_{\pi_1}\Big\{p^{-1}(F(u \wedge t) - F(u)F(t)) + \\ + (1-p)^{-1}\frac{f(u)f(t)}{g(u)g(t)}(G(u \wedge t) - G(u)G(t))\Big\} \quad (2.4)$$

(c) (Survey-weighted AUC)

$$\sqrt{n}\,(A_n - A_N) \to N(0, \delta^2)$$

in distribution, where

$$\delta^2 = \int_0^1 \int_0^1 \sigma^2\{G^{-1}(1{-}s), G^{-1}(1{-}t)\}\, ds\, dt = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \sigma^2(s,t)\, dG(s)\, dG(t)$$

The proof for Theorem 1 are presented in Appendix A of the Supplementary Material.

The results for the super-population inference can be obtained from the decomposition

$$\begin{aligned}\sqrt{n}\left(F_n \circ G_n^{-1} - F \circ G^{-1}\right) =& \sqrt{n}\left(F_n \circ G_n^{-1} - F_N \circ G_N^{-1}\right) + \\ & \sqrt{n}\left(F_N \circ G_N^{-1} - F \circ G^{-1}\right).\end{aligned}$$

From the results presented in Theorem 1, we have that the first component converges to a zero mean Gaussian process under $\mathbb{P}_{d,m}$ (and $\mathbb{P}_d$). Using similar arguments, combined with classical empirical processes results, we have that the second component also converges to a zero mean Gaussian process under $\mathbb{P}_m$. Theorem 5.1(iii) from Rubin-Bleuer et al. (2005) imply that the two components are asymptotically independent, leading in the following result:

Theorem 2 (SUPER POPULATION INFERENCE)**.** *Consider the estimators F_n, G_n, R_n and A_n as defined in (2.1), (2.2) and (2.3). Suppose that conditions (A1.1)-(A2.2) hold in the appendix hold and $n, N \to \infty$.*

(a) *(Survey-weighted ROC curve). Suppose that F and G have continuous positive densities f and g, respectively, on $[G^{-1}(a) - \epsilon, G^{-1}(b) + \epsilon]$, $\epsilon > 0$ and that $f(G^{-1})/g(G^{-1})$ is bounded on any subinterval (a, b), $0 < a < b < 1$. Then, for $0 < s < 1$*

$$\sqrt{n}\left(F_n \circ G_n^{-1}(s) - F \circ G^{-1}(s)\right) \rightsquigarrow \sqrt{\lambda(1+\mu_{\pi_1})}\Big\{p^{-1/2} B_2(F \circ G^{-1}(s)) + \\ + (1-p)^{-1/2} \frac{f(G^{-1}(s))}{g(G^{-1}(s))} B_1(s)\Big\}$$

where $B_1(.)$ and $B_2(.)$ denote two independent Brownian bridges. This result implies that $\sqrt{n}(R_n(s) - R(s)) \rightsquigarrow \mathbb{W}(G^{-1}(1-s))$, where $\mathbb{W}(u)$ is a Gaussian process with mean zero and covariance function $\mathbb{E}_{d,m}\{\mathbb{W}(u)\mathbb{W}(t)\} = \sigma^2(u,t)$ given by

$$\sigma^2(u,t) = \lambda(1+\mu_{\pi_1})\Big\{p^{-1}(F(u \wedge t) - F(u)F(t)) + \\ + (1-p)^{-1} \frac{f(u)f(t)}{g(u)g(t)} (G(u \wedge t) - G(u)G(t))\Big\}$$

(b) *(Survey-weighted AUC).*

$$\sqrt{n}\,(A_n - A) \to N(0, \delta^2)$$

in distribution, where

$$\delta^2 = \int_0^1 \int_0^1 \sigma^2\{G^{-1}(1-s), G^{-1}(1-t)\}\, ds\, dt = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \sigma^2(s,t)\, dG(s)\, dG(t)$$

