Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jeong Hoon Jang Date

Statistical Methods for Evaluating Continuous and Functional Diagnostic Markers

By

Jeong Hoon Jang
Doctor of Philosophy

Biostatistics and Bioinformatics

Amita K. Manatunga, Ph.D.
Advisor

Ying Guo, Ph.D.
Committee Member

Limin Peng, Ph.D.
Committee Member

Andrew T. Taylor, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Evaluating Continuous and Functional Diagnostic Markers

By

Jeong Hoon Jang

B.S., New York University, 2014

M.S., Emory University, 2018

Advisor: Amita K. Manatunga, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2019

Abstract

Statistical Methods for Evaluating Continuous and Functional Diagnostic Markers
By Jeong Hoon Jang

The proposed statistical research in this dissertation is motivated by a renal study conducted at Emory University. The study consists of kidneys with suspected obstruction, whose initial diagnoses were provided by nuclear medicine experts and residents. This study also includes functional markers (renogram curves) that have been collected as a noninvasive mean of interpreting kidney obstruction. The overarching scientific goal of this study is two-fold: (1) to understand the reliability of experts' and residents' interpretations of kidney obstruction; and (2) to evaluate the diagnostic utility of renogram curves for detecting kidney obstruction.

First research topic aims at developing new agreement indices based on root mean square of pairwise differences (RMSPD) that can be used to quantify agreement among multiple heterogeneous raters. The advantages of the proposed indices are: (1) interpretations are tied to the measurement scale; (2) satisfactory agreement is conveniently determined via pre-specified tolerable RMSPD. The proposed indices are applied to the Emory renal study to quantify the reliability in interpretations of kidney obstruction.

Quantitative features of functional markers (maximum, time to minimum, average velocity, etc.) are increasingly being used to diagnose diseases. Second research topic aims to study their alignment according to an ordinal reference test. I propose a class of summary functionals, which flexibly represent various quantitative features, and study its alignment via broad sense agreement (BSA, Peng et al., 2011). Asymptotic properties of the proposed BSA estimator are established. This work is applied to the Emory renal study to unveil quantitative features of renogram curves that closely replicate experts' interpretations.

Third research topic aims to assess the diagnostic accuracy of quantitative features based on area under the receiver operating characteristic curve (AUC). I propose a non-parametric AUC estimator that addresses discreteness and measurement error in functional data and establish its asymptotic properties. To describe the heterogeneity of AUC in different subpopulations, I propose a sensible adaptation of a semi-parametric regression model, whose parameters can be estimated by the proposed estimated estimating equations. This work is applied to the Emory renal study to identify quantitative features with high AUCs, and to investigate their relationship with patients' characteristics.

Statistical Methods for Evaluating Continuous and Functional Diagnostic Markers

By

Jeong Hoon Jang

B.S., New York University, 2014

M.S., Emory University, 2018

Advisor: Amita K. Manatunga, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2019

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Diagnostic markers are measurable indicators of the presence of a certain disease or medical state. Well-known examples include complete blood count, antigen level, oxygen level, diagnostic surveys (self-reported or physician-diagnosed) and many more. A quality of disease management and clinical decision-making heavily depends on the availability of good diagnostic markers. However, most markers are imperfect for detection of relevant disease or infection. Thus, rigorous evaluation of a diagnostic marker is a high priority in many clinical research programs.

An agreement study is concerned with assessing the performance, reliability and validity of a novel or generic marker by comparing its clinical measurements against the final true diagnoses or target (gold standard) values. In this dissertation, we aim to develop a set of novel agreement indices that can quantify inter-rater reliability among multiple heterogeneous raters and have simple interpretation tied to the original scale of measurement.

With advancements in technology, more and more cutting-edge, non-invasive medical devices are being used to diagnose and monitor diseases. The increasing complexity of data they generate, however, often pose unique statistical challenges for establishing a clinically interpretative relationship between the data-derived markers and disease pathology. In this dissertation, we focus on one such type of data, namely functional markers which are increasingly being produced by modern devices. Specifically, we develop novel statistical methods for systematically extracting important, interpretative features and patterns from functional markers, and rigorously evaluating their diagnostic utility.

1.2 Literature Review

1.2.1 Statistical methods for assessing agreement

The introduction of a new diagnostic marker is fundamental to the advancement of healthcare. The proposed adoption of a new marker into routine clinical practice essentially requires rigorous assessment of its acceptability, and this amounts to evaluating the accuracy and precision of its clinical measurements. As such, an agreement study is commonly conducted in clinical settings to evaluate the reliability and validity of a new marker by comparing its clinical measurements against the designated gold standard or target values taken on the same subjects (Lin et al., 2002). In this section, we review statistical methods used to assess agreement.

Categorical or ordinal scale

Cohen's kappa coefficient (κ) has been widely used assess the agreement of binary (Cohen, 1960) or categorical (Fleiss, 1971) outcomes between two raters. κ is a chance-corrected measure of agreement that is calculated from the observed and expected frequencies on the diagonal of a square contingency table. The formula to calculate κ is

$$\kappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e},$$

where p_0 is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement. A value of 0 for κ indicates agreement equivalent to chance and a value if 1 indicates perfect agreement.

Fleiss' kappa coefficient (Fleiss, 1971, Kraemer, 1980) is an extension of the Cohen's kappa coefficient that can assess agreement among more than two categorical raters. Similar to the original kappa coefficient, the Fleiss' kappa coefficient can be interpreted as expressing the extent to which the observed amount of agreement among raters exceeds what would be expected if all raters provide measurements completely

randomly.

In many reliability and validity studies, clinical measurements are often on an ordinal scale. With ordinal measurements, the weighted kappa coefficient (Cohen, 1968) is a popular chance-corrected measure of agreement. The measure is computed using a predefined table of weights which quantifies the degree of disagreement between the two raters; this approach allows counting disagreements differently by setting higher weights to represent higher disagreement.

The aforementioned measures can provide a good overall summary of agreement among categorical- or ordinal-scale raters, but may suffer from a loss of precision due to the potential variations of agreement among different subpopulations. Investigators thus may wish to assess the degree of agreement taking into account clinical and/or demographic covariates. Several generalized estimating equations (GEE) approaches have been proposed to model kappa or weighted kappa as a function of covariates (Gonin et al., 2000, Klar et al., 2000, Williamson et al., 2000). The GEE approach is particularly advantageous because it requires minimal assumption of the data and enables estimation and inferences for the kappa estimates to be done simultaneously (Banhart et al., 2001, Lin et al., 2007).

Continuous scale

Clinical measurements are usually in numerical forms or continuous data, such as blood pressure, glucose level, oxygen level, etc. To date, a considerable body of research has sought to develop statistical methods in assessing agreement between continuous raters. Bland and Altman (1986, 1999) advocated the use of a graphical method to plot the difference scores of two continuous measurements against the mean for each subject and to quantify agreement by studying the mean difference and constructing limits of agreement. Intraclass correlation (Bartko, 1966) and within-subject coefficient (Lee et al., 1989) are traditional measures of agreement for

continuous data.

The concordance correlation coefficient (CCC) is one of the most popular scaled indices for assessing agreement between paired continuous raters (Lin, 1989, 1992). Specifically, let Y_1 and Y_2 be denote a pair of continuous measurements produced by two raters from the same subject, with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12} (finite second moments). Lin (1989) defined the measure as

$$\text{CCC} = 1 - \frac{E[(Y_1 - Y_2)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}, \tag{1.1}$$

where $E[(Y_1 - Y_2)^2]$ is the mean squared deviation (MSD) that characterizes the degree of discordance between Y_1 and Y_2. CCC takes into account both accuracy and precision in measurements, and can be characterized by its ease of representation, in which 1 (-1) represents a perfect (perfectly reversed) agreement, and 0 represents no agreement. A GEE approach was introduced to model covariate-adjusted CCC (Banhart et al., 2001). Furthermore, Banhart et al. (2002) introduced the overall concordance correlation coefficient (OCCC), which is a generalization of the CCC in the presence of multiple raters.

Intuitively, a good diagnostic utility of a continuous marker may be warranted if a large proportion of its measurements are within a predetermined boundary from target values. In this context, a set of unscaled agreement indices, which directly incorporates MSD as a performance criterion, was proposed (Lin, 2000, Lin et al., 2002). Lin (2000) introduced the total deviation index (TDI) that describes an acceptable/tolerable range of absolute difference such that a predetermined proportion of the absolute differences between paired continuous measurements taken on the same subject is within that acceptable range. Specifically, let $|D| = |Y_1 - Y_2|$ denote the absolute difference so that $E(D^2)$ represents the MSD. Then, given predetermined proportion π_0, the solution to $\pi_0 = \Pr(|D| < x) = \Pr(D^2 < x^2)$ defines the TDI_{π_0},

that is,

$$\text{TDI}_{\pi_0} = \sqrt{G^{-1}(\pi_0)},$$

where $G(\cdot)$ is the cumulative distribution function (CDF) of D^2 and $G^{-1}(\cdot)$ is the inverse function of $G(\cdot)$. For estimation and inference, (Lin, 2000) assumed that D is normal with mean $\mu_d = \mu_1 - \mu_2$ and variance $\sigma_d^2 = \sigma_1^2 + \sigma_2^2 + \sigma_{12}$, so that G represents the cumulative noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter μ_d^2/σ_d^2. For non-normal data, several non-parametric approaches for estimation and inference of the TDI were introduced (Choudhary, 2010, Perez-Jaume and Carrasco, 2015, Lin et al., 2016).

Coverage probability (CP) is a reciprocal concept, in which a proportion of the absolute differences within a pre-specified acceptable range is computed. Both TDI and CP measures were extended through a mixed ANOVA model to allow evaluation of agreement among multiple raters (Lin et al., 2007). Recently, Banhart (2016) proposed relative area under the coverage probability (RAUCPC) as an aggregated agreement index in the presence of multiple predetermined acceptable/tolerable absolute differences. The index is scaled in nature but based on a series of CP values evaluated over the range of absolute differences.

Functional scale

With the advancement in data collection technology, more and more clinically applicable markers are being collected as functional curves. Herein the measurements on a subject are assumed to be realizations of a continuous underlying process that are sampled at dense discrete time points (or points on other continua). The individual datum is thus the whole function (curve), rather than its value at any particular point. A comprehensive overview of recently developed functional data analysis (FDA) tools and their interesting applications can be found in the book by Ramsay and Silverman (2005) and references therein.

Li and Chow (2005) extended the traditional CCC measure defined in (1.1) to allow assessment of agreement between paired functional markers. The authors characterized the degree of discordance between the two functional markers by their MSD, which was newly defined based on the functional inner product. Specifically, let Y_1 and Y_2 be the two functional markers defined on some probability functional space $\mathcal{F}$, and denote $Y_1(t)$ and $Y_2(t)$ as their respective realizations on $t \in \mathcal{T}$, a finite closed real interval. Then the functional inner product in $\mathcal{F}$ can be defined as (Li and Chow, 2005)

$$< Y_1, Y_2 >= E \int_{\mathcal{T}} Y_1(t) Y_2(t) w(t) dt,$$

where w is a nonrandom weight function that takes non-negative values on $\mathcal{T}$. Using this notion of inner product, Li and Chow (2005) defined the CCC for assessing agreement between Y_1 and Y_2 as

$$CCC = \frac{2 < Y_1 - E(Y_1), Y_2 - E(Y_2) >}{||E(Y_1) - E(Y_2)||^2 + ||Y_1 - E(Y_1)||^2 + ||Y_2 - E(Y_2)||^2},$$

where $||Y|| = \sqrt{< Y, Y >}$. This extended CCC measure possesses same characteristics as those of two continuous random variables (Lin, 1989); for instance, its value of 1 (-1) represents a perfect (perfectly reversed) agreement, and its value of 0 represents no agreement. Note that the weight function allows to assign different importance to different parts of $\mathcal{T}$.

More recently, Rathnayake and Choudhary (2016) proposed a methodology for constructing pointwise and simultaneous tolerance bands for functional measurements, as an extension of tolerance intervals for univariate measurements that have been widely used to assess individual bioequivalence (Brown et al., 1997, Chow and Liu, 2008).

Different scales

In some clinical studies, there is no guarantee that measurements produced by two raters are on the same scale, even though they measure and represent the same biological process or disease severity. For instance, in many mental health studies, researchers are very interested in replacing one diagnostic instrument with another less costly (surrogate) diagnostic instrument for more effective detection of psychiatric disorders; however, the two instruments often have different scales due to distinctive structures and point systems in their respective questionnaires (Peng et al., 2011, Rahman et al., 2017). All the approaches described above are not applicable in such cases, because they require measurements to be on the same scale.

Recently, Peng et al. (2011) proposed a broad sense agreement (BSA) framework, which is designed to evaluate the capability of interpreting a continuous measurement in an ordinal scale, and thus extends the classical framework of agreement. Let X and Y denote a continuous measurement and Y an ordinal measurement of a common outcome variable from the same subject, respectively. Peng et al. (2011) stated that order consistency is a crucial requirement for perfect broad sense agreement, that is, if $X_{(*k)}$ denotes the randomly selected X given $Y = k$ $(k = 1, 2, \ldots, K)$, a perfect broad sense agreement (disagreement) case implies $X_{(*1)} < X_{(*2)} < \cdots < X_{(*K)}$ $(X_{(*1)} > X_{(*2)} > \cdots > X_{(*K)})$ with probability 1.

Let $\{R_1, R_2, \ldots, R_K\}$ denote the ranks of $\{X_{(*1)}, X_{(*2)}, \ldots, X_{(*K)}\}$. Then the following index quantifies the degree of BSA between Y and X (Peng et al., 2011):

$$\rho_{\text{bsa}}(X, Y) = 1 - \frac{E\Big\{\sum_{k=1}^{K}(k - R_k)^2\Big\}}{E\Big\{\sum_{k=1}^{K}(k - R_k)^2 \mid X \perp Y\Big\}},$$

where $E(\cdot)$ denotes the expectation and $E(\cdot \mid X \perp Y)$ denotes the expectation given that X and Y are independent. This index is basically a scaled measure of discrep-

ancy between the observed ranks and the expected ranks under perfect BSA among continuous measurements. The index always takes a value between -1 and 1, with 1 (or -1) representing perfect broad sense agreement (disagreement), and 0 representing independence between X and Y.

To accommodate potential variations of BSA among different subpopulations, Rahman et al. (2017) recently proposed a non-parametric regression framework that allows for nonlinear covariate effects on BSA. This new method provides a robust tool for further investigating population heterogeneity in the alignment between ordinal and continuous measurements.

1.2.2 Statistical methods for evaluating diagnostic accuracy of markers

Accurate diagnosis and monitoring of diseases heavily rely upon an availability of good markers. Statistically determining whether a certain marker is good or not amounts to rigorously evaluating its ability to discriminate between the diseased and non-diseased status. In this section, we review statistical methods for evaluating markers based on their discriminating ability for diagnosing disease.

Binary scale

Let Y denote a binary marker value that gives either a positive ($Y = 1$) or negative ($Y = 0$) result for a particular patient whose disease status is given by D, where $D = 1$ if diseased and $D = 0$ if non-diseased. The most common way of reporting the diagnostic accuracy of a binary marker is using sensitivity and specificity. Sensitivity (true-positive) is the probability of a positive result given disease is present, denoted $Pr(Y = 1 \mid D = 1)$, and specificity (true-negative) is the probability of a negative result given disease is absent, denoted $Pr(Y = 0 \mid D = 0)$. Some people prefer to use the positive predictive value $Pr(D = 1 \mid Y = 1)$ and negative predictive value

$Pr(D = 0 \mid Y = 0)$ which indicate the likelihood of the disease and non-diseased state of a patient given the positive and negative marker values, respectively. These values, however, may give misleading conclusion regarding the accuracy of markers as they depend on the prevalence of disease $P(D = 1)$; extremely high (low) prevalence may result in spuriously large (small) positive predictive values (Altman and Bland, 1994).

Continuous scale

Receiver operating characteristic (ROC) analysis can be used to evaluate the diagnostic accuracy of continuous markers (Pepe, 2003). Now let Y denote the value of a continuous marker. We will assume that $Y > c$ indicates a classification into state $D = 1$, where c is a certain cutoff value. The ROC curve is a popular tool for visualizing the diagnostic accuracy of a continuous marker. It plots 1-specificity $Pr(Y > c \mid D = 0)$ on the x-axis versus the corresponding sensitivity $Pr(Y > c \mid D = 0)$ on the y-axis versus at each possible cutoff point c (Pepe, 2000).

There is a convenient mathematical form for the ROC curve which facilitates further investigation into many of its properties. Let $F(c) = Pr(Y > c \mid D = 1)$ and $G(c) = Pr(Y > c \mid D = 0)$ denote the survival functions for Y given the diseased and non-diseased status, respectively. Then, the ROC curve can be written as

$$\mathrm{ROC}(t) = F\{G^{-1}(t)\}, \tag{1.2}$$

where t represents a fixed level of 1-specificity (point on the x-axis of the ROC curve) on a given threshold c, that is, $G(c) = t$ (Pepe, 1997).

The empirical ROC curves can be obtained by connecting the observed 1-specificity and sensitivity pairs. But often the empirical ROC curve are quite jagged, and the smoothed version is desired. There are several parametric and non-parametric meth-

ods for the estimation of the smooth ROC curve (Zou et al., 1997, Pepe, 2003, Peng and Zhou, 2004). The most common way of obtaining a smooth ROC curve is using a binormal model. Suppose that continuous marker values given the disease status is normally distributed; that is, $Y \mid D = d \sim N(\mu_d, \sigma_d^2)$, $d = 0, 1$. Then the smooth ROC curve can be obtained via equation (1.2), which can be re-written under the binormal model as

$$\mathrm{ROC}(t) = \Phi\left(\frac{\mu_1 - \mu_0 + \sigma_0\Phi^{-1}(t)}{\sigma_1}\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function (CDF) of a standard normal distribution.

Often, the performance of a marker may depend on patient characteristics. As such, various regression modelling approaches have been proposed to assess possible covariate effects on the ROC curve. One popular approach is to formulate a regression model for the marker value given each disease status and induce the regression form of the ROC curve (Pepe, 1998, Farraggi, 2003, Rodríguez-Álvarez et al., 2011). For instance, under the binormal model, we can set $\mu_D = \alpha_0 + \alpha_1 D + \alpha_2 X + \alpha_3 DX$, where X denotes a covariate, and induce the regression model for the ROC curve as

$$\mathrm{ROC}_X(t) = \Phi\left(\beta_0 + \beta_1\Phi^{-1}(t) + \beta_2 X\right)$$

where $\beta_0 = -\alpha_1/\sigma_1$, $\beta_1 = \sigma_0/\sigma_1$ and $\beta_2 = -\alpha_3/\sigma_1$. This approach has been further extended to evaluate a longitudinal marker (Zheng and Heagerty, 2004) and adjust for functional covariates (Inácio et al., 2012). Another approach is to formulate a regression model that directly evaluates the covariate effects on the ROC curve (Alonzo and Pepe, 2002, Cai, 2004). More recently, Janes and Pepe (2009) introduced a covariate-adjusted ROC curve which is a measure of covariate-adjusted classification accuracy.

The area under the ROC curve (AUC) summarizes performance information of

a marker across all threshold values. An AUC of 1 represents a perfect marker, while AUC of 0.5 represents a worthless marker (e.g., coin flip). It has been shown that AUC is equivalent to the probability of a marker value of a randomly selected diseased subject is greater than that of a randomly selected non-diseased subject (Bamber, 1975). Empirical AUC can be obtained by numerically integrating (e.g., trapezoidal rule) the empirical ROC curve. AUC can be also computed based on the smooth ROC curve. For instance, under the binormal assumption, the AUC takes the form of

$$\text{AUC} = \Phi\left(\frac{\mu_1 - \mu_0}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right).$$

Several regression modelling approaches have been developed to systematically assess covariate effects on AUC (Pepe, 1998, 2003, Dodd and Pepe, 2003).

1.2.3 Latent class models for evaluating diagnostic accuracy of markers under no gold standard

Until now, we have focused on statistical methods for evaluating diagnostic accuracy of a marker when a gold standard is available for verifying disease status. For many diseases, however, neither the true disease status nor a gold standard test is available. In some cases, the disease itself is not easily detectable due to its complex biological mechanism underlying its trait; in other cases, a gold standard test may be too invasive or expensive to perform without a definitive symptom that directly reflects the presence of the disease. Disease diagnosis thus often relies upon information obtained from imperfect or subjective diagnostic markers. In this section, we review statistical methods using latent class models for evaluating diagnostic accuracy of single or multiple markers in the absence of a gold standard.