Theorem 2 implies that $\sqrt{n}(R_n(s) - R(s))$ converges in distribution to $N(0, \sigma^2(s))$, with $\sigma^2(s)$ given by

$$\sigma^2(s) = \lambda(1+\mu_{\pi_1}) \left\{ p^{-1}R(s)(1-R(s)) + (1-p)^{-1}\frac{f(G^{-1}(1-s))^2}{g(G^{-1}(1-s))^2} s(1-s) \right\} \tag{2.5}$$

Let $\hat{\sigma}^2(s)$ be the survey-weighted empirical version of $\sigma^2(s)$ with (R, p, F, G, f, g) replaced by their survey-weighted estimates. An approximate level $1-\alpha$ pointwise confidence interval for $R(s)$ is given by $R_n(s) \pm z_{1-\alpha/2}[\hat{\sigma}(s)^2/n]^{1/2}$, where z_α is such that $P(Z \leq z_\alpha) = \alpha$ with $Z \sim N(0,1)$.

Our proposed estimator for the ROC curve reduces to the empirical estimator for the ROC curve as defined in Bertail et al. (2008) for sampling without replacement (SWOR). Additionally, because $\mu_{\pi 1} = \lambda^{-1} - 1$ for SWOR, the result obtained in Theorem 2-a aligns with the corresponding result for the empirical estimator for the ROC curve in a simple random sampling setting also presented in Bertail et al. (2008).

3. Simulation studies

In the simulation studies, we investigate the performance of the proposed estimator for the ROC curve under stratified simple random sampling (SSRS) and stratified two-stage cluster sampling (STSCS). For each sampling scheme, a total of 8 scenarios were considered according to different finite population sizes N = 50,000 and 100,000, disease proportions p = 5%, 25% and sampling fractions λ = 5%, 10%.

We generated populations subdivided in five strata containing 5%, 10%, 25%, 30% and 30% of the observations. We set the AUC for the strata to 0.95, 0.9, 0.8, 0.7, 0.6 respectively, and for each stratum $h = 1, \cdots, 5$, we generated $X_h = \alpha_h D + \epsilon$, where $D \sim \text{Ber}(p)$ and $\epsilon \sim \text{N}(0, 1)$. The ROC curve in each stratum is given by the binormal model $R(s) = \Phi(\alpha_h + \Phi^{-1}(s))$, where $\Phi(.)$ is the standard normal cdf and α_h is determined from the corresponding AUC specified for the h-th stratum. For STSCS, M = 5,000 and 10,000 clusters of sizes 5, 10 and 15 were generated using quantiles of $X_h + \tau$, $\tau \sim N(0, 1)$, in addition to the steps already described.

Samples with size determined by the sampling fraction were drawn assuming uniform allocation. For the first stage of STSCS, m = 310 and 625 clusters were selected from the population of size N = 50,000, and m = 625 and 1250 clusters were sampled from the population of size N = 100,000

. At the second stage, 80% of the observations were sampled from each cluster.

We evaluated the performance at $s \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ in terms of Relative Bias (RB), Empirical Standard Error (ESE), Asymptotic Standard Error (ASE), and Coverage Probability (CP) for 95% confidence intervals. We compare our method (SVY) to the unweighted ROC curve (UN), where the sampling weights are ignored, and the asymptotic variance is computed following Hsieh et al. (1996). We also include the weighted estimator (WT), which shares the same point estimate as our proposed method, but the asymptotic variance is computed using the i.i.d. setting from Hsieh et al. (1996) by plugging in the survey-weighted estimators. Results are obtained by generating 2,000 finite populations and selecting one sample from each of the finite populations.

The results for the relative biases under SSRS and STSCS are reported in Figure 1 and Tables S1 and S2 in the Supplementary Materials. As expected, the relative bias for the UN estimator is quite large, especially at the beginning of the ROC curve, with relative biases close to 30%. In contrast, the values for the SVY and WT estimators never exceed 0.5%.

The estimates for the empirical and asymptotic standard errors under SSRS and STSCS are presented in Figure 2 and Tables S3 and S4 Supplementary Materials. For the SVY estimator, the values were obtained by plugging survey-weighted estimates of p, F, G, f, and g into the expression (2.5). For the UN and WT estimators, unweighted and survey-weighted estimates of p, F, G, f, and g were plugged into the asymptotic variance expression presented in Hsieh et al. (1996). In general, our method estimates are close to the ESE, with better performance for larger sample sizes and disease proportions. The variance estimator that ignores the complex-survey design leads to underestimated standard errors, even when the sampling weights are employed (WT estimator).

Figure 3 and Tables S5 and S6 in the Supplementary Materials give the coverage probabilities of the 95% confidence interval for the ROC curve. In general, the coverage probabilities based on our method are closer to 95% at the beginning of the ROC curve and decrease as we increase the FPR, except for FPR $= 0.9$ in the case of the smallest finite population, sample size, and disease proportion. The WT estimator presents coverage probabilities close to 92% at most, and the UN estimator performs poorly due to the significant bias and underestimated variances.

We also compared the performance of our proposed estimator with a parametric estimator based on the binormal model for the ROC curve Pepe et al. (2003). The point estimate for the binormal estimator (BIN) is derived by plugging in the survey-weighted estimators for the mean and variance for the diseased and non-diseased populations, which are readily available in standard software packages designed for complex survey sampling. Table 7 in the Supplementary Materials provide the RB for BIN under SSRS. In general, the binormal estimator exhibits a positive bias at the beginning of the ROC curve, with its RB gradually decreasing towards the end of the curve.

4. Application

Diabetes and its complications are major causes of morbidity and mortality worldwide. Currently, clinical practice guidelines recommend screening for pre-diabetes and type 2 diabetes with an informal assessment of risk factors or validated risk calculator in asymptomatic adults to guide providers on whether performing a definitive diagnostic test is necessary (Draznin et al., 2022). The current risk assessment tool used by the American Diabetes Association (ADA) to screen for pre-diabetes and type 2 diabetes is adapted from the algorithm developed in Bang et al. (2009) to estimate the risk of undiagnosed diabetes.

In this application, we wish to evaluate the discrimination of the algorithm developed by Bang et al. (2009) using the National Health and

Nutrition Examination Survey (NHANES) between 1999-2006. NHANES is an annual survey conducted by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) that utilizes a complex, multistage probability sampling design to select a representative sample of the non-institutionalized resident population of the United States.

Similarly as presented in Bang et al. (2009), we consider participants aged 20 years or more, excluding pregnant women, that had fasting plasma glucose (FPG) results. The participants are classified into four groups of diabetes status: known diabetes (if answered "yes" to the question "Other than during pregnancy, have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?"), normal glucose metabolism (FPG $<$ 100 mg/dL), pre-diabetes (FPG 100-125 mg/dL), and undiagnosed diabetes (FPG $>$ 125 mg/dL). The participants classified as "known diabetes" are not included in the analysis, and the undiagnosed diabetes was used as the binary outcome. The risk score was computed using age ($<$ 40, 40-49, 50-59, $>$ 59), sex (female, male), family history of diabetes (yes, no), history of hypertension (yes, no), obesity (not overweight, overweight, obese, extremely obese), physically active (yes, no).

In the 1999-2006 NHANES, 20,159 non-pregnant adults aged 20 years or more were enrolled. Out of this sample, 17,696 observations were classified as either normal glucose metabolism, pre-diabetes, and undiagnosed diabetes, and 7,348 observations had information for all variables needed to compute the risk score. In this final analytic sample, the proportion of undiagnosed diabetes is 3.1% (95% CI: 2.6, 3.5).

Figure 4 shows both survey-weighted and unweighted estimates of the ROC curve, as well as its corresponding AUC. The most considerable discrepancies between unweighted and survey-weighted estimates are observed between FPR 0.1-0.5, with the unweighted ROC curve being lower than the survey-weighted ROC curve. As a result, the AUC (survey-weighted = 0.83, unweighted = 0.80) is smaller when the survey weights are not considered.