Notations and general formulation of the model

Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{ip})^T$ denote the vector of p marker values for individual i ($i = 1, 2, \ldots, n$), with Y_{ij} denoting the j^{th} marker value ($j = 1, 2, \ldots, p$). Let D_i be the true unknown binary indicator of disease for patient i, where $D_i = 0$ means diseased, $D_i = 0$ means non-diseased, and $\pi = \Pr(D_i = 1)$ represents the prevalence of a disease. Assuming D_i as the latent class, a general formulation of the likelihood of the latent class model is given by

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \big\{\pi P(\mathbf{Y}_i|D_i = 1, \boldsymbol{\theta}) + (1 - \pi)P(\mathbf{Y}_i|D_i = 0, \boldsymbol{\theta})\big\}, \tag{1.3}$$

where $\boldsymbol{\theta}$ is a vector of unknown prevalence and marker parameters, and $P(\mathbf{Y}_i|D_i = d_i, \boldsymbol{\theta}) = P(Y_{i1}, Y_{i2}, \ldots, Y_{iP}|D_i = d_i, \boldsymbol{\theta})$ denotes the joint probability density function (PDF) of $\mathbf{Y}_i$ given $D_i = d_i$ ($d_i = 0, 1$). For each subject i, this is a finite mixture model with mixing proportions π and $1-\pi$ and two component distributions $P(\mathbf{Y}_i|D_i = 1, \boldsymbol{\theta})$ and $P(\mathbf{Y}_i|D_i = 0, \boldsymbol{\theta})$. Approach to parameter estimation and inference based on the likelihood function (1.3) differs depending the distribution of $\mathbf{Y}_i$, or more broadly, depending on whether the marker values are binary or continuous.

Binary scale

Suppose that we have binary markers for each i^{th} subject, that is, a positive result on the j^{th} marker is denoted by $Y_{ij} = 1$ and a negative result by $Y_{ij} = 0$. Then, $\alpha_j = \Pr(Y_{ij} = 1 \mid D_i = 1)$ is the sensitivity, and $\beta_j = \Pr(Y_{ij} = 0 \mid D_i = 0)$ is the specificity of the j^{th} marker. When the conditional independence can be assumed, that is, the p markers are independent of each other given the true disease status, j^{th} marker value given i^{th} individual disease status ($Y_{ij} \mid D_i = d_i$) follows an independent Bernoulli distribution with success probability $\Pr(Y_{ij} = 1|D_i = d_i)$. Thus, the likelihood in

(1.3) takes the simplest form:

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \Big[\pi \prod_{j=1}^{p} \left\{ \alpha_j^{y_{ij}} (1-\alpha_j)^{1-y_{ij}} \right\} + (1-\pi) \prod_{j=1}^{p} \left\{ (1-\beta_j)^{y_{ij}} \beta_j^{1-y_{ij}} \right\} \Big]. \quad (1.4)$$

This is called the conditional independence model or the two latent class model, and the parameters $\boldsymbol{\theta} = [\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_p, \pi]^T$ in (1.4) can be estimated by the EM algorithm (Dempster et al., 1977) or the quasi-Newton method (Thisted, 1988). The conditional independence model was first implemented in Hui and Walter (1980), where the authors provided identifiability conditions and maximum likelihood (ML) estimates with two diagnostic markers and two populations (with different prevalences).

The conditional independence assumption, however, is often violated in practice, especially among markers based on a common biological phenomenon. Several authors have demonstrated that it is important to account for such conditional dependence, if it exists, in order to achieve unbiased estimation of the prevalence of disease and accuracy of the diagnostic markers (Vacek, 1985, Torrance-Rynard and Walter, 1997). Qu et al. (1996) proposed a Gaussian random effects (GRE) model which induces a positive correlation among marker values. Specifically, the probability of testing positive on jth marker depends on both the disease status $D_i = d_i$ of the subject and the Gaussian latent variable u_i, through a probit regression model

$$\Pr(Y_{ij} = 1 \mid D_i = d_i, u_i) = \Phi(a_{jd_i} + b_{jd_i} u_i), \quad (1.5)$$

where u_i is a subject-specific random effect that follows a standard normal distribution, Φ is the CDF of a standard normal variate, a_{jd_i} and b_{jd_i} are diagnostic accuracy parameters, and the two unobserved random variables d_i and u_i are assumed to be independent of each other. By integrating the equation (1.5) over the standard normal variate u_i, we can estimate the sensitivity and specificity of the markers as

$\alpha_j = \Phi(a_{j1}/\sqrt{1+b_{j1}})$ and $\beta_j = \Phi(-a_{j0}/\sqrt{1+b_{j0}})$, respectively. The likelihood of the model (1.5) is given by

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{p} \int \Big(\pi \prod_{j=1}^{p} \left[\Phi(a_{j1} + b_{j1}u_i)^{y_{ij}} \{1 - \Phi(a_{j1} + b_{j1}u_i)\}^{1-y_{ij}} \right] + (1-\pi) \prod_{j=1}^{p} \left[\Phi(a_{j0} + b_{j0}u_i)^{y_{ij}} \{1 - \Phi(a_{j0} + b_{j0}u_i)\}^{1-y_{ij}} \right] \Big) \phi(u_i) du_i,$$

which can be easily derived from the equation (1.3) by noticing that the diagnostic markers are conditionally independent given both d_i and u_i. Herein, ϕ denotes the PDF of a standard normal variate, and EM algorithm can be used to obtain the ML estimates of the parameters $\boldsymbol{\theta} = [a_{10}, \ldots a_{p0}, a_{11}, \ldots a_{p1}, b_{10}, \ldots b_{p0}, b_{11}, \ldots b_{p1}, \pi]^T$

Several other latent class models have been proposed for evaluating the accuracy of binary markers. Torrance-Rynard and Walter (1997) introduced additional parameters in the joint probabilities of the marker values to capture pairwise conditional dependence between markers. Yang and Becker (1997) used marginal models to account for the dependencies within each latent class. Albert et al. (2001) proposed a finite mixture (FM) formulation to flexibly model the dependence between markers. More recently, Xu and Craig (2009) proposed a probit latent class model that allows a general correlation structure between diagnostic markers.

Some authors considered a Bayesian approach to parameter estimation and inference in the latent class model. The crux of this approach is to augment the likelihood function (1.3) with latent disease data, and consider the *complete-data* likelihood

$$L(\mathbf{Y}, D|\boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ \pi P(\mathbf{Y}_i|D_i = 1, \boldsymbol{\theta}) \right\}^{d_i} \left\{ (1-\pi) P(\mathbf{Y}_i|D_i = 0, \boldsymbol{\theta}) \right\}^{1-d_i},$$

which allows derivation of the augmented data posterior (Tanner and Wong, 1987). Then the Gibbs sampler algorithm which alternates between sampling $\boldsymbol{\theta}$ and D from the respective full conditional distributions can be adopted to obtain marginal pos-

terior densities of the parameters (Joseph et al., 1995, Dendukuri and Joseph, 2001). Under the conditional independence assumption, Joseph et al. (1995) proposed a Bayesian approach to obtain interpretative posterior distributions for each of the diagnostic accuracy parameters relative to a given prior distribution. This Bayesian framework has been further extended to allow dependence between markers via fixed and random effect models (Dendukuri and Joseph, 2001), incorporate multiple latent variables (Dendukuri et al., 2009) and facilitate meta-analysis of the accuracy of the markers (Dendukuri et al., 2012).

However, caution must be exercised in the use of latent class models to estimate diagnostic accuracy (Albert and Dodd, 2004, Collins and Albert, 2016). Firstly, one should always check whether the model is identifiable, that is, its number of parameters does not exceed its degrees of freedom (Collins and Huynh, 2014). The conditional independence model is identifiable if $p \geq 3$, and the GRE and FM models are identifiable if $p \geq 4$ (Albert and Dodd, 2004). If the model is unidentifiable, a good strategy is to adopt a Bayesian approach, which can naturally incorporate available information about each parameter in the form of a prior distribution and allow distinguishing between the numerous possible solutions by updating of its posterior (Dendukuri and Joseph, 2001).

Secondly, one should be aware that estimates of diagnostic accuracy are biased under a misspecified dependence structure between markers, and that existing model diagnostic checking tools (e.g., likelihood comparison) may not be able to distinguish between dependence structures unless there are a very large number of markers (Albert and Dodd, 2004, Collins and Albert, 2016). One approach for improving model performance would be to exploit results from the best available reference markers with high diagnostic accuracy, if they exist (Albert, 2009). Given that there is reasonably high consensus consensus among the best available markers, Zhang et al. (2012) showed that the model becomes remarkably robust to misspecification of the

conditional dependence structure.

Continuous scale

A latent class modeling approach has been proposed to estimate ROC curves and AUC statistics in the absence of a gold standard (Choi et al., 2006a, Wang et al., 2006). In this approach the dependent marker values are assumed to be jointly normally distributed conditional on unknown/latent disease status, that is, $\mathbf{Y}_i \mid D_i = d_i \sim N_p(\boldsymbol{\mu}_{d_i}, \Sigma_{d_i})$, where $\boldsymbol{\mu}_{d_i} = [\mu_{d_i,1}, \ldots, \mu_{d_i,p}]^T$ is a vector of conditional means of the marker values and $\Sigma_{d_i} = \{\sigma^2_{d_i,uv}\}_{p\times p}$ $(u, v = 1, \ldots, p)$ is a conditional covariance matrix of the marker values with $\sigma^2_{d_i,uv} = \mathrm{Cov}(Y_{iu}, Y_{iv})$. Accordingly, the likelihood function (1.3) for this model becomes

$$L(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \big\{\pi\phi(\mathbf{y}_i; \boldsymbol{\mu}_1, \Sigma_1) + (1-\pi)\phi(\mathbf{y}_i; \boldsymbol{\mu}_0, \Sigma_0)\big\},$$

where $\phi(\cdot; \boldsymbol{\mu}, \Sigma)$ is the PDF of a multivariate normal variate with mean $\boldsymbol{\mu}$ and covariance Σ.

Given the parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma_0, \Sigma_1\}$, the ROC curve of the jth marker $(j = 1, \ldots p)$ can be constructed by plotting the following pairs over the range of cutoff values $c \in (-\infty, \infty)$:

$$\Big[1 - \Phi\Big(\frac{c - \mu_{0j}}{\sigma_{0j}}\Big), 1 - \Phi\Big(\frac{c - \mu_{1,jj}}{\sigma_{1,jj}}\Big)\Big].$$

The AUC for the jth marker can be computed as

$$\mathrm{AUC}_j = \Phi\left(\frac{\mu_{1j} - \mu_{0j}}{\sqrt{\sigma^2_{0,jj} + \sigma^2_{1,jj}}}\right).$$

Choi et al. (2006a) and Wang et al. (2006) proposed a Bayesian approach to estimate and make inferences about the ROC curves and the AUC statistics in the absence of

a gold standard. More recently, a Bayesian latent class model for combining multiple markers under no gold standard has been proposed (Yu et al., 2011, Jafarzadeh et al., 2016).

One of the main goals in a statistical study of diagnostic markers is to develop a simple screening method that clinicians can use to make decisions about the disease status of patient. It is traditional to dichotomize marker values at the cutoff point that optimizes a trade-off between sensitivity and specificity (Pepe, 2003). This approach, however, has been criticized due to an inherent information loss and issue of replicability in dichotomization (Altman and Royston, 2006, Royston et al., 2006). Several authors recommended the use of predictive probability of disease based on values of single or multiple markers as an alternative diagnostic criterion, both for patients in the current dataset and for hypothetical future patients in the absence of a gold standard (Choi et al., 2006b, Jones et al., 2009, Jafarzadeh et al., 2016). Here, a Bayesian classifier that allocates each patient based on the posterior probability of disease given his/her marker values is developed; that is, a patient with marker value $\mathbf{Y}_i$ is diagnosed with the disease if $\Pr(D_i = 1 \mid \mathbf{Y}_i) \geq k$, for some pre-specified probability $k \in (0, 1)$.

1.3 Motivating Data

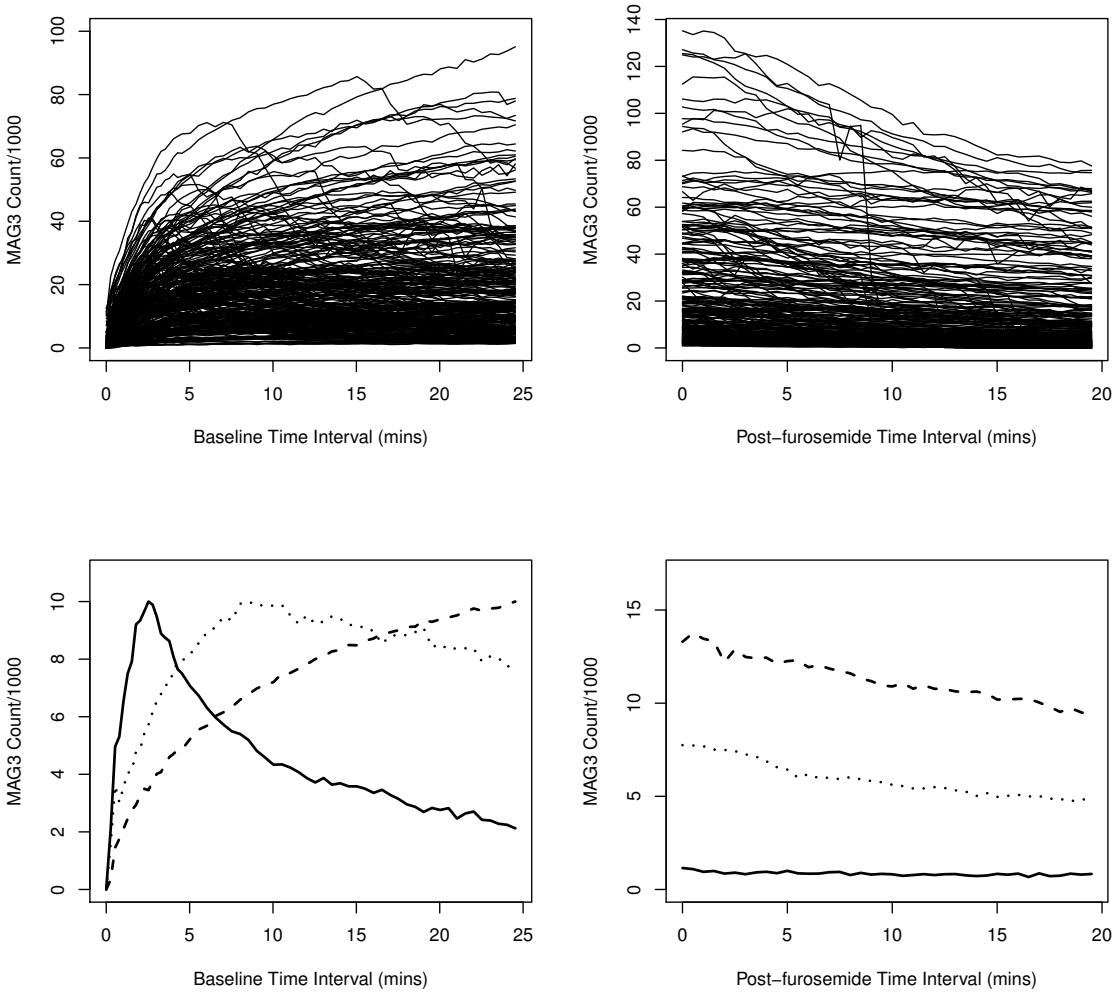
Obstruction to urine drainage from kidney (kidney obstruction) is a serious clinical problem that can lead to irreversible loss of renal function if not properly treated (Taylor, 2014). In recent years, diuresis renography have been widely adopted as an efficient, cost-effective and non-invasive approach to evaluate suspected kidney obstruction. Diuresis renography is performed by an intravenous injection of a gamma emitting tracer, ^{99m}Tc-Mercaptoacetyltriglycine (MAG3), that is rapidly removed from the blood by the kidneys and then travels down the ureters from the kidney to

the bladder. Photons emitted by tracer are then imaged and quantified in a region of interest (ROI) over each side of kidney, producing a set of renogram curves (Taylor et al., 2012). Baseline renogram curves are initially collected for patients referred for suspected obstruction. MAG3 photon counts over the region of interest (ROI) in each kidney are measured at 59 distinct time points over a period of 24 minutes. Each patient further receives an intravenous injection of furosemide, a potent diuretic, and a second (post-furosemide) renogram curve are obtained with an additional 20 minutes. Herein, MAG3 photon counts are measured at 40 time points using a framing rate of 30 seconds. The top left and right columns of Figure 1.1 respectively depict baseline and post-furosemide renogram curves of 275 kidneys stored in Emory University Hospital's archived database.

There are several important, interpretative patterns of the renogram curves that are known to strongly related to the renal function; for example, the speed of initial MAG3 uptake in the kidney, the rate of MAG3 excretion to the bladder, etc (Mettler and Guiberteau, 2012). To illustrate, consider the baseline renogram curve of a non-obstructed kidney in the bottom left panel of Figure 1.1 (see solid lines). The curve is characterized by a quick uptake and excretion of MAG3. On the other hand, the baseline renogram curve of an obstructed kidney is characterized by a prolonged period of MAG3 accumulation with no or poor excretion (see dashed lines in the bottom left panel of Figure 1.1), a trend which persists throughout the post-furosemide renogram (see dashed lines in the bottom right panel of Figure 1.1).

However, in practice, a high kidney-to-kidney variability in renogram curves is typical as seen from the top panel of Figure 1.1, and many show less distinctive patterns. For instance, the renogram of the "equivocal" kidney in Figure 1.1 (see dotted lines) show patterns somewhat between those of non-obstructed and obstructed kidneys. Therefore, in many cases, accurate diagnosis of kidney obstruction using quantitative features of renogram curves requires substantial expertise in renal physiology and

Figure 1.1: Top panel represents baseline (left) and post-furosemide (right) renogram curves of 275 kidneys. The bottom panel presents baseline (left) and post-furosemide (right) renogram curves of kidneys that are diagnosed as "non-obstructed" (solid lines), "obstructed" (dashed lines) and "equivocal" (dotted lines).



MAG3 pharmacokinetics (Taylor and Garcia, 2014). Unfortunately, a vast majority of diuresis renography scan interpretations are conducted by general radiologists in the United States at sites that perform fewer than 3 studies/week, and their lack of training and limited experience increase the error rate of the diagnosis (Taylor et al., 2008b, 2012, Taylor and Garcia, 2014, Taylor, 2014).

To assist practicing radiologists in limiting their errors and making correct inter-

pretation of kidney obstruction using diuresis renography, the researchers at Emory University undertook a project in which the goal was to develop decision support systems (DSS) and computer assisted diagnosis (CAD) tools (Taylor et al., 2008a). It is important to note that a gold standard for the detection of kidney obstruction, by which the CAD can be directly evaluated, is virtually nonexistent. A decision by a surgeon to operate or not operate may be considered a gold standard for the diagnosis of kidney obstruction; however, this surgical outcome is biased and cannot therefore be used, because it is directly influenced by the corresponding scan interpretation (obstructed versus non-obstructed). Thus, extra care is warranted when developing a CAD based on data collected from diuresis renography scans.

The study consisted of 275 kidneys from 145 patients (75 men [52%], 70 women [48%]; mean age, 58 years; SD, 16 years; range, 18-87 years), who were referred to the clinic with suspected kidney obstruction, and underwent a minor modification of the diuretic renography protocol recommended by an international consensus panel (O'Reilly et al., 1996). Baseline and post-furosemide renogram curve data were extracted from the ROI over each side of kidney. In addition, three nuclear medicine experts, each of whom had more than 20 years of experience in academic nuclear medicine, and three nuclear medicine residents, as a surrogate of practicing radiologists, were asked to provide both continuous ratings (from -1 to 1, with values approaching 1 indicating greater confidence in diagnosis of obstruction) and ordinal ratings (1: non-obstructed; 2: equivocal; 3: obstructed) on each kidney's obstruction status. Their ratings were based on a review of the images, renogram curve and their quantitative features, as well as other clinical variables. Note that although nuclear medicine experts are widely recognized to provide best available interpretations, inter-rater variability among them still exists as their interpretations do not always agree with each other (Taylor et al., 2008c).

1.4 Statistical Problems and Contributions

The overarching scientific goal of the Emory renal study described in Section 1.3 is two-fold: (1) to understand the reliability in experts and residents interpretations of kidney obstruction; and (2) to extract useful information from and evaluate the diagnostic utility of renogram curves for detection of kidney obstruction.