The observed discrepancy between survey-weighted and unweighted ROC curve estimates in Figure 4 illustrates the importance of incorporating the complex survey sampling when evaluating diagnostic tests using the ROC curve. In this application, failure to account for survey weights led to an underestimation of the test's accuracy, potentially misrepresenting its effectiveness. This underestimation is further illustrated by the differences in cutoff values derived from the survey-weighted and unweighted ROC curves, obtained by maximizing (sensitivity + specificity - 1). The cutoff obtained using the survey-weighted estimator (0.958) resulted in a higher estimated proportion of at-risk individuals diagnosed as diabetic (32.2%) compared to the cutoff from the unweighted estimator (0.965; 29.5%). This difference suggests that neglecting survey weights can lead to misclassification of individuals, potentially impacting patient care and resource allocation.

5. Discussion

In this paper, we studied a nonparametric estimator for the ROC curve in the context of complex survey data. We examined the asymptotic properties of the proposed estimator and evaluated its performance in finite samples through simulation studies. The asymptotic properties of the proposed estimator were developed using empirical process arguments in the super-population framework described in Rubin-Bleuer et al. (2005), where the sources of randomness from both model-based and design-based inference are jointly taken into account.

The uniform convergence for the ROC curve in the finite population and super-population levels were established using key results presented in Han and Wellner (2021), combined with empirical processes arguments. The asymptotic distribution for the finite population and the super-population level AUC was also presented. Simulation studies showed that our proposed estimator performed well in the practical situations considered. The estimator was then applied to a national-level health survey to evaluate the discriminatory ability of a traditional risk calculator of undiagnosed diabetes.

The methods presented in this paper serve as a basis for nonparametric estimation of the ROC curve in the context of complex survey data. The weakly convergence results make it possible to further compute confidence bands for the ROC curve in both super-population and finite population levels. The proposed estimator may serve as an option when using data arising from complex survey data, preventing from biased results and possibly misleading conclusions by ignoring the sampling design.

The developed methods encompass an expression for the variance of the ROC-AUC, accompanied by an explicit formula for estimating the variance of this summary quantity without resorting to resampling methods, as previously employed by Bisoffi et al. (2000) and Yao et al. (2015). Furthermore, our proposed method extends to a broader range of sampling design schemes, leveraging the framework developed by Han and Wellner (2021).

The proposed estimator is a discrete function, whereas the true ROC curve for continuous data is a continuous function. To have a smooth estimate, the study of semiparametric and parametric models for ROC curve estimation in the context of complex survey data deserves attention. In addition to smoothness, if the models are correctly specified, these alternative approaches might be more efficient in estimating the ROC curve in the context of complex survey data.

Our method also assumes that the sampling is noninformative, and further investigation for informative sampling will be worthwhile. There is also little literature exploring the accuracy of a diagnostic test that varies according to a set of characteristics in the context of complex survey data. To address this issue, the estimation of covariate-specific ROC curve for complex survey data is currently under investigation.

6. Figures and Tables

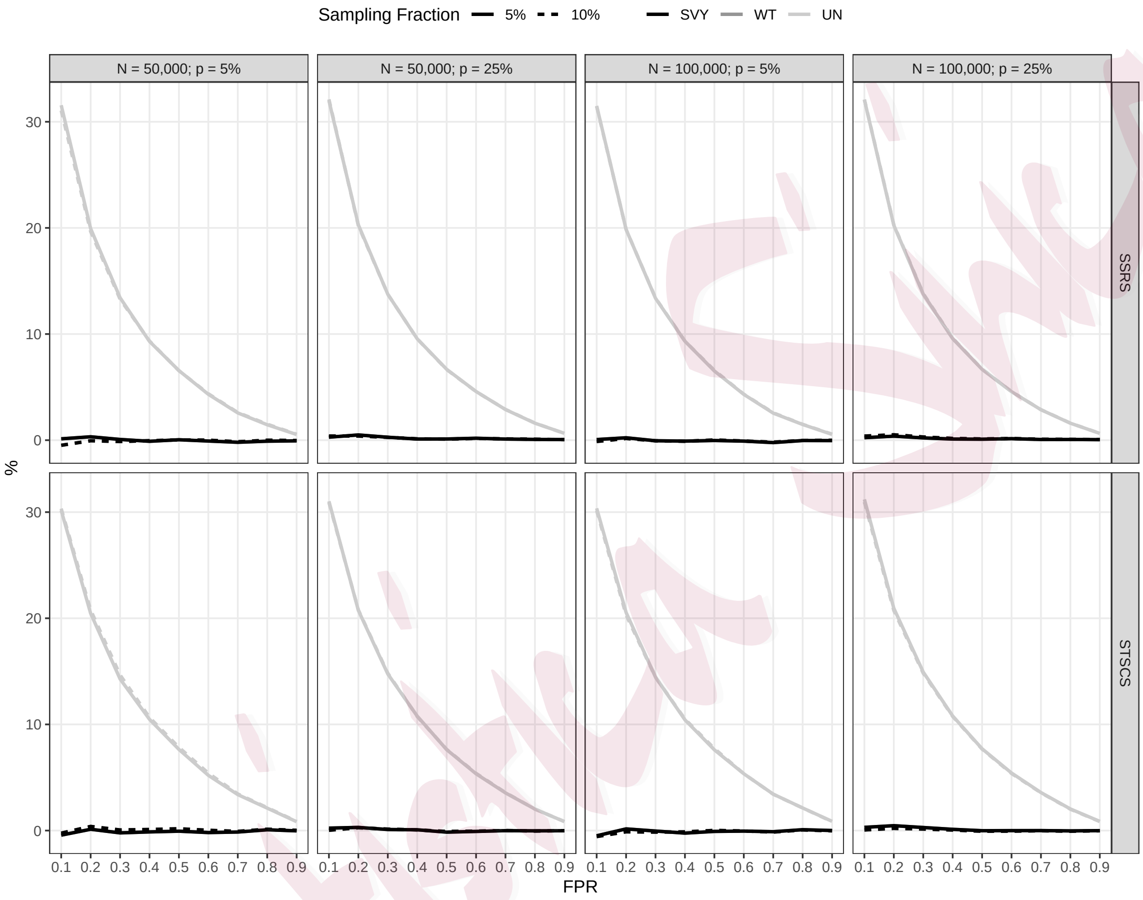

Figure 1: Relative Bias (in %) of the UN, WT, and SVY estimators for the super-population ROC curve with finite population size N, disease proportion p, and sampling fraction λ.