The need to assess agreement exists in various clinical studies where quantifying inter-rater reliability is of great importance. Use of unscaled agreement indices, such as total deviation index (TDI) and coverage probability (CP) are recommended for two main reasons: (1) they are intuitive in a sense that interpretations are tied to the original measurement unit; (2) practitioners can readily determine whether the agreement is satisfactory by directly comparing the value of the index to a prespecified tolerable coverage probability or absolute difference (Lin, 2000, Lin et al., 2002). However, the unscaled indices were only defined in the context of comparing two raters or multiple raters that assume homogeneity of variances across raters (Lin et al., 2007). However, this homogeneity is highly unlikely to hold in practice, especially when the goal especially when the goal is to assess inter-rater agreement among newly introduced raters with unknown measurement characteristics. For instance, every radiologist and expert has different experience and expertise, and there is no unscaled agreement index that can quantify agreement among heterogeneous multiple raters. In Chapter 2, we develop a set of new agreement indices based on root mean square of pairwise differences that can be used to quantify inter-rater reliability among multiple raters with heterogeneous measurement processes.

With advancements in technology, more and more cutting-edge, non-invasive medical devices are being used to diagnose and monitor diseases. The increasing complexity of data they generate, however, often pose unique statistical challenges for establishing a clinically interpretative relationship between the data-derived markers and disease pathology. In this dissertation, we specifically focus on one such type of

data, namely functional markers. The unit of observation of each functional marker is a smooth continuous curve (function) defined on a time or space continuum and its flexible and dynamic structure contains a rich source of clinical information (Ramsay and Silverman, 2005). It is thus typical in clinical research to describe and diagnose a disease using a set of "quantitative features" that characterize various dynamic, interpretative patterns of a functional marker, such as area under the curve, maximum value, time to reach maximum value and average velocity.

However, in many clinical settings, the selection and application of these features have been based on ad hoc blending of intuition and past practice without much scientific justification. For instance, in renal studies, although renogram curves and their several quantitative features (e.g., time to half MAG3 maximum) are frequently used to describe and diagnose kidney obstruction, establishing a scientifically justified relationship between these features and the underlying obstruction mechanism is of ongoing interest to prevent inappropriate patient management and unnecessary surgery (Bao et al., 2011, Taylor and Garcia, 2014). In Chapter 3, we develop a novel framework that can systematically extract various quantitative features of functional markers (renogram curves) and evaluate their diagnostic utility by rigorously assessing their alignment with an ordinal gold standard test (interpretations provided by nuclear medicine experts) based on BSA. In Chapter 4, we develop a novel statistical approach to assess the diagnostic accuracy of quantitative features based on AUC and describe the heterogeneity of AUC in different subpopulations by a sensible adaptation of a semi-parametric regression model.

In Chapter 5, we develop a novel statistical framework for systematically extracting dynamic changing patterns of functional markers via functional principal component analysis and evaluate their diagnostic utility absent gold standard under a latent binormal model. For multivariate functional markers, we propose to utilize a multivariate functional principal component analysis approach to characterize their

joint changing patterns. And if results from an imperfect reference test are available, we propose utilizing a functional partial least squares approach to exploit this information and achieve superior diagnostic performance.

Chapter 2

Overall Indices for Assessing Agreement Among Multiple Raters

Portions of this chapter were previously published as Jang JH, Manatunga AK, Taylor AT, Long Q. Overall indices for assessing agreement among multiple raters. *Statistics in Medicine*. 2018;37:4200–4215. https://doi.org/10.1002/sim.7912, and have been reproduced with permission.

2.1 Introduction

In various clinical studies, researchers are often interested in assessing agreement on clinical measurements taken on the same subjects using different raters. For continuous measurements, the use of a graphical method to plot the difference scores of two measurements against the mean for each subject has been advocated (Bland and Altman, 1986). However, this is a purely descriptive method and cannot provide inference regarding agreement. To overcome such limitations, scaled agreement indices, such as intraclass correlation coefficient (ICC) (Bartko, 1966), concordance correlation coefficient (CCC) (Lin, 1989) and its extensions (King and Chinchilli, 2001, Banhart et al., 2002, 2005, Lin et al., 2007) were introduced to assess agreement among two or more raters. Use of these scaled agreement indices has gained popularity in practice for their simplicity and ease of representation.

While being simple, scaled agreement indices have been criticized for several limitations. The main problem with these methods is that they are very sensitive to sample heterogeneity, sometimes resulting in counterintuitive interpretations (Bland and Altman, 1986, Atkinson and Nevill, 1997). For example, absurdly high values of ICC and CCC can be obtained even for a highly varied sample, where relative magnitude of between-subject variability to the total population variability is large. Moreover, scaled indices do not provide the interpretation terms of the original measurement unit and there is not a set ground for determining how high these indices should be in order to qualify as satisfying agreement.

As a formal alternative, unscaled agreement indices such as total deviation index (TDI) (Lin, 2000) and coverage probability (CP) (Lin et al., 2002) were introduced. TDI describes an acceptable/tolerable range of absolute difference such that a pre-specified proportion of the absolute differences between paired measurements is within the acceptable range. CP is a reciprocal concept, in which a proportion of the absolute differences within a pre-specified acceptable range is computed. Using unscaled indices has three key advantages: a) they provide direct intuitive interpretation tied with the original measurement unit; b) satisfactory agreement can be easily determined by directly comparing their values to a pre-specified acceptable range of distance or coverage probability; c) formal statistical inferences can be made based on their estimates. CP has been recommended as the preferred choice of agreement index for assessing reproducibility in a core lab setting (Banhart et al., 2016).

All of aforementioned unscaled agreement indices were only defined in the context of comparing a pair of raters. In the presence of multiple raters and replicated measurements for each subject, several extended unscaled agreement indices such as inter- and total-TDI (inter- and total-CP) have been proposed (Lin et al., 2007). These indices were expressed as functions of variance components through a mixed analysis of variance (ANOVA) model. Although it is possible to quantify agreement among multiple raters using inter- and total-TDI (inter- and total-CP), the ANOVA model assumption severely restricts the degree of heterogeneity that actual measurement processes may exhibit. Specifically, this assumption imposes a compound symmetry covariance structure shared by all measurements from different raters. However, it is highly unlikely for the assumption to hold in practice, especially when the goal is to assess inter-rater agreement among newly introduced raters with unknown measurement characteristics.

Consequently, a formal tool that is unscaled in nature and able to assess inter-rater agreement among multiple raters that exhibit highly heterogeneous variabilities

in measurement processes would be desirable, but has been lacking in literature. For example, data from a renal study demonstrate the need of such a statistical framework. In absence of a gold standard, it is generally accepted that the best available interpretation of renal scans comes from experienced experts, but inter-observer variability still exists as their interpretations do not always agree with each other (Taylor et al., 2008c). Practicing radiologists at U.S. hospitals often have marked variability in their interpretations compared to experienced readers due to the fact that their training in nuclear medicine was limited to 3-4 months (Taylor et al., 2008c, Taylor and Garcia, 2014). An analysis was thus carried out to quantify the interobserver agreement among practicing radiologists and better understand the nature of diagnostic variability present in a real-world clinical practice with renal scans. It is also of interest to determine if a new intervention with educational training called CAD (computer-assisted diagnosis) would reduce the interobserver variability among practicing radiologists. However, every radiologist and expert has different experience and expertise, and there is no unscaled agreement index that can incorporate possible heterogeneous variabilities in respective interpretation processes.

In this chapter, we propose a set of overall indices based on root mean square of pairwise differences (RMSPD) that are unscaled and can be used to assess agreement among multiple raters in the presence of heterogeneity of measurements. Recently, relative area under coverage probability curve (RAUCPC) was introduced as a summary agreement index that is scaled and summarizes agreement based on more than one pre-specified absolute differences (Banhart, 2016). Accordingly, we propose another overall agreement index based on RMSPD that is scaled and extends the concept of RAUCPC in the presence of multiple raters. A challenging aspect of using RMSPD to define an agreement index is that its explicit analytical expression for the inverse distribution is unavailable when a general covariance structure is considered. To this end, we propose to adopt an approximate distribution and the details are described

in Section 2.2. We propose maximum likelihood and bootstrap approaches for estimation and inference. In Section 2.3, we conduct simulation studies to evaluate the performance of the proposed approaches. In Section 2.4, we illustrate the application of our methods via application to a renal study. We present a summary in Section 2.5.

2.2 Methods

2.2.1 Existing unscaled and summary agreement indices for two raters

Let Y_1 and Y_2 be measurements from the same subject taken by first and second rater, respectively. Then the absolute difference $|D| = |Y_1 - Y_2|$ represents a distance, the extent to which paired measurements deviate from each other. Under this setting, higher proportion of paired measurements with smaller $|D|$ implies better agreement between the two raters. TDI is defined as the range of absolute difference between paired measurements such that a pre-specified proportion (π_0) of observations has absolute differences within that range Banhart et al. (2005). In other words, for $0 < \pi_0 < 1$, TDI_{π_0} is defined as the solution to $\pi_0 = P(|D| < \text{TDI}_{\pi_0}) = P(D^2 < \text{TDI}^2_{\pi_0})$, that is,

$$\text{TDI}_{\pi_0} = \sqrt{G^{-1}(\pi_0)},$$

where $G(\cdot)$ is the cumulative distribution function of D^2 and $G^{-1}(\cdot)$ is the inverse function of $G(\cdot)$.

CP is the reciprocal of TDI Lin et al. (2002). Here, the practitioner first specifies the maximum acceptable/tolerable range of absolute difference between paired measurements and computes the proportion of observations within this predetermined range. Let d denote the pre-specified acceptable absolute difference between paired

measurements. CP is defined as

$$\text{CP}_d = P(|D| < d) = P(D^2 < d^2) = G(d^2).$$

Recently, Banhart (2016) proposed relative area under the coverage probability (RAUCPC) as a summary agreement index between two raters in the presence of multiple acceptable absolute differences. The index is scaled in nature, but utilizes information based on a series of estimated coverage probabilities. For example, a practitioner may be interested in quantifying agreement based on certain varying acceptable distance criteria: a) $100\pi_0^{(1)}\%$ of observations should have absolute difference less than $d^{(1)}$; b) $100\pi_0^{(2)}\%$ of observations should have absolute difference less than $d^{(2)}$; c) $100\pi_0^{(3)}\%$ of observations should have absolute difference less than $d^{(3)}$. Denote $\delta_{\max}$ as *a priori* maximum acceptable range of distance such that $P(D < \delta_{\max}) \approx 1$. Rather than comparing CP_d to the preset $\pi_0^{(s)}$ at every pre-specified absolute difference $d^{(s)}$, $s = 1, 2, 3$, the area under the coverage probability curve $\text{CP}(d) = P(D < d)$, $0 \leq d \leq \delta_{\max}$, can be used for simultaneous comparison. Specifically, RAUCPC is defined as

$$\text{RAUCPC} = \frac{\int_0^{\delta_{\max}} \text{CP}(x)dx}{\delta_{\max}},$$

where the area under the coverage probability curve is scaled relative to $\delta_{\max}$ so that $0 \leq \text{RAUCPC} \leq 1$.

2.2.2 Overall agreement indices for multiple raters

Let Y_j denote a random variable representing a measurement from rater j, $(j = 1, ..., k)$. We assume that the $k \times 1$ vector of measurements $\mathbf{Y} = [Y_1, Y_2, ..., Y_k]^T$ has finite first and second moments with $k \times 1$ mean vector $\mu = [\mu_1, \mu_2, ..., \mu_k]$ and $k \times k$ covariance matrix $\mathbf{\Sigma}$. The covariance matrix $\mathbf{\Sigma}$ may take an unstructured form so that all k raters can exhibit heterogeneous measurement processes. In this manuscript, we

consider root mean square of pairwise differences (RMSPD) as an extended measure of distance that describes overall deviance among measurements taken by k (≥ 2) raters:

$$D_k = \sqrt{\frac{2}{k(k-1)} \sum_{1 \leq p < q \leq k} (Y_p - Y_q)^2}. \tag{2.1}$$

D_k is the square root of the average squared difference between all possible pairs of k raters, where the square root is taken to preserve the measurement unit. We note that D_k reduces to $|D|$ when $k = 2$, but in (2.1), we have expressed the deviation of measurements between any two raters in terms of the squared difference as opposed to the absolute difference used in conventional definitions of TDI and CP. The proposed RMSPD (D_k) basically summarizes the degree of deviation of measurements among multiple raters by taking into account all possible pairwise comparisons of their measurements based on the squared difference.

Based on (2.1), we propose a novel unscaled agreement index, the overall deviation index (ODI), for measuring agreement among k raters. For $0 < \pi_0 < 1$, $\text{ODI}_{\pi_0,k}$ is defined as the solution to $\pi_0 = P(D_k < \text{ODI}_{\pi_0,k}) = P(D_k^2 < \text{ODI}_{\pi_0,k}^2)$, that is,

$$\text{ODI}_{\pi_0,k} = \sqrt{F^{-1}(\pi_0)}, \tag{2.2}$$

where $F(\cdot)$ is the cumulative distribution of D_k^2 and $F^{-1}(\cdot)$ is the inverse function of $F(\cdot)$. Putting into words, this means that $100\pi_0\%$ of observations have RMSPDs among k raters smaller than or equal to $\text{ODI}_{\pi_0,k}$. Thus, the lower the ODI value, the better the agreement among measurements from multiple raters.

As in the case of the original CP, we propose the overall coverage probability (OCP) as the reciprocal of ODI. Initially, the acceptable RMSPD among k raters (d_k) is predetermined. Then the proportion of observations within this acceptable

range is computed to quantify agreement. Specifically, OCP is defined as

$$\text{OCP}_{d_k,k} = P(D_k < d_k) = P(D_k^2 < d_k^2) = F(d_k^2). \tag{2.3}$$

$\text{OCP}_{d_k,k}$ thus measures the proportion of observations that have RMSPDs among k raters less than or equal to d_k. Thus, higher OCP value suggests better agreement among measurements from multiple raters.

As for a scaled summary agreement index (Banhart, 2016) we propose to use the relative area under the overall coverage probability curve (RAUOCPC) in the presence of multiple raters. For example, consider the three varying acceptable distance criteria as presented in section 2.1. For the case of multiple raters, each absolute difference $d^{(s)}$ is now replaced by RMSPD $d_k^{(s)}$, $s = 1, 2, 3$. Denote $\delta_{\text{max},k}$ as *a priori* maximum acceptable RMSPD such that $P(D_k < \delta_{\text{max,k}}) \approx 1$. Rather than comparing $\text{OCP}_k(d_k)$ to the preset $\pi_0^{(s)}$ at every pre-specified RMSPD $d_k^{(s)}$, the area under the overall coverage probability curve $\text{OCP}_k(d_k) = P(D_k < d_k)$, $0 \leq d \leq \delta_{\text{max},k}$, can be used for simultaneous comparison. Specifically, RAUOCPC is defined as

$$\text{RAUOCPC}_k = \frac{\int_0^{\delta_{\text{max},k}} \text{OCP}_k(x)dx}{\delta_{\text{max},k}}, \tag{2.4}$$

so that $0 \leq \text{RAUOCPC} \leq 1$, with higher values indicating better agreement. Therefore, RAUOCPC can be used as a convenient tool to simultaneously compare each OCP to the multiple predetermined acceptable/tolerable RMSPDs.

Note that, when $k = 2$, $\text{ODI}_{\pi_0,2} = \text{TDI}_{\pi_0}$, $\text{OCP}_{d_2,2} = \text{CP}_d$ and $\text{RAUOCPC}_2 = \text{RAUCPC}$. Thus, the ODI, OCP and RAUOCPC are natural extensions of TDI, CP and RAUCPC, respectively.

Parametrization

In previous literature Lin (2000), Lin et al. (2002), Banhart (2016), D was assumed to follow a normal distribution (D^2 to follow a non-central chi-square distribution) in order to define, estimate and perform inference on TDI, CP and RAUCPC. Likewise, formulations of ODI, OCP and RAUOCPC as in definitions (2.2), (2.3) and (2.4), respectively, require an appropriate parametrization of $F(d_k^2)$ and its inverse $F^{-1}(\pi_0)$.

Define a $(k-1)\times k$ matrix $\mathbf{A} = \{a_{u,v}\}_{(k-1)\times k}$, where $a_{u,v} = 1$ for $u = v$, $a_{u,v} = -1$ for $u+1 = v$ and $a_{u,v} = 0$ otherwise. Then $\mathbf{X} = \mathbf{AY} = [Y_1 - Y_2, Y_2 - Y_3, ..., Y_{k-1} - Y_k]^T = [X_1, X_2, ...X_{k-1}]^T$ represents the $(k-1)\times 1$ vector of distinct pairwise differences with $(k-1)\times 1$ mean vector $\boldsymbol{\mu_d} = \mathbf{A}\boldsymbol{\mu} = [\mu_1 - \mu_2, \mu_2 - \mu_3, ..., \mu_{k-1} - \mu_k]^T = [\mu_{d1}, \mu_{d2}, ..., \mu_{d,k-1}]^T$ and $(k-1)\times(k-1)$ covariance matrix $\boldsymbol{\Sigma_d} = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A^T}$. We assume that $\mathbf{X}$ is normally distributed as $\mathrm{MN}_{k-1}(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$, a weaker assumption than imposing normality on $\mathbf{Y}$. Then D_k^2 can be expressed as a quadratic form in normal variates $\mathbf{X}$ (see Appendix A.1). Specifically,

$$D_k^2 = \frac{2}{k(k-1)} \sum_{1\le p<q\le k} (Y_p - Y_q)^2 = \mathbf{X^T}\Big\{\frac{2}{k-1}(\mathbf{AA^T})^{-1}\Big\}\mathbf{X} = \mathbf{X^T B X}, \tag{2.5}$$

where $\mathbf{B} = \frac{2}{k-1}(\mathbf{AA^T})^{-1}$ with $\mathrm{rank}(\mathbf{B}) = k-1$. If $\boldsymbol{\Sigma_d}$ is non-singular, it can be shown that the exact distributional form of D_k^2 can be expressed as a weighted sum of chi-square variables (Imhof, 1961):

$$D_k^2 \quad \sim \quad \sum_{r=1}^{k-1} \lambda_r \chi^2_{h_r,\delta_r}. \tag{2.6}$$

The λ_r are the distinct non-zero eigenvalues of $\mathbf{B}\boldsymbol{\Sigma_d}$, the h_r their respective orders of multiplicity, the δ_r are squares of certain linear combinations of $\mu_{d1}, \mu_{d2}, ..., \mu_{d,k-1}$, the $\chi^2_{h_r,\delta_r}$ are independent non-central chi-square random variables with h_r degrees of freedom and non-centrality parameter δ_r.