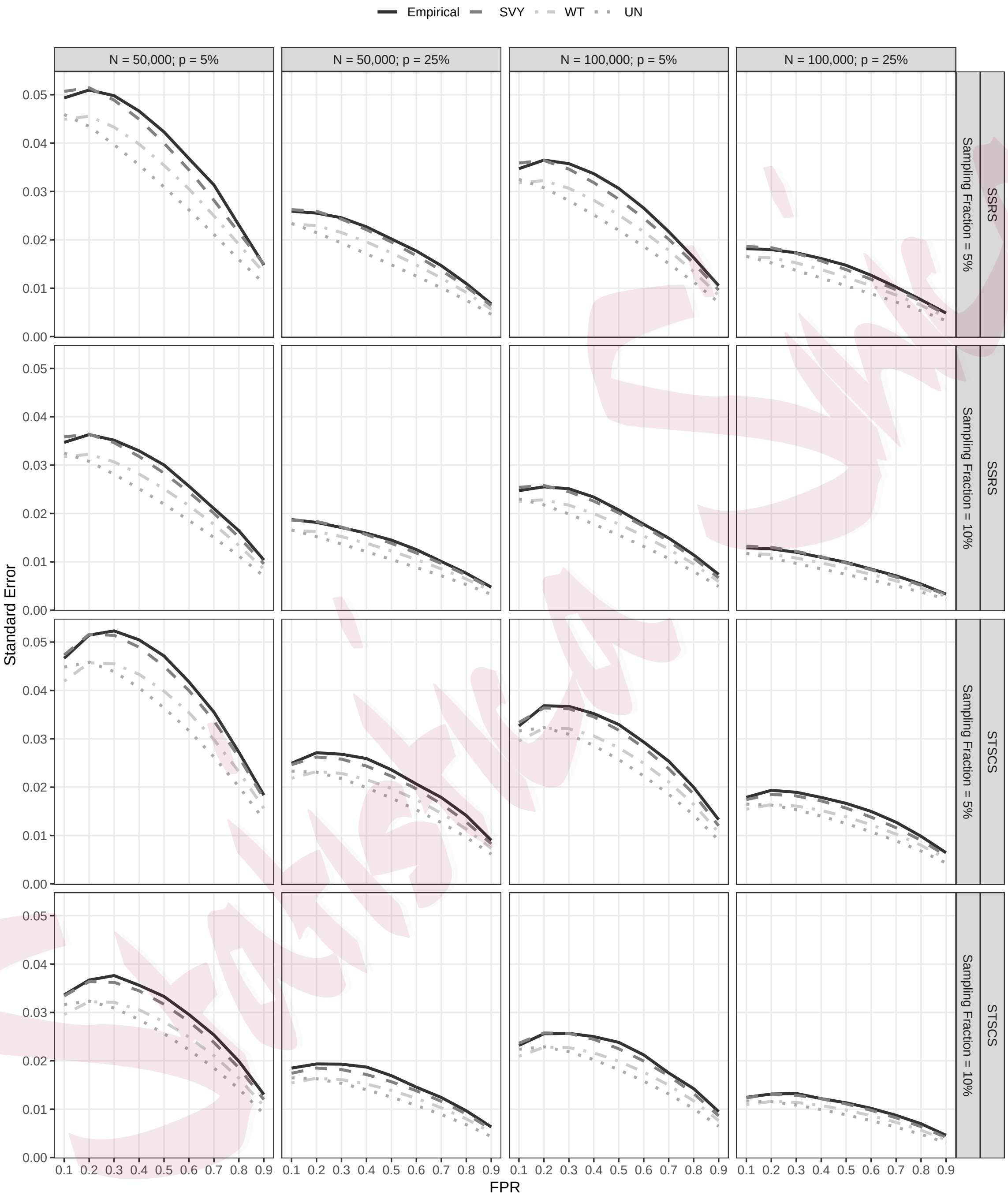

Figure 2: Empirical and Asymptotic Standard Error (in %) of the UN, WT, and SVY estimators for the super-population ROC curve with finite population size N, disease proportion p, and sampling fraction λ.

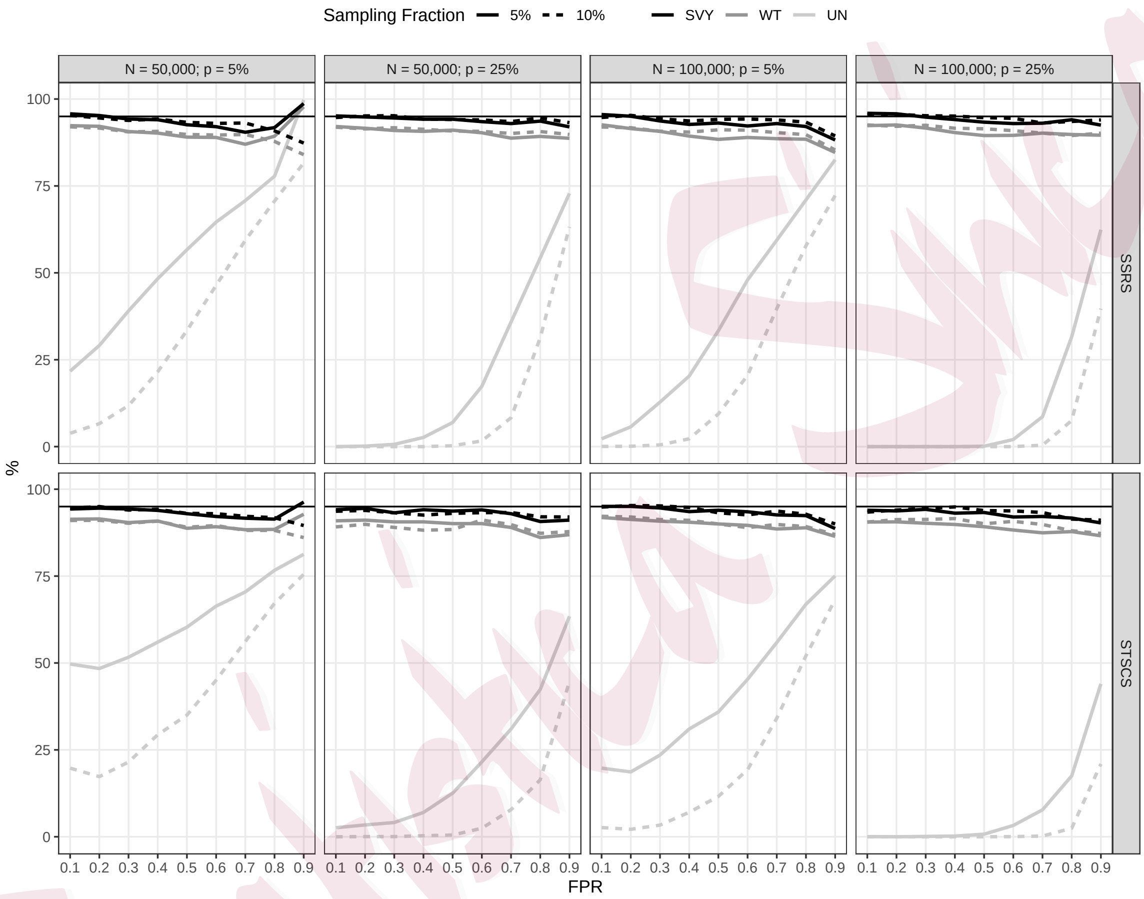

Figure 3: Coverage Probabilities (in %) of the UN, WT, and SVY estimators for the super-population ROC curve with finite population size N, disease proportion p, and sampling fraction λ.