However, computing $F(d_k^2)$ and its inverse $F^{-1}(\pi_0)$ using exact distributional form (2.6) is not straightforward except in some special cases. In order to readily define and estimate the proposed overall unscaled agreement indices, we propose to adopt the approximate distribution of D_k^2 as opposed to its exact distribution (see Section 2.5 for a more detailed discussion). Specifically, we propose to approximate $F(d_k^2)$ and $F^{-1}(\pi_0)$ using a single non-central chi-square random variable $\chi^2_{l,\delta}$, where the degrees of freedom l and the non-centrality parameter δ are determined by the first four cumulants of D_k^2 Liu et al. (2009). Specifically, let κ_t denote t^{th} cumulant of D_k^2. Then κ_t can be directly expressed as a function of parameters $(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ from the assumed multivariate normal distribution on the distinct pairwise differences (Liu et al., 2009, Provost and Mathai, 1992):

$$\kappa_t(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}) = 2^{t-1}(t-1)!\Big[\text{trace}\{(\mathbf{B}\boldsymbol{\Sigma_d})^t\} + t\boldsymbol{\mu_d}^T(\mathbf{B}\boldsymbol{\Sigma_d})^{t-1}\mathbf{B}\boldsymbol{\mu_d}\Big].$$

Accordingly, the mean, standard deviation, skewness and kurtosis of the distribution of D_k^2 can be defined in terms of the cumulants. We omit $(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ for ease of representation:

$$\mu_Q = \kappa_1, \quad \sigma_Q = \sqrt{\kappa_2}, \quad \beta_1 = \frac{\kappa_3}{\kappa_2^{3/2}} \quad \beta_2 = \frac{\kappa_4}{\kappa_2^2}.$$

We can initially write

$$F(d_k^2) = P(D_k^2 < d_k^2) = P\left(\frac{D_k^2 - \kappa_1}{\sqrt{\kappa_2}} < \frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right).$$

Then, the above probability can be approximated using a single non-central chi-square random variable $\chi^2_{l,\delta}$ as

$$P\left(\frac{\chi^2_{l,\delta} - \mu^*}{\sigma^*} < \frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right) = P\left\{\chi^2_{l,\delta} < \left(\frac{d_k^2 - \kappa_1}{\sqrt{\kappa_2}}\right)\sigma^* + \mu^*\right\},$$

where $\mu^* = \mathrm{E}(\chi^2_{l,\delta}) = l+\delta$ and $\sigma^* = \mathrm{SD}(\chi^2_{l,\delta}) = \sqrt{2(l+2\delta)}$. Here, parameters l and δ are determined so that skewnesses of D_k^2 and $\chi^2_{l,\delta}$ are equal and the difference between their kurtoses are minimized. Let $s_1 = \kappa_3/\sqrt{8}\kappa_2^{3/2}$, $s_2 = \kappa_4/12\kappa_2^2$ and $a = \sqrt{l+2\delta}$. It can be shown that if $s_1^2 > s_2$ Liu et al. (2009),

$$a = \frac{1}{(1-\sqrt{s_1^2 - s_2})}, \quad \delta = s_1 a^3 - a^2 \quad \text{and} \quad l = a^2 - 2\delta,$$

and if $s_1^2 \leq s_2$,

$$a = \frac{1}{s_1}, \quad \delta = 0 \quad \text{and} \quad l = a^2.$$

Thus, $F(d_k^2)$ and its inverse $F^{-1}(\pi_0)$ can be approximated as

$$F(d_k^2) \approx$$
$$\chi^2\left\{\left(\frac{d_k^2 - \kappa_1(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})}{\sqrt{\kappa_2(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})}}\right)\sqrt{2}a(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}) + l(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}) + \delta(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}), l(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}), \delta(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})\right\}$$

and

$$F^{-1}(\pi_0) \approx \frac{\sqrt{\kappa_2(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})}[\chi^{2(-1)}\{\pi_0, l(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}), \delta(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})\} - l(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}) - \delta(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})]}{\sqrt{2}a(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})}$$
$$+ \kappa_1(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d}),$$

where $\chi^2(\cdot, l, \delta)$ is the cumulative distribution function of the non-central chi-square distribution with l degrees of freedom and non-centrality parameter δ, and $\chi^{2(-1)}(\cdot, l, \delta)$ is the inverse function of $\chi^2(\cdot, l, \delta)$. It is important to note that both quantities are completely determined by the parameters $(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$.

By adopting the proposed parametrization and plugging in approximated values of $F(d_k^2)$ and $F^{-1}(\pi_0)$, definitions (2.2), (2.3) and (2.4) become

$$\mathrm{ODI}_{\pi_0,k} = \left[\frac{\sqrt{\kappa_2}\{\chi^{2(-1)}(\pi_0, l, \delta) - l - \delta\}}{\sqrt{2}a} + \kappa_1\right]^{1/2}, \tag{2.7}$$

$$\mathrm{OCP}_{d_k,k} = \chi^2\Big\{\Big(\frac{d_k^2-\kappa_1}{\sqrt{\kappa_2}}\Big)\sqrt{2}a + l + \delta, l, \delta\Big\}, \tag{2.8}$$

and

$$\mathrm{RAUOCPC}_k = \frac{\displaystyle\int_0^{\delta_{\max,k}} \chi^2\left\{\left(\frac{x^2-\kappa_1}{\sqrt{\kappa_2}}\right)\sqrt{2}a + l + \delta, l, \delta\right\}dx}{\delta_{\max,k}}. \tag{2.9}$$

Compound Symmetry Case

Suppose $\mathbf{Y}$ has a mean vector $\boldsymbol{\mu}$ and a compound symmetry covariance structure $\boldsymbol{\Sigma} = \sigma^2(1-\rho)\mathbf{I_k} + \sigma^2\rho\mathbf{1_k}\mathbf{1_k^T}$. Note that $\mathbf{I_k}$ is a $k \times k$ identity matrix and $\mathbf{1_k}$ is a $k \times 1$ vector with only 1's as its elements. This represents the case in which multiple raters share common variabilities in respective measurement processes. Assume that the vector of distinct pairwise differences $\mathbf{X}$ is normally distributed as $\mathrm{MN}_{k-1}(\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ with $\boldsymbol{\Sigma_d} = \mathbf{A\Sigma A^T} = \sigma^2(1-\rho)\mathbf{AA^T}$. Consequently, by (2.5) and the relationship between normal and chi-square distribution, and noting that $\mathbf{\Sigma_d^{-1}} = \frac{1}{\sigma^2(1-\rho)}(\mathbf{AA^T})^{-1}$, $D_k^2 = \mathbf{X^T\Sigma_d^{-1}X} = \frac{\sum_{1\le p<q\le k}(Y_p-Y_q)^2}{k\sigma^2(1-\rho)} \sim \chi^2_{k-1,\gamma}$, where $\chi^2_{k-1,\gamma}$ denotes a non-central a chi-square random variable with $k-1$ degrees of freedom and non-centrality parameter $\gamma = \frac{\sum_{1\le p<q\le k}(\mu_p-\mu_q)^2}{k\sigma^2(1-\rho)}$. In other words, when measurements follow a compound symmetry covariance structure, $F(d_k^2)$ and its inverse $F^{-1}(\pi_0)$ can be computed using exact distributional form as

$$F(d_k^2) = P\left\{\frac{(k-1)D_k^2}{2\sigma^2(1-\rho)} < \frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}\right\} = \chi^2\left\{\frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1\le p<q\le k}(\mu_p-\mu_q)^2}{k\sigma^2(1-\rho)}\right\}$$

and

$$F^{-1}(\pi_0) = \left\{\frac{2\sigma^2(1-\rho)}{k-1}\right\}\chi^{2(-1)}\left\{\pi_0, k-1, \frac{\sum_{1\le p<q\le k}(\mu_p-\mu_q)^2}{k\sigma^2(1-\rho)}\right\}.$$

By adopting the exact parametrization assuming compound symmetry covariance

structure, definitions (2.2), (2.3) and (2.4) become

$$\mathrm{ODI}^{(C)}_{\pi_0,k} = \left[\left\{ \frac{2\sigma^2(1-\rho)}{k-1} \right\} \chi^{2(-1)} \left\{ \pi_0, k-1, \frac{\sum_{1\le p<q\le k} (\mu_p - \mu_q)^2}{k\sigma^2(1-\rho)} \right\} \right]^{1/2}, \tag{2.10}$$

$$\mathrm{OCP}^{(C)}_{d_k,k} = \chi^2 \left\{ \frac{(k-1)d_k^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1\le p<q\le k} (\mu_p - \mu_q)^2}{k\sigma^2(1-\rho)} \right\}, \tag{2.11}$$

and

$$\mathrm{RAUOCPC}^{(C)}_{k} = \frac{\int_0^{\delta_{\max,k}} \chi^2 \left\{ \frac{(k-1)d_{ik}^2}{2\sigma^2(1-\rho)}, k-1, \frac{\sum_{1\le p<q\le k} (\mu_p - \mu_q)^2}{k\sigma^2(1-\rho)} \right\} dx}{\delta_{\max,k}}. \tag{2.12}$$

When there are no replicates from respective raters, definitions (2.10) and (2.11) are the same quantities as Inter-TDI and Inter-CP (or Total-TDI and Total-CP) proposed by Lin et al. (2007), respectively, which are based on the mixed ANOVA model.

2.2.3 Estimation

Let $\mathbf{Y_i}$ $(i = 1, ..., n)$ be the vector of measurements for subject i. Then $\mathbf{X_i} = \mathbf{A}\mathbf{Y_i}$ is the vector of distinct pairwise differences for the same subject. Denote $\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ as the (bias-adjusted) maximum likelihood estimators for the parameters in $\mathrm{MN}(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$ distribution. Specifically, $\hat{\boldsymbol{\mu}}_{\boldsymbol{d}} = [\hat{\mu}_{d1}, \hat{\mu}_{d2}, ..., \hat{\mu}_{d,k-1}]^T = [\overline{X}_1, \overline{X}_2, ..., \overline{X}_{k-1}]^T$
$= \frac{1}{n}\Big[\sum_{i=1}^{n} X_{i1}, \sum_{i=1}^{n} X_{i2}, ..., \sum_{i=1}^{n} X_{i,k-1} \Big]^T$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{d}})(\boldsymbol{X_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{d}})^T$.

We propose to estimate $\mathrm{ODI}_{\pi_0,k}$, $\mathrm{OCP}_{d_k,k}$ and $\mathrm{RAUOCPC}_k$ by replacing parameters $\kappa_1(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$, $\kappa_2(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$, $a(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$, $l(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$ and $\delta(\boldsymbol{\mu}_{\boldsymbol{d}}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$ in definitions (2.7), (2.8) and (2.9) by their maximum likelihood (ML) estimates $\hat{\kappa_1} = \kappa_1(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$, $\hat{\kappa_2} = \kappa_2(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$, $\hat{a} = a(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$, $\hat{l} = l(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$ and $\hat{\delta} = \delta(\hat{\boldsymbol{\mu}}_{\boldsymbol{d}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})$. Therefore by the invariance property of ML estimators, we can express the ML estimators of the

proposed overall unscaled agreement indices as

$$\widehat{\text{ODI}}_{\pi_0,k} = \left[\frac{\sqrt{\hat{\kappa_2}}\{\chi^{2(-1)}(\pi_0, \hat{l}, \hat{\delta}) - \hat{l} - \hat{\delta}\}}{\sqrt{2}\hat{a}} + \hat{\kappa_1}\right]^{1/2}, \tag{2.13}$$

$$\widehat{\text{OCP}}_{d_k,k} = \chi^2\Big\{\Big(\frac{d_k^2 - \hat{\kappa_1}}{\sqrt{\hat{\kappa_2}}}\Big)\sqrt{2}\hat{a} + \hat{l} + \hat{\delta}, \hat{l}, \hat{\delta}\Big\}, \tag{2.14}$$

and

$$\widehat{\text{RAUOCPC}}_k = \frac{\displaystyle\int_0^{\delta_{\max,k}} \chi^2\left\{\left(\frac{x^2 - \hat{\kappa_1}}{\sqrt{\hat{\kappa_2}}}\right)\sqrt{2}\hat{a} + \hat{l} + \hat{\delta}, \hat{l}, \hat{\delta}\right\}dx}{\hat{\delta}_{\max,k}}. \tag{2.15}$$

RAUOCPC can also be estimated using the non-parametric method as suggested by Banhart for the RAUCPC case (Banhart, 2016). We first order the unique observed RMSPDs among k raters as $d_{1,k} < d_{2,k} < ... < d_{n,k}$ with $d_{n,k} < \delta_{\max,k}$. Define $ocp_{1,k} < ocp_{2,k} < ... < ocp_{n,k}$ as the estimated overall coverage probabilities, where $ocp_{i,k}$ denotes the proportion of all possible RMSPDs less than or equal to $d_{i,k}, i = 1, 2, ..., n$. Then the empirical overall coverage probabilities can be drawn as a series of straight lines connecting $(d_{0,k}, ocp_{0,k}), (d_{1,k}, ocp_{1,k}), ..., (d_{n,k}, ocp_{n,k}), (d_{n+1,k}, ocp_{n+1,k})$ where $d_{0,k} = 0, ocp_{0,k} = 0, d_{n+1,k} = \delta_{\max,k}$ and $ocp_{n+1,k} = ocp_{n,k}$. The non-parametric estimator for RAUOCPC_k is the area under these straight lines scaled by $\delta_{\max,k}$. Specifically, the estimator is given as

$$\widehat{\text{RAUOCPC}}_k^{\text{non-parm}} = \frac{\displaystyle\sum_{i=1}^{n+1}(d_i - d_{i-1})(ocp_{i-1} + \tfrac{ocp_i - ocp_{i-1}}{2})}{\delta_{\max,k}}. \tag{2.16}$$

2.2.4 Inference

One-Sample

Let θ_k be one of the three proposed overall unscaled agreement indices and $\boldsymbol{\beta} = (\boldsymbol{\mu_d}, \boldsymbol{\Sigma_d})$ be its associated parameter. Denote $\hat{\theta}_k$ and $\hat{\boldsymbol{\beta}}$ as their ML estimators, re-

spectively. Also let I be the observed Fisher information matrix for $\boldsymbol{\beta}$ and $G = \frac{\partial g\{\theta_k(\boldsymbol{\beta})\}}{\partial \boldsymbol{\beta}}\Big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ be the gradient vector of the index evaluated at the ML estimator, where g represents a monotone transformation of the parameter that is adopted to accelerate convergence to asymptotic normality. Since $\mathrm{ODI}_{\pi_0,k} \in [0,\infty)$ and $\mathrm{OCP}_{d_k,k}, \mathrm{RAUOCPC}_k \in [0,1]$, we use the natural log transformation for the former index, and the logit transformation for the latter indices. Then, from asymptotic normality of ML estimators and delta method, we have $g(\hat{\theta}_k) \sim AN(g(\theta_k), G^T I^{-1} G)$, where $(G^T I^{-1} G)^{1/2} = \widehat{\mathrm{SE}}\{g(\hat{\theta}_k)\}$ denotes the standard error estimate.

Since the analytical form of $(G^T I^{-1} G)^{1/2}$ is complicated, we propose bootstrap approach for standard error estimation. Specifically, we can take B bootstrap samples from the observed data at the subject level with replacement, compute $g(\hat{\theta}_k)^{(b)}$ for each bootstrap sample $b = 1, 2, ..., B$, and obtain bootstrap estimate of the standard error,

$$\widehat{\mathrm{SE}}_B\{g(\hat{\theta}_k)\} = \left[\frac{1}{B}\sum_{b=1}^{B}\left\{g(\hat{\theta}_k)^{(b)} - \overline{g(\hat{\theta}_k)}_B\right\}^2\right]^{1/2}, \tag{2.17}$$

where $\overline{g(\hat{\theta}_k)}_B = \frac{1}{B}\sum_{b=1}^{B} g(\hat{\theta}_k)^{(b)}$. Note that sampling on the subject level is essential as we should account for correlated measurements within a subject.

Suppose we postulate that agreement among k raters based on the ODI is satisfactory if approximately $100\pi_0\%$ of observations have RMSPDs among the raters less than a predetermined constant L_0. Here, L_0 denotes the maximum RMSPD that we are willing to tolerate, and accept that all k raters exhibit homogeneity in the measurement processes. We would accept satisfactory agreement with a type I error α if the $100(1-\alpha)\%$ upper confidence limit of $\mathrm{ODI}_{\pi_0,k}$, that is,

$$U_{\mathrm{ODI}_{\pi_0,k,1-\alpha}} = \exp\Big\{\log(\widehat{\mathrm{ODI}}_{\pi_0,k}) + z_{1-\alpha}(\widehat{\mathrm{SE}}_B\{(\log(\widehat{\mathrm{ODI}}_{\pi_0,k})\}\Big\}, \tag{2.18}$$

is less than L_0, where bootstrap standard error estimate is calculated from (2.17) and

$z_{1-\alpha}$ denotes the $100(1-\alpha)^{\text{th}}$ percentile of a standard normal distribution.

For the OCP, the lower confidence limit is preferably calculated because ensuring acceptable agreement with respect to OCP often involves a null proportion π_0, which we would deem too small as to conclude satisfactory agreement. Specifically, using the bootstrap standard error estimate (2.17), the $100(1-\alpha)\%$ lower confidence limit of $\text{OCP}_{d_k,k}$ is computed as

$$L_{\text{OCP}_{d_k,k,1-\alpha}} = h\Big[\text{logit}(\widehat{\text{OCP}}_{d_k,k}) - z_{1-\alpha}\widehat{\text{SE}}_B\Big\{\text{logit}(\widehat{\text{OCP}}_{d_k,k})\Big\}\Big], \tag{2.19}$$

where $h(\cdot) = \frac{exp(\cdot)}{1+exp(\cdot)}$. Then given type I error rate of α and tolerable RMSPD d_k, we accept that k raters produce reasonably homogeneous ratings on a given subject if (2.19) is greater than π_0.

For the RAUOCPC, consider three multiple acceptable RMSPDs $(d_k^{(1)}, d_k^{(2)}, d_k^{(3)})$ and denote $\delta_{\text{max},k}$ as *a priori* maximum acceptable RMSPD such that $P(D_k < \delta_{\text{max,k}}) \approx 1$. Initially, the area under the $\widehat{\text{OCP}}_k(d_k)$, $0 \le d_k \le \delta_{\text{max},k}$, can be visually compared to the area under straight lines that connect points formed by a series of preset RMSPDs with the corresponding overall coverage probabilities, for example connecting points (0,0), $(d_k^{(1)}, \pi_0^{(1)})$, $(d_k^{(2)}, \pi_0^{(2)})$, $(d_k^{(3)}, \pi_0^{(3)})$ and $(\delta_{\text{max},k}, 1)$. The larger size of the former area would suggest satisfying agreement among k raters. Testing whether the difference between sizes of the two areas is statistically significant is equivalent to testing whether RAUOCPC_k is greater than T_0, which denotes the size of the latter area scaled by $\delta_{\text{max},k}$. Thus, we can focus on deriving the $100(1-\alpha)\%$ lower boundary. Specifically, after obtaining the bootstrap standard error estimate from (2.17), the $100(1-\alpha)\%$ lower confidence limit of RAUOCPC_k is computed as

$$L_{\text{RAUOCPC}_{k,1-\alpha}} = h\Big[\text{logit}(\widehat{\text{RAUOCPC}}_k) - z_{1-\alpha}\widehat{\text{SE}}_B\Big\{\text{logit}(\widehat{\text{RAUOCPC}}_k)\Big\}\Big]. \tag{2.20}$$

If (2.20) is greater than T_0, we can conclude satisfactory agreement among k raters

based on the three varying acceptable distance criteria. Note that $100(1-\alpha)\%$ lower boundary based on the non-parametric RAUOCPC estimate can be computed in a similar manner.

Two-Sample

Now suppose we are interested in comparing degrees of inter-rater agreement among measurements on the same set of subjects between two groups of raters using one of the three proposed overall unscaled indices. This scenario often arises when the goal of a study is to evaluate the performance of a group of new raters relative to a group of best standard raters in terms of inter-rater agreement. Suppose $\theta_k^{(1)}$ and $\theta_k^{(2)}$ measure agreement among the first and second groups of k raters, respectively. We form a null hypothesis

$$H_0 : \theta_k^{(1)} = \theta_k^{(2)}, \quad \text{or equivalently,} \quad H_0 : g(\theta_k^{(1)}) = g(\theta_k^{(1)}),$$

against the alternative hypothesis

$$H_1 : \theta_k^{(1)} \neq \theta_k^{(2)}, \quad \text{or equivalently,} \quad H_0 : g(\theta_k^{(1)}) \neq g(\theta_k^{(1)}).$$

Using the asymptotic property of ML estimators, we can formulate the Wald test statistic as

$$\frac{g(\hat{\theta}_k^{(1)}) - g(\hat{\theta}_k^{(2)})}{\text{SE}\{g(\hat{\theta}_k^{(1)}) - g(\hat{\theta}_k^{(2)})\}} = \frac{g(\hat{\theta}_k^{(D)})}{\text{SE}\{g(\hat{\theta}_k^{(D)})\}} \sim AN(0,1),$$

under the null hypothesis. Since the analytical form of the standard error is complicated, we estimate the standard error by bootstrap approach. See Appendix A.2 for detailed steps of the two-sample hypothesis testing procedure.

2.3 Simulations

We conducted simulation studies to assess the performance of the proposed approaches to evaluate agreement via overall unscaled agreement indices. We assumed that there are four raters and the data are generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ and covariance matrix $\boldsymbol{\Sigma}$, for both compound symmetry and unstructured scenarios. Under both scenarios, a total of 1000 simulated data sets were generated. We considered sample sizes of 20 and 60. To evaluate the performance of inference based on bootstrap approach, 1000 bootstrap samples were used to compute standard error estimates and one-sided 95% confidence intervals.

Under the compound symmetry scenario, data were generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4) = (3.5, 3.5, 3.5, 4.2)$ and compound symmetry covariance matrix with common variance $\sigma^2 = 0.25$ and common correlation coefficient $\rho = 0.8$. Specifically, the compound symmetry covariance matrix was defined as $\boldsymbol{\Sigma} = \{\sigma_{u,v}\}$ where $\sigma_{u,v} = 0.25$ if $u = v$ and $\sigma_{u,v} = 0.2$ if $u \neq v$, $u, v = 1, 2, 3, 4$. Thus first three raters exhibit homogeneity in respective measurement processes. However, the fourth rater represents a heterogeneous measurement process in which ratings are consistently overestimated as evidenced by its larger mean.