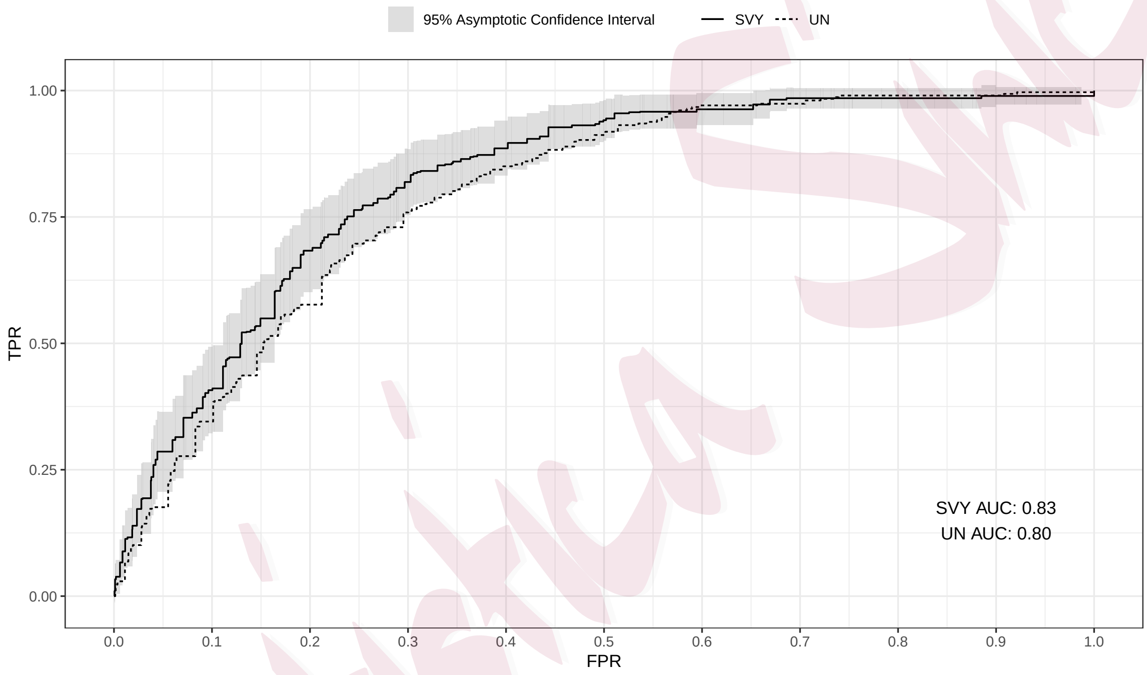

Figure 4: Unweighted (UN) and survey weighted (SVY) estimates ROC curves and survey weighted 95% confidence interval for NHANES data.

Supplementary Materials

The Supplementary Materials include the proofs for Theorems 1 and 2, and supplemental simulation tables.

Acknowledgements

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - File Number 88881.128210/2016-01.

References

Bang, H., A. M. Edwards, A. S. Bomback, C. M. Ballantyne, D. Brillon, M. A. Callahan, S. M. Teutsch, A. I. Mushlin, and L. M. Kern (2009). A patient self-assessment diabetes screening score:: development, validation, and comparison to other diabetes risk assessment scores. *Annals of internal medicine 151*(11), 775.

Bertail, P., S. Clémençcon, and N. Vayatis (2008). On bootstrapping the roc curve. *Advances in Neural Information Processing Systems 21*.

Bisoffi, G., M. A. Mazzi, and G. Dunn (2000). Evaluating screening questionnaires using receiver operating characteristic (roc) curves from two-phase (double) samples. *International Journal of Methods in Psychiatric Research 9*(3), 121–133.

Boistard, H., H. P. Lopuhaä, A. Ruiz-Gazen, et al. (2017). Functional central limit theorems for single-stage sampling designs. *The Annals of Statistics 45*(4), 1728–1758.

DeBoer, M. D. and M. J. Gurka (2014). Low sensitivity of the metabolic syndrome to identify adolescents with impaired glucose tolerance: an analysis of nhanes 1999–2010. *Cardiovascular diabetology 13*(1), 1–8.

Draznin, B., V. R. Aroda, G. Bakris, G. Benson, F. M. Brown, R. Freeman, J. Green, E. Huang, D. Isaacs, S. Kahan, et al. (2022). 2. classification and diagnosis of diabetes: Standards of medical care in diabetes-2022. *Diabetes Care 45*(Supplement_1), S17–S38.

Han, Q. and J. A. Wellner (2021). Complex sampling designs: Uniform limit theorems and applications. *The Annals of Statistics 49*(1), 459–485.

Hsieh, F., B. W. Turnbull, et al. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics 24*(1), 25–40.

Inácio, V., M. X. Rodríguez-Álvarez, and P. Gayoso-Diz (2021). Statistical evaluation of medical tests. *Annual Review of Statistics and Its Application 8*, 41–67.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

Pandya, A., M. C. Weinstein, and T. A. Gaziano (2011). A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the nhanes iii population. *PloS one 6*(5), e20416.

Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association 95*(449), 308–311.

Pepe, M. S. et al. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Pfeffermann, D. (2000). *Handbook of Statistics_29B: Sample Surveys: Inference and Analysis*, Volume 29. Elsevier.

Rubin-Bleuer, S., I. S. Kratina, et al. (2005). On the two-phase framework for joint model and design-based inference. *The Annals of Statistics 33*(6), 2789–2810.

Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer.

Yao, W., Z. Li, and B. I. Graubard (2015). Estimation of roc curve with complex survey data. *Statistics in medicine 34*(8), 1293–1303.

Zhou, X.-H., D. K. McClish, and N. A. Obuchowski (2009). *Statistical methods in diagnostic medicine*, Volume 569. John Wiley & Sons.

Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

E-mail: tamy.tsujimoto@gmail.com

Department of Biostatistics, Gillings School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

E-mail: cai@bios.unc.edu