Under the unstructured scenario, data were generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4) = (3.5, 3.5, 3.5, 4.2)$ and unstructured covariance matrix defined as $\boldsymbol{\Sigma} = \{\sigma_{u,v}\}$, where $\sigma_{1,1} = \sigma_{2,2} = \sigma_{3,3} = 0.30$, $\sigma_{1,2} = \sigma_{1,3} = \sigma_{2,3} = 0.24$, $\sigma_{4,4} = 0.15$ and $\sigma_{1,4} = \sigma_{2,4} = \sigma_{3,4} = 0.08$. Again, the first three raters exhibit homogeneity in measurement process with common covariances among themselves. However, the fourth rater is allowed to exhibit stronger heterogeneity in the measurement process by adding further flexibility in the data generation process. Not only its mean is larger compared to first three, but its variance is smaller, and its

linear relationship with others is relatively weak as evidenced by smaller covariances in relation to other raters.

Table 2.1: Simulation results for ODI based on 1000 simulated data sets under compound symmetry (CS) and unstructured scenarios (UN). "Relative bias" represents sample mean of 1000 values of $100 * \{(\widehat{\text{ODI}} - \text{ODI})/\text{ODI}\}$, where $\widehat{\text{ODI}}$s are obtained through anti-transformations. "Std of estimate" represents standard deviation of 1000 $\widehat{\text{ODI}}$s. "Mean of SE" estimate represents mean of 1000 bootstrap standard errors. "CP" represents the proportion of 1,000 estimated 95% upper confidence limits computed by (2.18) that are greater than the true value.

Scenario	n	Statistics	True value	Relative bias(%)	Std of estimate	Mean of SE estimate	CP
CS	20	$\text{ODI}_{0.80,4}$	0.7056	-0.2474	0.0642	0.0611	0.916
		$\text{ODI}^{(C)}_{0.80,4}$		-0.5809	0.0608	0.0597	0.923
		$\text{ODI}_{0.90,4}$	0.7827	0.4463	0.0659	0.0631	0.932
		$\text{ODI}^{(C)}_{0.90,4}$		-0.7953	0.0608	0.0587	0.922
	60	$\text{ODI}_{0.80,4}$	0.7056	-0.0707	0.0369	0.0364	0.942
		$\text{ODI}^{(C)}_{0.80,4}$		-0.0731	0.0368	0.0357	0.940
		$\text{ODI}_{0.90,4}$	0.7827	0.0493	0.0382	0.0376	0.941
		$\text{ODI}^{(C)}_{0.90,4}$		-0.0229	0.0346	0.0348	0.946
UN	20	$\text{ODI}_{0.80,4}$	0.8416	0.2438	0.0966	0.0952	0.923
		$\text{ODI}^{(C)}_{0.80,4}$		-3.5240	0.0940	0.0896	0.854
		$\text{ODI}_{0.90,4}$	0.9855	0.0933	0.1025	0.0948	0.910
		$\text{ODI}^{(C)}_{0.90,4}$		-6.7998	0.0918	0.0867	0.764
	60	$\text{ODI}_{0.80,4}$	0.8416	-0.0169	0.0565	0.0565	0.944
		$\text{ODI}^{(C)}_{0.80,4}$		-3.0690	0.0548	0.0534	0.841
		$\text{ODI}_{0.90,4}$	0.9855	0.1131	0.0566	0.0563	0.948
		$\text{ODI}^{(C)}_{0.90,4}$		-6.4116	0.0519	0.0515	0.621

Table 2.1 shows the simulation results for ODI and $\text{ODI}^{(C)}$ under both scenarios. We considered two widely used pre-specified coverage probabilities: $\pi_0 = 0.80$ and 0.90. Under the compound symmetry scenario, both ODI and $\text{ODI}^{(C)}$ estimates

yielded almost identical results, as expected for normal data with compound symmetry covariance structure. All absolute relative biases are less than 1%, implying that our proposed estimation approach provides reasonable unbiased estimates even for a sample size as small as 20. The 95% coverage is slightly less than the nominal level for sample size of 20, possibly due to the underestimation of standard errors as compared to empirical counterparts. The 95% coverage is very close to the nominal level for sample sizes of 60 or greater (not shown). Under the unstructured scenario, ODI estimates again have negligible bias. The coverage of true ODI based on one-sided 95% confidence interval is 92% or less with sample size of 20, but a coverage of 94% or 95% is generally achieved for sample sizes of 60 or greater. However, $\text{ODI}^{(C)}$ estimates are biased underestimating the true ODI. This results in very poor coverage probabilities, especially when $\pi_0 = 0.90$. Under both scenarios, estimated standard errors rapidly approach their empirical counterparts as the sample size increases, confirming that our bootstrap procedure provides valid and robust standard error estimates.

Table 2.2 shows the simulation results for OCP and $\text{OCP}^{(C)}$ under both scenarios. We considered two different preset RMSPDs among four raters: $d_4 = 0.8$ and 0.9. Under the compound symmetry scenario, both OCP and $\text{OCP}^{(C)}$ estimates show good performances, as expected under the correct model specification. Each absolute relative bias is less than 1%, indicating the consistency of the point estimates. The 95% coverage is slightly greater than the nominal level, possibly due to the overestimation of standard errors. However, coverage probabilities are all generally around 95% in almost every situation. Under the unstructured scenario, OCP estimates are virtually unbiased even for a small sample size of 20. The 95% coverage of true OCP is close to the nominal level for all sample sizes. However, $\text{OCP}^{(C)}$ estimates are biased overestimating the true OCP, and the bias in fact increases for larger sample sizes. As a result, coverage probabilities are in general noticeably below the desired nominal rate, especially when $d_4 = 0.9$. We should also remark that standard error

Table 2.2: Simulation results for OCP based on 1000 simulated data sets under compound symmetry (CS) and unstructured scenarios (UN). "Relative bias" represents sample mean of 1000 values of $100 * \{(\widehat{\text{OCP}} - \text{OCP})/\text{OCP}\}$, where $\widehat{\text{OCP}}$s are obtained through anti-transformations. "Std of estimate" represents standard deviation of 1000 $\widehat{\text{OCP}}$s. "Mean of SE" estimate represents mean of 1000 bootstrap standard errors. "CP" represents the proportion of 1,000 estimated 95% lower confidence limits computed by (2.19) that are smaller than the true value.

Scenario	n	Statistics	True value	Relative bias(%)	Std of estimate	Mean of SE estimate	CP
CS	20	$\text{OCP}_{0.80,4}$	0.9162	-0.6576	0.6210	0.6339	0.959
		$\text{OCP}^{(C)}_{0.80,4}$		0.2728	0.6090	0.6359	0.953
		$\text{OCP}_{0.90,4}$	0.9742	-0.6919	0.8839	0.9061	0.965
		$\text{OCP}^{(C)}_{0.90,4}$		-0.1471	0.8606	0.8760	0.951
	60	$\text{OCP}_{0.80,4}$	0.9162	-0.3581	0.3485	0.3516	0.958
		$\text{OCP}^{(C)}_{0.80,4}$		-0.0220	0.3350	0.3373	0.950
		$\text{OCP}_{0.90,4}$	0.9742	-0.2610	0.4860	0.5002	0.963
		$\text{OCP}^{(C)}_{0.90,4}$		-0.0546	0.4282	0.4465	0.959
UN	20	$\text{OCP}_{0.80,4}$	0.7620	-0.1722	0.4467	0.4683	0.964
		$\text{OCP}^{(C)}_{0.80,4}$		1.6989	0.5597	0.6017	0.939
		$\text{OCP}_{0.90,4}$	0.8465	-0.4474	0.5333	0.5643	0.963
		$\text{OCP}^{(C)}_{0.90,4}$		3.8343	0.6981	0.7353	0.898
	60	$\text{OCP}_{0.80,4}$	0.7620	-0.1676	0.2177	0.2214	0.951
		$\text{OCP}^{(C)}_{0.80,4}$		2.3916	0.3147	0.3261	0.914
		$\text{OCP}_{0.90,4}$	0.8465	-0.2431	0.2996	0.2999	0.950
		$\text{OCP}^{(C)}_{0.90,4}$		4.0765	0.3728	0.3840	0.807

estimate increases as d_4 increases in any scenario, and this indicates less precision in OCP estimates for a relatively large pre-specified RMSPD compared to the range of data. Under both scenarios, estimated standard errors quickly approach their empirical counterparts as the sample size increases.

Table 2.3 shows the simulation results for RAUOCPC under both scenarios. We

Table 2.3: Simulation results for RAUOCPC based on 1000 data sets under compound symmetry (CS) and unstructured scenarios (UN). "Relative bias" represents sample mean of 1000 values of $100*\{(\widehat{\text{RAUOCPC}}-\text{RAUOCPC})/\text{RAUOCPC}\}$, where $\widehat{\text{RAUOCPC}}$s are obtained through anti-transformations. "Std of estimate" represents standard deviation of 1000 $\widehat{\text{RAUOCPC}}$s. "Mean of SE" estimate represents mean of 1000 bootstrap standard errors. "CP" represents the proportion of 1,000 estimated 95% upper confidence limits computed by (2.20) that are greater than the true value.

Scenario	n	Statistics	True value	Relative bias(%)	Std of estimate	Mean of SE estimate	CP
CS	20	RAUOCPC_4	0.4889	-0.5872	0.1388	0.1318	0.940
		$\text{RAUOCPC}_4^{(C)}$		0.1358	0.1356	0.1326	0.924
		$\text{RAUOCPC}_4^{\text{non-parm}}$		4.0914	0.1332	0.1322	0.848
	60	RAUOCPC_4		-0.1392	0.0789	0.0781	0.948
		$\text{RAUOCPC}_4^{(C)}$		-0.0930	0.0799	0.0777	0.939
		$\text{RAUOCPC}_4^{\text{non-parm}}$		1.4879	0.0792	0.0785	0.892
UN	20	RAUOCPC_4	0.4544	-0.7209	0.2080	0.1978	0.928
		$\text{RAUOCPC}_4^{(C)}$		-5.0134	0.2368	0.2223	0.955
		$\text{RAUOCPC}_4^{\text{non-parm}}$		4.5911	0.2220	0.2199	0.861
	60	RAUOCPC_4		-0.1624	0.1195	0.1167	0.941
		$\text{RAUOCPC}_4^{(C)}$		-5.0183	0.1317	0.1339	0.988
		$\text{RAUOCPC}_4^{\text{non-parm}}$		0.7797	0.1271	0.1301	0.928

first fixed $\delta_{\max} = 1.1$. Under the compound symmetry scenario, both parametric methods provide virtually unbiased estimates for each sample size, as expected for normal data generated under the compound symmetry covariance structure. Empirical coverage rate of the one-sided 95% confidence interval is reasonably close to the nominal value even for a small sample size of 20, indicating fast convergence of the estimates to the normal distribution. On the other hand, the non-parametric method overestimates the true RAUOCPC, but the bias decreases as sample size increases. 95% coverage rate is 90% or less, though reasonable coverage of 92% is generally achieved for a sample size of 100 (not shown). Under the unstructured scenario, it

should be highlighted that the RAUOCPC$^{(C)}$ estimates significantly underestimate the true RAUOCPC. Moreover, their empirical coverage rates indicate that the nominal confidence tends to be overestimated, with values close to 100%. In contrast, the proposed parametric approach results in much smaller values of bias and coverage probabilities that approach 95% as sample size increases. We should also point out that the performance of non-parametric method is relatively poor for a sample size of 20, but it significantly improves in terms of bias and coverage rate as the sample size increases. Standard error estimates are close to their empirical counterparts in any situation.

Our simulation studies show that estimates of the compound symmetry case overall unscaled agreement indices are largely biased and have very poor coverage of the true values when the underlying covariance structure is not strictly compound symmetry. Because it is usually rare to assess agreement among multiple raters that assume homogeneity of all variabilities in measurement processes, we recommend using the proposed approach, defined absent any restriction on the covariance structure, for estimation and inference in most practical situations.

2.4 Renal Study

Renal scans in nuclear medicine (diuresis renography) play an important role in the determination of kidney obstruction which is a condition that may lead to loss of kidney function (Taylor, 2014), Diagnosis of kidney obstruction using renal scans is generally a difficult problem for the following reasons: (1) there is currently no good gold standard for detecting kidney obstruction; and (2) correct interpretation of renal scans requires a deep understanding of renal physiology and technetium-99m-mercaptoacetyltriglycine (99mTc-MAG3) pharmacokinetics.

In the absence of a gold standard, it is generally accepted that the best available

interpretation comes from an expert with broad expertise and extensive experience in academic nuclear medicine. Although interobserver variability still exists between different nuclear medicine experts, it is generally considered to be minimal (Taylor and Garcia, 2014). The vast majority of the 590,000 renal scans performed annually in the United States are interpreted by general radiologists, who have less than 4 months training and experience in interpretations of renal scans and thus have marked variability in their interpretations compared to experienced readers (Taylor et al., 2008c, Taylor and Garcia, 2014). For instance, a survey shows that different practicing radiologist may disagree on the interpretation of the same scan between 9% and 72% of the time (Jaksić et al., 2005).

Given the background, a pilot study was conducted at Emory University with the goal of gaining better insight into the nature of diagnostic variability present in a real-world clinical practice using renal scans. In this respect, the goal of the study is to quantify the interobserver variability in the population of practicing radiologists. The nuclear medicine residents with a minimum of one year of formal training in nuclear medicine were recruited in the pilot study as a surrogate for practicing radiologists. Three nuclear medicine experts who have more than 20 years of experience in nuclear medicine were also recruited. It has been recognized that residents may not perform as well as experts due to their lack of training and limited experience (Taylor et al., 2008c, Taylor and Garcia, 2014, Erdogan et al., 2014).

An intervention called "Computer Assisted Diagnosis" (CAD) was recently introduced by Emory researchers to minimize errors and reduce the interobserver variability among practicing radiologists. The CAD analyzes renal image data and provides a second opinion about the diagnosis with reasoning. Having a second opinion with reasoning is thought to minimize the interobserver variability among radiologists. Thus, it is also of interest to determine if the CAD intervention would reduce the interobserver variability among radiologists.

Thirty five patients with suspected obstruction (20 females, mean age ± SD, 58.7 ± 15.8y) in either right or left kidney were randomly selected. Their scans from 70 kidneys (35 left kidneys and 35 right kidneys) were independently interpreted by the three groups of raters: a) three nuclear medicine experts each with 20+ years of experience ("Experts"); b) three nuclear medicine residents each having completed a minimum one of their three year nuclear medicine residency ("Residents"); c) same three residents with subsequent access to CAD ("Residents + CAD"). Raters scored each kidney of a scale of -1 to +1.

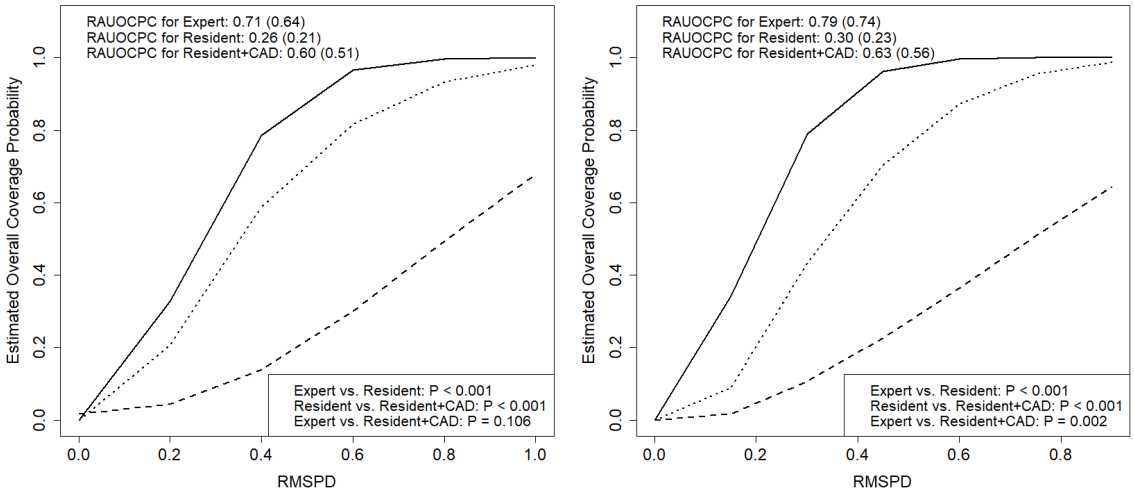
We first performed a series of preliminary data analyses to investigate the structure of data and check relevant statistical assumptions. The multivariate normality assumption of pairwise differences within each group of raters was assessed based on chi-square Q-Q plots (not shown) and Doornik-Hansen's test of multivariate normality Doornik and Hansen (2008), and no significant departure from the assumption was found for all five sets of pairwise differences (p-value range: 0.07 to 0.51) except within experts in left kidneys (p-value = 0.01). Heterogeneous variabilities appeared to exist, especially among the three residents, whose sample variances ranged from 0.2 to 0.5 in both kidneys. Presence of such heterogeneity was further confirmed by the Morgan-Pitman test for comparing variances between correlated samples (all p-values < 0.05).

Table 2.4 presents the estimates of $\text{ODI}_{0.9,3}$ and $\text{OCP}_{0.5,3}$ for the three groups of raters, and results of their pairwise across-group comparisons based on two-sample hypothesis tests. The pre-specified RMSPD of $d_3 = 0.5$ was determined based on a clinically important squared pairwise difference between any paired ratings that may disagree on the obstruction status, with the consultation of a nuclear medicine expert, Dr. Taylor, from Emory University. Statistical inference is based on 1000 bootstrap samples. Note that all ODI estimates were rounded to one decimal place to reflect the unit of measurement, while all OCP estimates were rounded to two decimal places.

The three experts showed high agreement for both kidneys as expected. Specifically, based on the ODI estimates, 90% of ratings from the three experts had RMSPDs less than 0.5 (95% upper limit = 0.6) for the left kidney and less than 0.4 (95% upper limit = 0.5) for the right kidney. Or equivalently, based on the OCP estimates, about 91% (95% lower limit = 72%) of ratings for the left kidney and almost 98% (95% lower limit = 92%) of ratings for the right kidney had RMSPDs less than 0.5. Agreement among the three residents is significantly worse than that of the three experts for both left and right kidneys as evidenced by significantly higher ODI estimates and significantly lower OCP estimates (all p-values < 0.001). However, agreement among residents dramatically improves after given access to the CAD. For this group ("Residents + CAD"), approximately 90% of ratings had RMSPDs less than 0.7 and 0.6 (95% upper limits = 1.0 and 0.8) in left and right kidneys, respectively. Or equivalently, at least 72% (95% lower limit = 53%) of ratings for the left kidney and 77% (95% lower limit = 61%) of ratings for the right kidney had RMSPDs less than 0.5. Results of pairwise across-group comparisons based on two sample hypotheses tests show that agreement among residents with CAD was significantly better than that among the same residents before access to CAD (all p-values < 0.001). Moreover, the interobserver agreement among the residents with CAD resembles that among the experts, as we find that the differences between two groups are not significant for the left kidney (p-values = 0.107 and 0.191).

Figure 2.1 shows three estimated overall coverage probability curves with parametric RAUOCPC estimates (95% lower limits) and results of their pairwise across-group comparisons comparisons for left and right kidneys, respectively. Note that *a priori* maximum acceptable difference was set to 1. We can first visually verify that the experts have the highest agreement as evidenced by the largest area under the curve, closely followed by the residents with CAD. The area under the curve is smallest for the residents without access to CAD, indicating that their agreement is worst among

Figure 2.1: Overall coverage probability curves based on left (Left) and right (Right) kidneys from renal study data. The solid lines indicate the estimated overall coverage probability curves for experts; the dotted lines indicate the estimated overall coverage probability curves for residents + CAD; and the dashed lines are indicate estimated overall coverage probability curves for residents.



the three groups. The three experts have highest estimated RAUOCPC as 0.71 (95% lower limit = 0.64) for the left kidney and as 0.79 (95% lower limit = 0.74) for the right kidney. The residents have significantly lower estimates as compared to those of experts for both left and right kidneys (all p-values < 0.001), though the agreement significantly improves with CAD (all p-values < 0.001). Moreover, agreement among residents with CAD closely matches that of the experts for left kidney (p-value = 0.106). Our preliminary results suggest that although there exists high interobserver variability in the interpretations of renal scans among practicing radiologists, use of CAD significantly reduces their interobserver variability and results in the degree of variability being close to that among the experts.

2.5 Discussion

We have proposed several overall agreement indices (ODI, OCP and RAUOCPC), which are practically useful for assessing agreement among multiple raters. The key is to quantify disagreement among measurements using a new comprehensive measure of distance that represents the root mean square of pairwise differences (RMSPD) among measurements provided multiple raters. The proposed overall unscaled indices defined without any assumption regarding the covariance structure among measurements provide great flexibility in a sense that practitioners can quantify agreement even among multiple raters with highly heterogeneous variabilities in respective measurement processes. Overall unscaled indices are tied to the original measurement scale and can be compared against pre-specified acceptable/tolerable RMSPDs to determine satisfying agreement. Thus, they are easily explained to non-statistical practitioners and can serve as a useful alternative or complement to existing scaled indices in various clinical study settings.

Definitions and interpretations of unscaled agreement indices depend on the choice of an extended measure of distance that quantifies the degree of disagreement among multiple raters. In this manuscript, we have proposed the use of RMSPD (D_k) as defined in (2.1). A possible alternative to this measure is the average of pairwise absolute differences among multiple raters, that is, $D_k^* = \frac{2}{k(k-1)} \sum_{1 \le p < q \le k} |Y_p - Y_q|$. D_k and D_k^* quantify inherently different aspects of the spread of data and each has its own merits. One advantage of D_k^* is its robustness (less sensitive to the outliers) compared to D_k, which depends on moments of the distribution. The proposed measure D_k, on the other hand, has better mathematical and asymptotic properties in connection with various widely-used statistical models and probability distributions.

For practitioners, we note that the square root of acceptable/tolerable squared difference that one is willing to impose between a typical pair of measurements can serve as an interpretable reference level based on RMSPD. In practice, reference

standard data can serve as a guidance to set up an acceptable RMSPD value. For instance, expert rating data on selected patients with suspected obstruction have been available over the years at Emory University Hospital, and RMSPD value based on such data may guide the choice of the acceptable value.

We used Liu et al.'s approach (Liu et al., 2009) to approximate the distribution of D_k^2 (squared RMSPD) under two rationales: 1) we can rely on nearly all statistical packages to easily obtain its quantile function (inverse of a cumulative non-central chi-square distribution function), which is used to define a series of overall unscaled agreement indices; 2) the reasonably high accuracy of approximation was demonstrated in simulation studies. We may consider using exact distributional form (2.6) or other approximation approaches. However, these alternatives either have a complicated form from which the quantile function is not readily obtainable or yield poor approximation results as compared to Liu et al.'s approach (Liu et al., 2009). For instance, since exact distribution form (2.6) has as an infinite series representation (Kotz et al., 1967), we can compute F by truncating the series after N terms, but obtaining F^{-1} is a burdensome task because it involves inverting a infinite series. Another approach uses numerical inversion of the characteristic function of D_k^2 to approximate the cumulative distribution function (Imhof, 1961). However, obtaining F^{-1} is extremely difficult as it involves inverting an analytically intractable integral. There exist other moment-based approximation approaches, such as the extension of Pearson's three-moment central χ^2 method (Imhof, 1961, Pearson, 1959). This approach essentially provides a central χ^2 approximation to non-negative D_k^2 by matching the third-order moments. Its approximate distribution form is simple, as it only requires an inversion of a central χ^2 cumulative distribution function. However, the current approach produces better approximation results for the tail probabilities since it further requires a best match of the fourth-order moments (Liu et al., 2009).

Our proposed approach assumes distinct pairwise differences to follow a multi-

variate normal distribution in defining overall unscaled agreement indices, which is a weaker assumption then imposing normality on measurements themselves. Such assumption is a natural conceptual extension to the traditional TDI case in which the difference between paired measurements is assumed to follow a univariate normal distribution. In the event that there are potential violations to the normality assumption, appropriate transformation can be applied to the measurements to ensure that the distributional assumption likely holds. It is important to note that interpretations should then be in terms of the transformed scale. Recently, several novel non-parametric methods for estimation and inference of the TDI were introduced (Choudhary, 2010, Perez-Jaume and Carrasco, 2015, Lin et al., 2016). It is of future interest to develop similar non-parametric approach for our proposed indices to deal with non-normal data.

Table 2.4: Estimated ODIs and OCPs from Renal Study Data. 95% upper and lower confidence limits are used for ODI and OCP estimates, respectively. P-values denote results from two-sample hypothesis tests.

Statistics	Kidney	Experts (95% Limit)	Residents (95% Limit)	Resid.+CAD (95% Limit)	Experts vs. Residents	Residents vs. Resid.+CAD	Experts vs. Resid.+CAD
$ODI_{0.9,3}$	L	0.5 (0.6)	1.4 (1.6)	0.7 (1.0)	< 0.001	< 0.001	0.107
	R	0.4 (0.5)	1.3 (1.5)	0.6 (0.8)	< 0.001	< 0.001	0.002
$OCP_{0.5,3}$	L	0.91 (0.72)	0.21 (0.15)	0.72 (0.53)	< 0.001	< 0.001	0.191
	R	0.98 (0.92)	0.27 (0.18)	0.77 (0.61)	< 0.001	< 0.001	0.014

Chapter 3

Assessing Alignment Between Functional Markers and Ordinal Outcomes Based on Broad Sense Agreement

Portions of this chapter were previously published as Jang JH, Peng L, Manatunga AK. Assessing alignment between functional markers and ordinal outcomes based on broad sense agreement. *Biometrics*. 2019;1-13. https://doi.org/10.1111/biom.13063, and have been reproduced with permission.

3.1 Introduction

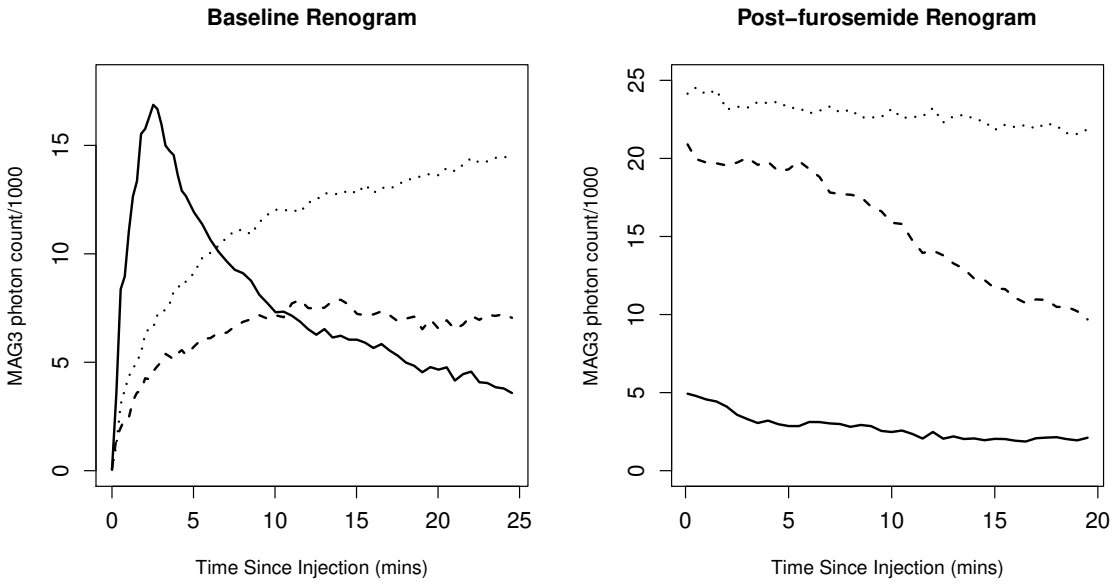
Statistical methods for characterizing alignment between paired measurements in the same scale are well-established in the agreement literature. For instance, with paired categorical or ordinal data, the kappa coefficient or the weighted kappa coefficient (Cohen, 1960, 1968, Fleiss, 1971, Kraemer, 1980); with paired continuous measurements, intraclass correlation coefficient (ICC) (Bartko, 1966) and concordance correlation coefficient (CCC) (Lin, 1989) are popular measures of agreement. Some raters, however, may use different measurement processes with distinctive point systems, resulting in paired measurements with different scales (e.g., continuous and ordinal). The aforementioned methods cannot be applied in such cases, because they require measurements to be on the same scale. Recently, Peng et al. (2011) proposed a broad sense agreement (BSA) framework that is specifically designed to characterize alignment between continuous and ordinal measurements. The BSA measure proposed by Peng et al. (2011) is scaled between -1 and 1, with its value equaling 1 (or -1) representing a perfect broad sense agreement (or disagreement). The high value of the BSA measure closer to 1, the higher the capability of interpreting the continuous scale according to the ordered categories of interest.

With the advancement in data collection technology, more and more observations are being collected as functional markers, each of which consists of repeated measurements that are densely sampled over a time or other continua (Ramsay and Silverman,

2005). In literature on traditional agreement, a few popular indices have been generalized in the presence of functional markers. Li and Chow (2005) proposed an extended CCC that can evaluate agreement between paired functional markers. Following the formulation of a traditional CCC that involves paired univariate measurements (Lin, 1989), the authors characterized the degree of agreement between the two functional markers by their expected squared distance, which is defined based on the functional inner product. More recently, Rathnayake and Choudhary (2016) proposed a concept of tolerance bands for functional markers, as an extension of univariate tolerance intervals that have been used to evaluate agreement in clinical measurement methods (Choudhary, 2008). Specifically, the authors proposed simultaneous bands that always contain a certain proportion of *entire* individual curves with pre-specified confidence. These methods, however, are limited to comparing between functional measurements. To our knowledge, no systematic research addresses the question of how to assess alignment between a functional marker and an ordinal outcome. The recent work by Peng et al. (2011) provides a promising tool to address such a question, which is the focus of this paper.

Our work is motivated by data collected in a renal study. Obstruction to urine drainage from kidney (kidney obstruction) is a serious clinical problem that can lead to irreversible loss of renal function if not properly treated. In the diagnosis of kidney obstruction, ^{99m}Tc-Mercaptoacetyltriglycine (MAG3) is injected to a patient and photon counts in each kidney are measured during the renal scan period, producing a set of renogram curves (Taylor et al., 2008c). The first renogram curves (called baseline) represent the MAG3 photon counts detected during the initial period of 24 minutes (see the left panel in Figure 3.1). Second renogram curves (called post-furosemide) are also obtained with an additional 20 minutes after an intravenous injection of furosemide, a potent diuretic (see the right panel in Figure 3.1). In the absence of a gold standard for the presence of kidney obstruction, consensus ordinal ratings on

Figure 3.1: Representative Renogram curves for three kidneys. The solid lines are from a kidney rated as "non-obstructed" by expert consensus; the dashed lines are from a kidney rated as "equivocal"; and the dotted lines are from a kidney rated as "obstructed".



each subject's obstruction status (1: non-obstructed; 2: equivocal; 3: obstructed) were further collected from a group of nuclear medicine experts as the best available standard.

A good alignment between the renogram curves and the consensus ordinal ratings would suggest an improved diagnostic utility of renogram curves in detecting suspected kidney obstruction. One ad-hoc approach is to compute the BSA measures between the observed photon counts on the curves and the ordinal ratings at each discrete time point. However, the interpretation of a set of pointwise BSA estimates that fluctuate in time is not always straightforward. For instance, crossings of the baseline renogram curves in Figure 3.1 will imply their varying degrees of alignment with the ordinal ratings over time. The resulting pointwise BSA estimates at different time points may only provide an inconclusive picture of the overall diagnostic utility of the renogram curves themselves.

Often in clinical practice, several quantitative features of functional curves (e.g., pharmacokinetic area under the curve, AUC) are used for the interpretation of markers or diseases. This is indeed the case with renogram curves which can be characterized by several important features that are inherent in its functional nature and are also related to the severity of the kidney obstruction. Common examples include maximum MAG3 photon count, time to reach maximum MAG3 photon count, etc., which are frequently derived from the renogram curves to help physicians evaluate possible kidney obstruction (Bao et al., 2011). From this perspective, a more substantive interest in studying the relationship between functional and ordinal scales is to identify an important quantitative feature of the functional marker that aligns well with the corresponding ordinal rating. Thus, our goal is to develop a framework based on BSA that can assess and compare alignment of various quantitative features of functional markers according to their ordinal outcomes, and ultimately help identify quantitative features that have good diagnostic utility.

In this manuscript, our strategy is to adopt a general class of *summary functionals*, each of which flexibly captures a different quantitative feature of a functional marker, such as AUC, the evaluation of a function or its derivatives at certain points or the point that reaches a maximum/minimum of a functional marker. This approach allows studying alignment between a large class of important quantitative features of a functional marker and an ordinal outcome. Following this idea, we provide an inferential framework for comparing a pair of candidate summary functionals in terms of their importance on the ordinal outcome. It is worth noting that there are some complications in the estimation and inference of the proposed framework. That is, each functional marker is not directly observable continuously in time; rather, each observation is collected at discrete time points with some possible measurement error. In such a situation, extra work in constructing the functional estimate is warranted to ensure desirable asymptotic properties of the corresponding BSA estimator.

The remainder of the Chapter is organized as follows. In Section 3.2, we first review the existing BSA framework, followed by an introduction of a general class of summary functionals and our proposed framework based on BSA for measuring alignment of functional markers according to their ordinal outcomes. Nonparametric estimators and their asymptotic properties, and subsequent inferential procedures including variance estimation and construction of confidence intervals are also presented. In Section 3.3, we illustrate the proposed framework based on BSA using several concrete classes of summary functionals. In Section 3.4, we describe the inferential framework for comparing summary functionals regarding their alignment with the ordinal outcomes. In Section 3.5, we report the results of a simulation study conducted to evaluate the performance of the proposed approaches. The application of our methods to a renal study is illustrated in Section 3.6. Finally, we conclude with some remarks in Section 3.7.

3.2 Methods

3.2.1 Review of broad sense agreement

The concept of broad sense agreement (BSA) was introduced by Peng et al. (2011) as to characterize the alignment of continuous measurements X according to their established ordered categories Y. Let $\mathcal{D}_X$ and $\mathcal{D}_Y$ be the domain of X and Y, respectively. Perfect broad sense agreement (or disagreement) between X and Y is defined as the existence of an increasing (or decreasing) step function Ψ from $\mathcal{D}_X$ and $\mathcal{D}_Y$ such that $Y = \Psi(X)$ with probability of 1. That is, if $X_{(*k)}$ denotes the randomly selected X given $Y = k$ $(k = 1, 2, \ldots, K)$, a perfect broad sense agreement (disagreement) case implies $X_{(*1)} < X_{(*2)} < \cdots < X_{(*K)}$ $(X_{(*1)} > X_{(*2)} > \cdots > X_{(*K)})$ with probability 1.

An index for measuring the degree of BSA was proposed (Peng et al., 2011). The index quantifies the discrepancy between the observed and expected ranks under per-

fect BSA among a group of continuous measurements. Specifically, denote the ranks of $\{X_{(*1)}, X_{(*2)}, \ldots, X_{(*K)}\}$ by $\{R_1, R_2, \ldots, R_K\}$. Then the proposed BSA measure is defined as

$$\rho_{\text{bsa}}(X,Y) = 1 - \frac{E\Big\{\sum_{k=1}^{K}(k-R_k)^2\Big\}}{E\Big\{\sum_{k=1}^{K}(k-R_k)^2 \mid X \perp Y\Big\}}, \tag{3.1}$$

where $E(\cdot)$ denotes the expectation and $E(\cdot \mid X \perp Y)$ denotes the expectation given that X and Y are independent. $\rho_{\text{bsa}}(X,Y)$ always takes a value between -1 and 1, with 1 (or -1) representing perfect broad sense agreement (disagreement). A value close to 0 indicates independence between X and Y.

A non-parametric estimator of the BSA measure $\rho_{\text{bsa}}(X,Y)$ was proposed (Peng et al., 2011). The basic idea is to adopt the stratified resampling idea and examine all possible groups of K observations of (X,Y) with distinct Y values. Define Θ_K as the sample space of $\{R_1, \cdots, R_K\}$ which consists of $K!$ permutations of the elements of $\overrightarrow{\mathbf{k}} = [1, \ldots, K]'$. Suppose the observed data consist of n pairs of (X_i, Y_i), $i = 1, \ldots, n$. Let $n_k = \sum_{i=1}^{n} I(Y_i = k)$ and $\sum_{k=1}^{K} n_k = n$, and denote $X_{(*k),s_k}$ as the s_k^{th} $(1 \le s_k \le n_k)$ continuous measurement among those that fall into the k^{th} ordinal category. Denote $\overrightarrow{\mathbf{R}}(\cdot)$ as a mapping from $\mathcal{D}_X^K$ to Θ_K: $\overrightarrow{\mathbf{R}}(\overrightarrow{\mathbf{x}}) = [r_1, \ldots, r_K]'$, where $\overrightarrow{\mathbf{x}} = [x_1, \ldots, x_K]' \in \mathcal{D}_X^K$ and $r_k = \sum_{m=1}^{K} I(x_k \ge x_m)$ representing the rank of x_k among $\{x_1, \ldots, x_K\}$. Then it can be shown that the estimator takes the form of (Peng et al., 2011)

$$\widehat{\rho}_{\text{bsa}}(X,Y) = 1 - \frac{\Big(\prod_{k=1}^{K} n_k\Big)^{-1} \sum_{s_1=1}^{n_1} \cdots \sum_{s_K=1}^{n_K} \Big|\Big|\overrightarrow{\mathbf{k}} - \overrightarrow{\mathbf{R}}(X_{(*1),s_1}, \ldots, X_{(*K),s_K})\Big|\Big|^2}{(K^3-K)/6}, \tag{3.2}$$

where $||\cdot||$ is a Euclidean norm in $\mathbb{R}^K$. Asymptotic properties of the estimator have also been established (Peng et al., 2011).

3.2.2 General formulation of the summary functional

Without loss of generality, we take the domain of the function marker X as a time interval $\mathcal{T}$ and accordingly denote $X(t)$ as a continuous measurement at time $t \in \mathcal{T}$. For a nonnegative integer ω, define $\mathcal{F}_\omega = \{f : \mathcal{T} \to \mathbb{R};\ f$ is square integrable and ω-times continuously differentiable at any $t \in \mathcal{T}\}$, and assume $X \in \mathcal{F}_\omega$. A general formulation of the *summary functional* of X is simply defined as a map from $\mathcal{F}_\omega$ to $\mathbb{R}$, that is, $\phi : \mathcal{F}_\omega \to \mathbb{R}$. Our approach is to consider this general formulation of the summary functional that encompasses a wide class of quantitative features of a functional marker; we defer discussion of its specific examples until Section 3.3.

3.2.3 Proposed BSA framework

First, we aim to investigate how a chosen quantitative feature of functional markers is informative about their corresponding ordinal outcomes. To achieve this, we propose an extended BSA framework, under which the degree of alignment between the chosen quantitative feature of a functional marker and the ordinal outcome is characterized by

$$\rho_{\text{bsa}}(\phi(X), Y),$$

where $\phi(X)$ is a summary functional. We see that the index is now essentially based on the comparison between the ranks of $\{\phi(X_{(*1)}), \phi(X_{(*2)}), \cdots, \phi(X_{(*K)})\}$ and their anticipated ranks under the perfect BSA scenario $(1, 2, \cdots, K)$ (see (3.1)). Thus, the closer $\rho_{\text{bsa}}(\phi(X), Y)$ is to 1, the better the alignment between the chosen quantitative feature of a functional marker captured by the summary functional $\phi(X)$ and the ordinal outcome Y.

3.2.4 Nonparametric estimation

Suppose that n functional markers $X_1, X_2, \ldots X_n$ are directly observable. Then, given a reasonable choice of a quantitative feature to be analyzed, this in turn implies that n summary functionals $\phi(X_1), \phi(X_2), \ldots, \phi(X_n)$ can be obtained for each subject. In such a case, it is straightforward to assess their alignment with the given ordinal outcomes Y by estimating the BSA measure using (3.2) with X replaced by $\phi(X)$.

In reality, however, each functional marker X_i $(i = 1, \ldots, n)$ is not observed continuously in time; instead, a set of continuous measurements $X_i(t_{ij})$ $(j = 1, \ldots, N_i)$ are observed with possible measurement error as $W_i(t_{ij})$ at N_i discrete time points $\{(t_{i1}, t_{i2}, \ldots, t_{iN_i}) \in \mathcal{T} : t_{i1} < t_{i2} < \cdots < t_{iN_i}\}$. We can express this using the following model

$$W_i(t_{ij}) = X_i(t_{ij}) + \epsilon_i(t_{ij}), \tag{3.3}$$

where the random measurement error $\epsilon_i(t)$ follows an independent and identical distribution with $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_e^2 < \infty$ for each t, mutually independent of the true function X_i. We assume that $N_i \to \infty$ as $n \to \infty$ for all i, that is, $N_i = N_{i,n}$ is a sequence that tends to infinity. For ease of presentation, we omit the subscript n.

Accordingly, the true value of each summary functional $\phi(X_i)$ is unknown but can be estimated based on the observed data as $\phi_{N_i}(W_i)$. For instance, a summary functional that captures the AUC of a functional marker, that is, $\phi(X_i) = \int X_i(t)dt$, cannot be directly computed; instead, it may be constructed as a Riemann sum of the observed data points as $\phi_{N_i}(W_i) = \sum_{j=1}^{N_i - 1} W_i(t_{ij})(t_{i,j+1} - t_{ij})$. The observed data thus consist of n independently distributed pairs of estimated summary functionals of interest and their respective ordinal outcomes $\{\phi_{N_1}(W_1), Y_1\}, \{\phi_{N_2}(W_2), Y_2\}$, $\ldots, \{\phi_{N_n}(W_n), Y_n\}$.

We propose to estimate $\rho_{\mathrm{bsa}}(\phi(X), Y)$ by $\widehat{\rho}_{\mathrm{bsa}}(\phi_N(W), Y)$, which, under formu-

lation (3.2), is based on the stratified resampling scheme that examines all possible groups of K observations $\{\phi_N(W), Y\}$ with distinct Y values. Specifically, the non-parametric estimator takes the form of

$$\begin{aligned}&\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)\\ &= 1 - \frac{\Big(\prod_{k=1}^{K} n_k\Big)^{-1} \sum_{s_1=1}^{n_1} \cdots \sum_{s_K=1}^{n_K} \big|\big| \overrightarrow{\mathbf{k}} - \overrightarrow{\mathbf{R}}\{\phi_N(W_{(*1),s_1}), \ldots, \phi_N(W_{(*K),s_K})\}\big|\big|^2}{(K^3 - K)/6},\end{aligned} \tag{3.4}$$

where $\overrightarrow{\mathbf{k}} = [1, \ldots, K]'$, $\overrightarrow{\mathbf{R}}(\cdot)$ is mapping from $\mathcal{D}_\phi^K$ to Θ_K: $\overrightarrow{\mathbf{R}}(\overrightarrow{\mathbf{w}}) = [r_1, \ldots, r_K]'$, $\overrightarrow{\mathbf{w}} = [\phi_N(w_{(*1)}), \ldots, \phi_N(w_{(*K)})]' \in \mathcal{D}_\phi^K$ and $r_k = \sum_{m=1}^{K} I\{\phi_N(w_{(*k)}) \geq \phi_N(w_{(*m)})\}$ representing the rank of $\phi_N(w_{(*k)})$ among $\{\phi_N(w_{(*1)}), \ldots, \phi_N(w_{(*K)})\}$. Note that all the subscripts of N have been dropped in the right-hand side of the equation (3.4) for ease of representation; that is, $\phi_{N_{(*k),s_k}}(W_{(*k),s_k}) \equiv \phi_N(W_{(*k),s_k})$ for all k.

There are two potential sources of sampling error in the estimation of $\rho_{\text{bsa}}(\phi(X), Y)$ using $\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ given by (3.4). The first source of error stems from the basic formulation of the BSA statistic where a stratified resampling scheme is adopted to estimate the similarity between the observed ranks of summary functionals and their anticipated ranks under the scenario of perfect BSA. The second source of error comes from replacing the true summary functionals $\phi(X)$ with their estimates $\phi_N(W)$ and is thus specific to our situation involving functional markers. It is important that both sources of error are taken into consideration when studying the (asymptotic) properties of the proposed estimator (3.4).

3.2.5 Asymptotic properties

In Theorem 3.2.1, we establish the consistency and asymptotic normality of the proposed estimator $\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ given by (3.4), provided that the consistency of the estimator $\phi_{N_i}(W_i)$ for $\phi(X_i)$ holds for all i. Its proof is provided in Appendix B.1.

Theorem 3.2.1. *Suppose* $\sup_{1\le i\le n}|\phi_{N_i}(W_i)-\phi(X_i)| \le \mathcal{O}_p(P_N)$, *where* P_N *is a nonnegative sequence. (i) If* $P_N \to 0$ *as* $n\to\infty$ *and the regularity conditions A1-A3 provided in Appendix B.1 hold,* $\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)$ *is a consistent estimator for* $\rho_{\text{bsa}}(\phi(X),Y)$. *(ii) If* $\sqrt{n}P_N \to 0$ *as* $n\to\infty$ *and the regularity conditions A1, A2 and A4 provided in Appendix B.1 hold, when* $|\rho_{\text{bsa}}(\phi(X),Y)|<1$, $\sqrt{n}\{\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)-\rho_{\text{bsa}}(\phi(X),Y)\}$ *has an asymptotic normal distribution with mean zero and variance* σ^2_{bsa}, *where* σ^2_{bsa} *is defined in Appendix B.1.*

The key idea of the proof is to consider the decomposition $\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)-\rho_{\text{bsa}}(\phi(X),Y)=T_1+T_2$, where $T_1=\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)-\rho_{\text{bsa}}(\phi_N(W),Y)$ and $T_2=\rho_{\text{bsa}}(\phi_N(W),Y)-\rho_{\text{bsa}}(\phi(X),Y)$. The consistency and asymptotic normality of the first term T_1 can be readily established as for the univariate case (Peng et al., 2011) given any fixed N as $n\to\infty$. T_2 can be shown negligible provided that $\sup_{1\le i\le n}|\phi_{N_i}(W_i)-\phi(X_i)|\le\mathcal{O}_p(P_N)$, where P_N and $\sqrt{n}P_N$ approach 0 as n goes to infinity.

3.2.6 Estimation of standard error and confidence interval

We propose to estimate asymptotic variance of $\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)$ using the jackknife method, given the rather complicated analytic form of σ^2_{bsa}. The consistency of the jackknife estimator is guaranteed by the fact that $\sqrt{n}\{\widehat{\rho}_{\text{bsa}}(\phi_N(W),Y)-\rho_{\text{bsa}}(\phi(X),Y)\}$ is, asymptotically, a U-statistic (Arvesen, 1969). Specifically, let $\hat{\rho}^{(-i)}_{\text{bsa}}(\phi_N(W),Y)$ be the BSA estimate based on the sample with the i^{th} pair $\{\phi_{N_i}(W_i),Y_i\}$ removed. The jackknife variance estimator is then given by

$$\hat{s}^2_J=\frac{n-1}{n}\sum_{i=1}^{n}\left\{\hat{\rho}^{(-i)}_{\text{bsa}}(\phi_N(W),Y)-\frac{1}{n}\sum_{p=1}^{n}\hat{\rho}^{(-p)}_{\text{bsa}}(\phi_N(W),Y)\right\}^2. \tag{3.5}$$

Note that the validity of the jackknife formula (3.5) is due to the fact that both $\text{Var}(T_2)$ and $\text{Cov}(T_1,T_2)$ (T_1 and T_2 are defined in Section 3.2.5 under Theorem 3.2.1)

are asymptotically negligible given a consistent estimator of $\phi(X)$. Furthermore, other non-parametric methods such as the bootstrap, half-sampling or subsampling can be used for estimating the asymptotic variance; see Efron (1981) for details of other applicable methods.

One may use normal approximation to construct confidence intervals (CIs) of $\rho_{\text{bsa}}(\phi(X), Y)$. Since $\rho_{\text{bsa}}(\phi(X), Y) \in [-1, 1]$, adopting Fisher's Z-transformation may accelerate the convergence of $\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ to asymptotic normality, especially when $\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ is close to the boundary. Specifically, let $g(a) = 0.5 \times \ln\{(1 + a)/(1 - a)\}$, $g'(a) = dg(a)/da$, and $g^{-1}(\cdot)$ denote the inverse function of $g(\cdot)$. Using the delta method, the $100(1-\alpha)\%$ CI for $\rho_{\text{bsa}}(\phi(X), Y)$ can be constructed as

$$\Big[g^{-1}\big(\widetilde{g} - z_{1-\alpha/2} \cdot \tilde{g}'\widehat{s}_J\big), g^{-1}\big(\widetilde{g} + z_{1-\alpha/2} \cdot \tilde{g}'\widehat{s}_J\big)\Big], \tag{3.6}$$

where $\widetilde{g} \equiv g\{\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)\}$, $\tilde{g}' \equiv g'\{\widehat{\rho}_{\text{bsa}}(\phi_N(W), Y)\}$ and $z_{1-\alpha/2}$ denotes the $100(1-\alpha/2)^{\text{th}}$ percentile of $N(0, 1)$.

3.3 Illustration of the Proposed BSA Framework

In this section, we illustrate the proposed framework based on BSA using three special classes of summary functionals that are relevant and of importance in various clinical settings.

3.3.1 Three special cases of summary functionals

Suppose that the functional marker X is ω-times continuously differentiable (see Section 3.2.2). We denote $X^{(\nu)}$ as its ν^{th} derivative ($0 \leq \nu \leq \omega - 2$), with $X^{(0)} = X$.

AUC-type funtionals. AUC-type functionals are often used in practice to summarize a functional marker. Specifically, they take the form

$$\phi(X) \coloneqq \phi^{[\nu]}_{\mathrm{AUC}}(X) = \int_{\mathcal{T}} X^{(\nu)}(t)dt.$$

Setting $\nu = 0$ and $\nu = 1$ above gives the area under a crude curve (crude AUC) and the area under the first derivative of a curve (first derivative AUC), respectively.

Magnitude-specific functionals. Another important quantitative feature of a functional marker or its (higher-order) derivative is its magnitude associated with a specific argument value t. Accordingly, given $t^* \in \mathcal{T}$, a magnitude-specific functional can be expressed as

$$\phi(X) \coloneqq \phi^{[\nu]}_{\mathrm{MAG}(t^*)}(X) = X^{(\nu)}(t^*).$$

A unique maximum or minimum magnitude of a functional marker sometimes provides useful information and can be expressed as a summary functional as defined below:

$$\phi(X) \coloneqq \phi^{[\nu]}_{\mathrm{MAX}}(X) = \sup_{t\in\mathcal{T}} X^{(\nu)}(t) \quad \text{or} \quad \phi(X) \coloneqq \phi^{[\nu]}_{\mathrm{MIN}}(X) = \inf_{t\in\mathcal{T}} X^{(\nu)}(t).$$

Time-specific functionals. Time to attain a certain threshold value η of a functional marker or its (higher-order) derivative is often of great interest for researchers. Such a quantitative feature can be readily captured by a time-specific functional that maps the space of functional markers to the relevant time domain, i.e., $\phi : \mathcal{F}_\omega \to \mathcal{T} \subset \mathbb{R}$:

$$\phi(X) \coloneqq \phi^{[\nu]}_{\mathrm{TIME}(\eta)}(X) = \inf\{t \in \mathcal{T} : X^{(\nu)}(t) = \eta\}.$$

In many practical situations, researchers are interested in investigating the timing

of a unique maximum of a curve. This quantitative feature can be appropriately captured using a time-specific functional of the form

$$\phi(X) \coloneqq \phi^{[\nu]}_{\text{tMAX}}(X) = \arg\sup_{t\in\mathcal{T}} X^{(\nu)}(t).$$

An analogous form holds for the time at which a unique minimum value is achieved.

The interpretation of a perfect BSA scenario (i.e. $\rho_{\text{bsa}}(\phi(X), Y) = 1$) differs depending on a chosen summary functional. For instance, if $\phi_{\text{AUC}}(X)$ is adopted, a perfect BSA scenario implies $\phi_{AUC}(X_{(*1)}) < \cdots < \phi_{AUC}(X_{(*K)})$. In other words, with probability 1, functional markers that are indexed with higher ordinal values have greater crude AUC than those of other functional markers indexed with lower ordinal values. Analogous interpretations hold with respect to the magnitude-specific and time-specific summary functionals.

3.3.2 Nonparametric estimation of the special-case summary functionals

Assume without loss of generality that $\mathcal{T} = [0, 1]$. As illustrated in model (3.3), we do not observe the true functional marker X_i $(i = 1, \ldots, n)$ in practice, but collect its realized values at N_i discrete times points $0 = t_{i1} < t_{i2} < \cdots < t_{iN_i} = 1$ with measurement error as W_i.

Several smoothing techniques are available to estimate the true functional marker X_i based on noisy observations W_i (e.g., kernel smoothing, spline, moving average and so on). For instance, a popular approach is to use smoothing splines (e.g., cubic B-splines) to approximate the true function. The coefficients for the spline basis functions are estimated as the solution to the penalized least squares problem that aims to explicitly control the trade-off between fidelity to the data and roughness of the function estimate. An excellent reference on smoothing splines is Green and

Silverman (1994).

In our work, we opt to a non-parametric smoothed estimate of the true underlying function $X_i^{(\nu)}$ using the following kernel estimator (Gasser and Müller, 1979, 1984, Müller, 1984, 1985):

$$\widehat{W}_i^{[\nu]}(t) = \frac{1}{b_{N_i}^{\nu+1}} \sum_{j=1}^{N_i} \int_{d_{i,j-1}}^{d_{ij}} K_\nu\Big(\frac{t-u}{b_{N_i}}\Big) du \cdot W_i(t_{ij}), \tag{3.7}$$

where $d_{i,j-1} = (t_{ij} + t_{i,j-1})/2$ ($d_{i0} = 0, d_{N_i} = 1$) and b_{N_i} is a smoothing parameter (bandwidth) satisfying $b_{N_i} \to 0$, $N_i b_{N_i} \to \infty$ as $N_i \to \infty$ for all i. K_ν is a kernel function of order (ν, ω) defined on a compact support $[-1, 1]$ and takes on zero values on the boundary points (see Appendix B.2). This so-called Gasser-Müller kernel estimator is widely recognized for its computational efficiency and good asymptotic properties (Gasser and Müller, 1984, Müller, 1984). Furthermore, it provides relatively accurate first or higher-order derivative estimates even with a number of observed time points as small as 15 (Gasser et al., 1991). Thus, we propose to build a non-parametric estimator $\phi_{N_i}(W_i)$ for each of the three special-case summary functionals $\phi(X_i)$ based on the Gasser-Müller kernel estimator (3.7) as following.

AUC-type funtionals. This type of summary functional can be estimated by the following Riemann sum of $\widehat{W}_i^{[\nu]}$ with respect to the output design points $\{t_{i1}, \ldots, t_{i,N_i}\}$:

$$\phi_{N_i}(W_i) \coloneqq \phi_{N_i,\mathrm{AUC}}^{[\nu]}(W_i) = \sum_{j=1}^{N_i-1} \widehat{W}_i^{[\nu]}(t_{ij})(t_{i,j+1} - t_{ij}).$$

Magnitude-specific funtionals. Given a specific time point $t^* \in \mathcal{T}$, a general magnitude-specific functional can be directly estimated by its empirical counterpart of $\widehat{W}_i^{[\nu]}$:

$$\phi_{N_i}(W_i) \coloneqq \phi_{N_i,\mathrm{MAG}(t^*)}^{[\nu]}(W_i) = \widehat{W}_i^{[\nu]}(t^*).$$

Analogously, the unique maximum or minimum value of a functional marker and its (higher-order) derivatives can be estimated as (Gasser and Müller, 1984)

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{MAX}}(W_i) = \sup_{t\in\mathcal{T}} \widehat{W}_i^{[\nu]}(t) \quad \text{or} \quad \phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{MIN}}(W_i) = \inf_{t\in\mathcal{T}} \widehat{W}_i^{[\nu]}(t).$$

Time-specific funtionals. Assume that $X_i^{(\nu)}$ attains a certain threshold value η. Then, its timing can be non-parametrically estimated by its empirical counterpart of $\widehat{W}_i^{[\nu]}$:

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{TIME}(\eta)}(W_i) = \inf\{t \in \mathcal{T} : \widehat{W}_i^{[\nu]}(t) = \eta\}.$$

If $\widehat{W}_i^{[\nu]}$ does not attain η, we define $\phi^{[\nu]}_{N,\mathrm{TIME}(\eta)}(W_i) = 0$. Similarly, the timing of a unique maximum can be estimated by (Gasser and Müller, 1984):

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{tMAX}}(W_i) = \inf\{t \in \mathcal{T} : \widehat{W}_i^{[\nu]}(t) = \sup_{u\in\mathcal{T}} \widehat{W}_i^{[\nu]}(u)\}.$$

In Appendix B.3, we provide and prove a theorem that states the consistency of each of the above estimators. Then by Theorem 3.2.1 from Section 3.2.5, this in turn establishes the consistency and asymptotic normality of the corresponding non-parametric BSA estimator.

3.4 Statistical Test for Selecting a Summary Functional

Our aim for this section is to identify quantitative features of a functional marker that are well-aligned with an ordinal scale of interest. Given that the ordinal scale reasonably reflects the severity of a certain clinical outcome, this effort amounts to producing sensible function-based biomarkers for understanding and assessing the same clinical mechanism in future studies. To address this objective, we develop a hypothesis test-

ing procedure for comparing the BSAs of different competing quantitative features of a functional marker. Specifically, suppose we are interested in determining whether a particular type of a summary functional $\phi_1(X)$ leads to a significantly better alignment with an ordinal outcome than that of a competing summary functional $\phi_2(X)$. For simplicity, let $\rho_{\text{bsa},1}$ and $\rho_{\text{bsa},2}$ denote the true BSA measures based on two different summary functionals $\phi_1(X)$ and $\phi_2(X)$, respectively. The null and alternative hypotheses can be formulated as

$$H_0 : \rho_{\text{bsa},1} = \rho_{\text{bsa},2} \quad \text{vs.} \quad H_1 : \rho_{\text{bsa},1} > \rho_{\text{bsa},2}.$$

Using the asymptotic property of the proposed estimator and the delta method, we can formulate the Wald test statistic as

$$T_{N,n} = \frac{\sqrt{n}\big\{g(\widehat{\rho}_{\text{bsa},1}) - g(\widehat{\rho}_{\text{bsa},2})\big\}}{\widehat{V}_J} = \frac{D_{N,n}}{\widehat{V}_J} \xrightarrow{d} N(0,1), \tag{3.8}$$

where $g(a) = 0.5 \times \ln\{(1+a)/(1-a)\}$ (Fisher's Z-transformation) and $\widehat{V}_J$ denotes the estimated asymptotic standard error of $D_{N,n}$. Since the analytical form of the standard error is complicated, the jackknife method is used to estimate V_J. Specifically, let $\widehat{D}_{N,n}^{(-i)}$ denote the estimate for $D_{N,n}$ obtained from the data excluding (W_i, Y_i). Then the jackknife estimate of the asymptotic variance of $D_{N,n}$ is given by

$$\widehat{V}_J^2 = \frac{n-1}{n} \sum_{i=1}^{n} \left(\widehat{D}_{N,n}^{(-i)} - \frac{1}{n} \sum_{p=1}^{n} \widehat{D}_{N,n}^{(-p)} \right)^2.$$

Therefore, the null hypothesis can be rejected when the absolute value of $T_{N,n}$ is greater than the $100(1-\alpha)^{\text{th}}$ percentile of the standard normal distribution.

3.5 Simulations

We conducted simulations studies to assess the performance of the proposed approaches to evaluate alignment between functional markers and ordinal outcomes. Specifically, finite-sample performances of BSA estimators based on three special cases of summary functionals (AUC-type, magnitude-specific, and time-specific) were assessed. Initially, for the ordinal outcomes, we set $K = 3$ and generate Y from the multinomial distribution with equal probabilities, that is, $\Pr(Y = k) = 1/3$, $k = 1, 2, 3$.

Given each $Y = k$, the true functional markers X are generated over a time interval $\mathcal{T} = [0, 1]$ under five different scenarios depending on the type of a summary functional to be analyzed. For the AUC-type summary functionals, we generate $X(t)$ as a Gaussian process with mean functions $\mu(t) = k$ (scenario 1) and $\mu(t) = kt$ (scenario 2). Scenarios 1 and 2 represent a constant and improving degrees of alignment in terms of the crude AUC over the time interval, respectively. Performances based on the magnitude-specific summary functionals are evaluated using a Gaussian process with mean function $\mu(t) = k\sin(\pi t)$, whose unique maximum value 1 is attained at time 1/2 (scenario 3). Note that all Gaussian processes are generated with a common covariance function $\mathrm{Cov}(X(s), X(t)) = \exp\{-(s-t)^2\}$, $s, t \in \mathcal{T}$. We consider two scenarios for evaluating the finite-sample performance based on the time-specific summary functionals. In scenario 4, if $Y = 1$, $X(t) = \sin(2\pi t)$ with probability 1; if $Y = 2$, $X(t) = \sin(0.25\pi t)$ with probability 1; and if $Y = 3$, $X(t) = \sin(0.5\pi t)$ with probability 1. In scenario 5, if $Y = 1$, $X(t) = \sin(2\pi t)$ with probability 1; if $Y = 2$, $X(t) = \sin(0.66\pi t)$ with probability 1; and if $Y = 3$, $X(t) = \sin(\pi t)$ with probability 1. Time to reach $\eta = 1/2$ and time to reach the maximum value 1 are the quantitative features of interest in scenarios 4 and 5, respectively. Figure 3.2 illustrates the representative curve sample (one for each ordinal category), the type(s) of summary functional we are targeting and the corresponding true BSA value(s) for

each of the five scenarios.

To assess the sensitivity of the proposed framework to varying density of time points, we consider the following five study designs: (a) unbalanced design with N_i following a Poisson distribution with mean 20; (b) unbalanced design with N_i following a Poisson distribution with mean 40; (c) balanced design with $N_i = 20$ per subject; (d) balanced design with $N_i = 40$ per subject; and (e) balanced design with $N_i = 60$ per subject. Except for the two fixed endpoints ($t_{i0} = 0$ and $t_{iN_i} = 1$), the N_i number of observation times in all these study designs are randomly drawn from a uniformly distributed grid $\mathcal{T}_{\text{grid}} = \{(u-1)/119,\ u = 1, \ldots, 120\}$ separately for each subject.

In order to mimic as closely as possible a real situation, we further contaminate the generated functional markers based on the model (3.3) at each time point, assuming that the measurement errors ϵ are independent and identically distributed N(0, 0.1) random variables. In all configurations, we obtain the Gasser-Müller kernel estimators (3.7) evaluated on 300 output design points using a polynomial kernel of degree 2 (Müller, 1984) and an automatically adapted global "plug-in" bandwidth that is asymptotically optimal with respect to the mean integrated square error (MISE) (Gasser et al., 1991). Standard error estimates and 95% CIs are computed based on the formulas (3.5) and (3.6), respectively. Results presented in Table 3.1 are based on 1000 simulated datasets of size $n = 40$ and 60.

From Table 3.1, we see that the proposed method exhibits satisfactory finite-sample performance. Empirical biases are generally low, implying that the corresponding BSA estimates quickly converge to the respective true values. But they do tend to be slightly larger for magnitude- and time-specific summary functionals when the data are highly sparse; see cases (a) and (c). Therefore, when magnitude-specific or time-specific summary functional is considered, we recommend using functional markers that are collected on at least average of 25 time points to produce reliable

Figure 3.2: Representative curve sample (3 for each ordinal category), the type(s) of summary functional we are targeting and the corresponding true BSA value(s) for each of the five scenarios. The solid lines denote functional markers paired with $Y = 1$; the dashed lines denote functional markers paired with $Y = 2$; and the dotted lines denote functional markers paired with $Y = 3$.

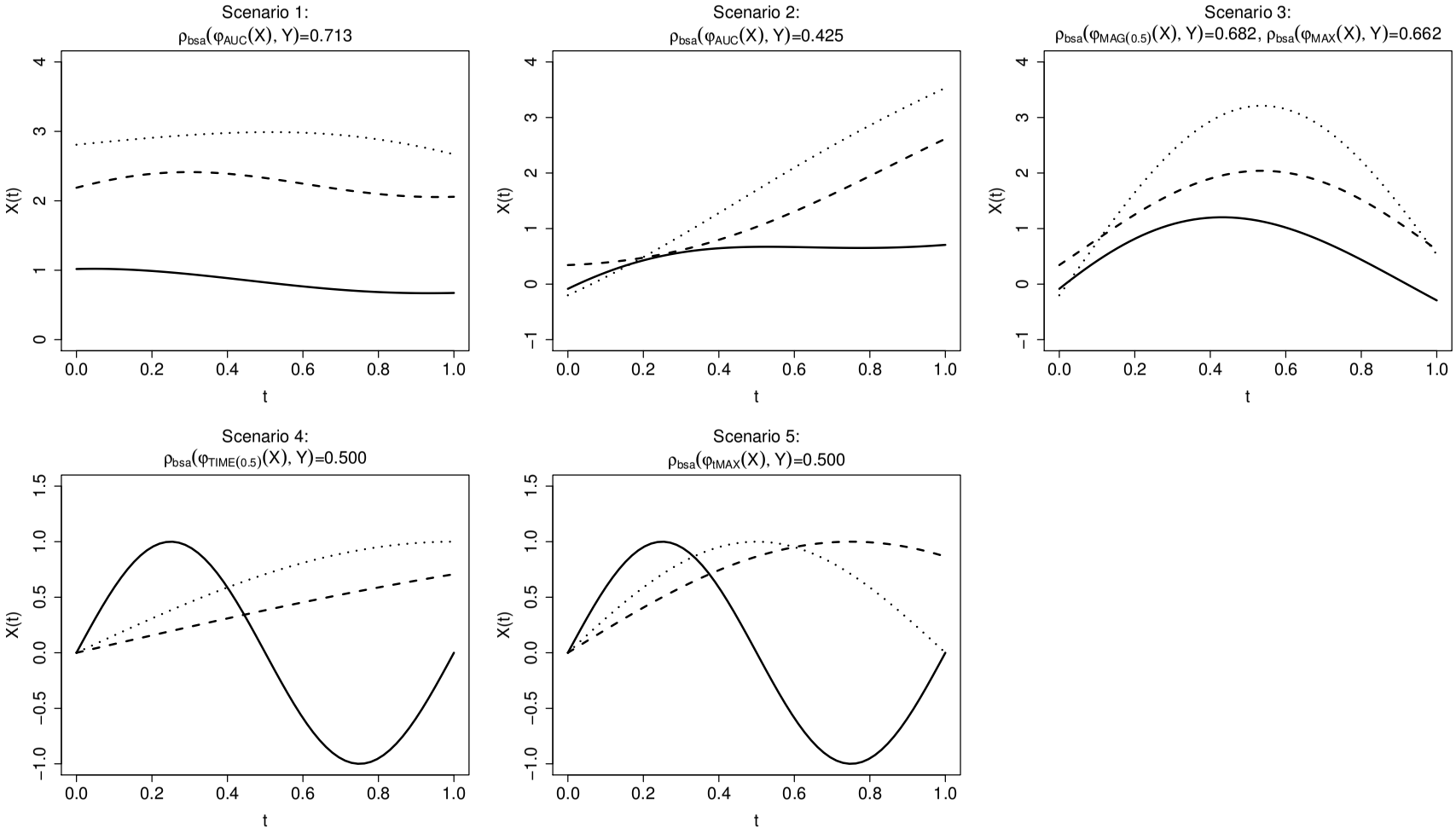


Table 3.1: Simulation results on proposed BSA measures: mean of 1000 biases (EmpBias), standard deviation of 1000 BSA estimates(EmpSD), mean of 1000 standard error estimates(EstSE) and proportion of 95% CIs containing the true BSA value (Cov95). N denotes the five study designs: (a) unbalanced design with N_i following a Poisson distribution with mean 20; (b) unbalanced design with N_i following a Poisson distribution with mean 40; (c) balanced design with $N_i = 20$; (d) balanced design with $N_i = 40$; and (e) balanced design with $N_i = 60$.

			$n = 40$				$n = 60$			
Scenario	True BSA Values	N	EmpBias	EmpSD	EstSE	Cov95	EmpBias	EmpSD	EstSE	Cov95
1	$\rho_{\text{bsa}}(\phi_{\text{AUC}}(X), Y)$	(a)	-0.001	0.096	0.095	0.938	-0.001	0.076	0.077	0.944
	$= 0.713$	(b)	0.003	0.096	0.095	0.942	0.001	0.077	0.076	0.936
		(c)	-0.001	0.094	0.096	0.945	-0.005	0.074	0.076	0.948
		(d)	0.001	0.099	0.096	0.918	-0.001	0.075	0.076	0.940
		(e)	-0.002	0.094	0.096	0.945	-0.002	0.075	0.077	0.946
2	$\rho_{\text{bsa}}(\phi_{\text{AUC}}(X), Y)$	(a)	-0.001	0.149	0.151	0.947	-0.006	0.119	0.122	0.961
	$= 0.425$	(b)	0.002	0.152	0.151	0.935	0.000	0.121	0.121	0.945
		(c)	-0.003	0.146	0.153	0.952	-0.008	0.117	0.122	0.956
		(d)	0.005	0.154	0.151	0.942	-0.002	0.118	0.121	0.948
		(e)	-0.004	0.147	0.153	0.950	0.000	0.115	0.121	0.956
3	$\rho_{\text{bsa}}(\phi_{\text{MAG(1/2)}}(X), Y)$	(a)	-0.012	0.103	0.106	0.942	-0.014	0.081	0.085	0.959
	$= 0.682$	(b)	-0.010	0.104	0.106	0.948	-0.008	0.081	0.084	0.942
		(c)	-0.011	0.106	0.105	0.938	-0.013	0.082	0.085	0.960
		(d)	-0.007	0.102	0.104	0.945	-0.003	0.081	0.083	0.935
		(e)	-0.003	0.102	0.103	0.942	-0.007	0.086	0.084	0.939
3	$\rho_{\text{bsa}}(\phi_{\text{MAX}}(X), Y)$	(a)	-0.011	0.108	0.111	0.936	-0.014	0.084	0.089	0.956
	$= 0.662$	(b)	-0.008	0.108	0.110	0.948	-0.006	0.086	0.087	0.933
		(c)	-0.011	0.111	0.110	0.944	-0.011	0.085	0.088	0.952
		(d)	-0.006	0.107	0.108	0.939	-0.002	0.084	0.086	0.938
		(e)	0.000	0.105	0.107	0.943	-0.005	0.089	0.087	0.933
4	$\rho_{\text{bsa}}(\phi_{\text{TIME(1/2)}}(X), Y)$	(a)	0.008	0.065	0.060	0.963	0.011	0.050	0.048	0.918
	$= 0.500$	(b)	0.010	0.039	0.030	0.952	0.010	0.031	0.026	0.971
		(c)	0.009	0.063	0.058	0.972	0.009	0.051	0.048	0.904
		(d)	0.008	0.039	0.030	0.971	0.010	0.029	0.025	0.972
		(e)	0.007	0.026	0.018	0.900	0.006	0.020	0.015	0.967
5	$\rho_{\text{bsa}}(\phi_{\text{tMAX}}(X), Y)$	(a)	0.020	0.044	0.039	0.978	0.019	0.036	0.033	0.916
	$= 0.500$	(b)	0.018	0.024	0.021	0.966	0.019	0.020	0.018	0.973
		(c)	0.020	0.043	0.039	0.981	0.019	0.034	0.031	0.926
		(d)	0.016	0.025	0.021	0.969	0.016	0.018	0.017	0.980
		(e)	0.012	0.018	0.014	0.947	0.011	0.014	0.012	0.982

BSA estimates. The estimated standard errors rapidly approach the empirical standard deviations as sample size increases in all configurations, suggesting that the jackknife procedure based on the Fisher's Z-transformation works well regardless of the study design and choice of a summary functional. Likewise, the 95% CIs have coverage probabilities that are all close to the nominal level.

We further evaluate the performance of the hypothesis testing procedure presented in Section 3.4. Specifically, empirical rejection rates of H_0 are obtained under the two selected scenarios from above and are presented in Table B.1 of Appendix B.4. In summary, the empirical rejection rates are very close to the nominal level of 0.05 when H_0 is correct and demonstrates adequate power if otherwise. Furthermore, the simulation study in Table 3.1 is repeated at the level of the first derivative of functional markers, and its results are presented in Table B.2 of Appendix B.4.

3.6 Renal Study

In this section, we apply the proposed approaches to the motivating renal study data described in Section 3.1. In the absence of a gold standard for detection of kidney obstruction, it is generally accepted that the nuclear medicine experts provide the best available interpretation of renal scans (Taylor and Garcia, 2014). Unfortunately, a vast majority of scan interpretations are conducted by general radiologists in the United States, and their lack of training and limited experience increase the error rate of the diagnosis (Taylor et al., 2008c). Under such circumstances, several quantitative features, such as maximum MAG3 photon count, time to reach maximum MAG3 photon count, etc., are derived from the baseline and post-furosemide renogram curves to assist readers arrive at correct diagnosis of kidney obstruction (Taylor et al., 2008c, Bao et al., 2011), and it is of ongoing interest to rigorously establish their connection with the underlying obstruction mechanism to prevent inappropriate patient man-

agement and unnecessary surgery.

The study was thus designed to assess and improve the diagnostic utility of the baseline and post-furosemide renogram curves under the proposed framework. A total of 108 patients (54 men [50%], 54 women [50%]; mean age, 57 years; SD, 17 years; range, 18-87 years), that is, 216 kidneys (108 kidneys from each side), with suspected kidney obstruction were enrolled in the study. Three selected nuclear medicine experts were asked to provide an ordinal rating of the obstruction status in each kidney. Their consensus ordinal rating was determined by majority of vote unless there was substantial disagreement. At baseline, 145 kidneys (68 left kidneys and 77 right kidneys) were rated as "non-obstructed" ($Y = 1$), 12 kidneys (7 left kidneys and 5 right kidneys) were rated "equivocal" ($Y = 2$) and 59 kidneys (33 left kidneys and 26 right kidneys) were rated as "obstructed" ($Y = 3$).

Baseline renogram curves were initially collected for patients referred for suspected obstruction (see the left panel of Figure 3.1). MAG3 photon counts in the region of interest (ROI) around each kidney were measured at 59 distinct time points over a period of 24 minutes. Each patient further received an intravenous injection of furosemide, a potent diuretic, and a second (post-furosemide) renogram curve was obtained with an additional 20 minutes (see the right panel of Figure 3.1). Herein, MAG3 photon counts were measured at 40 time points using a framing rate of 30 seconds. Note that both curves have equally-sparsed domains for all subjects; that is, $t_{ij} \equiv t_j$ and $N_i \equiv N$ for all i.

Our choice of quantitative features for both renogram curves was mainly guided by available *a priori* scientific information. Specifically, Bao et al. (2011)'s study suggests that a set of quantitative features that reflects the degree of MAG3 excretion from kidneys at baseline is strongly related to the obstruction status. Based on this information, we considered four quantitative features of the baseline renogram: a) crude AUC (ϕ_{AUC}); b) first derivative AUC ($\phi^{[1]}_{\text{AUC}}$); c) time to maximum of the crude

curve (ϕ_{tMAX}); and d) minimum rate of change ($\phi^{[1]}_{\text{MIN}}$). Furthermore, Eskild-Jensen et al. (2004) showed that MAG3 accumulates in the ROI without any excretion for the first 2-3 minutes regardless of the obstruction status. By combining this information with the empirical findings we drew from the patterns of the baseline renogram curves, we not only estimated each AUC-type functional on the entire time period $\mathcal{T}$, but also estimated it over the two sub-time intervals, $\mathcal{T}_1 = [0, 10]$ and $\mathcal{T}_2 = [10, 24]$, dichotomized at the ten minute milestone. For the post-furosemide renogram curves, Bao et al. (2011) suggests the importance of their overall MAG3 intensity in detecting kidney obstruction. Accordingly, two quantitative features were chosen: a) crude AUC (ϕ_{AUC}); and b) maximum of the crude curve (ϕ_{MAX}).

We first obtained the Gasser-Müller kernel estimates (3.7) of the crude ($\nu = 0$) renogram curves and their first derivatives ($\nu = 1$) using a polynomial kernel of degree 2 and 3, respectively (Müller, 1984). In both cases, the Gasser-Müller kernel estimators were evaluated on 300 design points using a data-driven global bandwidth that is asymptotically optimal with respect to MISE (Gasser et al., 1991).

Table 3.2 presents the BSA estimates between the four selected summary functionals (SFs) of baseline renogram curves and the experts' consensus ordinal ratings in each side of the kidney. Crude AUCs exhibit poor alignment with the experts ratings in both left (estimated BSA = 0.04; 95% CI = -0.13 to 0.20) and right (estimated BSA = -0.02; 95% CI = -0.15 to 0.12) kidneys. Similar conclusions can be drawn at the level of each sub-time interval. First derivative AUCs show a slightly better alignment in both left (estimated BSA = 0.32; 95% CI = 0.15 to 0.47) and right (estimated BSA = 0.15; 95% CI = -0.03 to 0.32) kidneys, but each of the BSA estimates is not large enough to conclude its diagnostic utility. However, a further analysis at the sub-time interval level unveils a noticeably better alignment of the first derivative AUCs evaluated on $\mathcal{T}_2$ ($\phi^{[1]}_{\text{AUC},\mathcal{T}_2}$), especially in the left kidneys (estimated BSA = 0.76; 95% CI = 0.60 to 0.86). Results of the hypothesis tests suggest that the degree of

Table 3.2: Estimated BSA measures based on four types of summary functionals (SFs) and results of hypothesis tests comparing their BSA values for baseline renogram data. P-values listed in the last column are from testing equality of BSA measures evaluated on two different sub-scan periods.

Kidney	SF	Estimated BSA (95% CI)			P-value
		$\mathcal{T} = [0, 24]$	$\mathcal{T}_1 = [0, 10]$	$\mathcal{T}_2 = [10, 24]$	
Left	ϕ_{AUC}	0.04 (-0.13, 0.20)	-0.15 (-0.31, 0.01)	0.19 (0.02, 0.35)	<.001
	$\phi^{[1]}_{\text{AUC}}$	0.32 (0.15, 0.47)	0.02 (-0.15, 0.18)	0.76 (0.60, 0.86)	<.001
	ϕ_{tMAX}	0.58 (0.38, 0.73)	–	–	–
	$\phi^{[1]}_{\text{MIN}}$	0.69 (0.52, 0.80)	–	–	–
Right	ϕ_{AUC}	-0.02 (-0.15, 0.12)	-0.14 (-0.26, -0.01)	0.07 (-0.07, 0.20)	<.001
	$\phi^{[1]}_{\text{AUC}}$	0.15 (-0.03, 0.32)	-0.03 (-0.17, 0.12)	0.51 (0.30, 0.67)	<.001
	ϕ_{tMAX}	0.37 (0.16, 0.54)	–	–	–
	$\phi^{[1]}_{\text{MIN}}$	0.45 (0.27, 0.60)	–	–	–

* *P-values from testing equality of BSA measures ($\phi^{[1]}_{\text{AUC},\mathcal{T}_2}$ vs. Other type SFs)*

		ϕ_{tMAX}	$\phi^{[1]}_{\text{MIN}}$
Left	$\phi^{[1]}_{\text{AUC},\mathcal{T}_2}$	0.045	0.314
Right	$\phi^{[1]}_{\text{AUC},\mathcal{T}_2}$	0.178	0.544

alignment of the first derivative AUCs evaluated on $\mathcal{T}_2$ is significantly stronger than those evaluated on $\mathcal{T}_1$ in both kidneys (both P-values $< .001$). Furthermore, both times to maximum value of the crude curves and minimum rates of change exhibit good alignment with the experts consensus in both kidneys. Hypothesis test results shown at the bottom of Table 3.2 suggest that their BSA values are as good as those of the first derivative AUCs evaluated on $\mathcal{T}_2$ (all P-values are close to or greater than 0.05).

Table 3.3: Estimated BSA measures based on two types of summary functionals (SFs) and results of hypothesis tests comparing their BSA values (P-value) for post-furosemide renogram data.

	Estimated BSA (95% CI)		
Kidney	ϕ_{AUC}	ϕ_{MAX}	P-value
Left	0.73 (0.57, 0.84)	0.67 (0.49, 0.80)	0.004
Right	0.55 (0.37, 0.69)	0.48 (0.30, 0.63)	0.002

Table 3.3 presents the BSA estimates between the two selected summary functionals (SFs) of baseline renogram curves and the experts ratings in each side of the kidney. Crude AUCs over the entire scan period are well aligned according to the expert consensus in both left (estimated BSA = 0.73; 95% CI = 0.57 to 0.84) and right (estimated BSA = 0.55; 95% CI = 0.37 to 0.69) kidneys. Maximum values are also well aligned with the expert ratings, but its degree falls short of that of the crude AUCs (both P-values < 0.01).

These results suggest a high diagnostic utility of the first derivative AUCs in the baseline renogram curves during the last 15 minutes of the scan period. Specifically, a relatively high positive overall rate of change in the baseline renogram curve at this period strongly suggests that the kidney is obstructed as implied by the experts. Considering the significant time and cost involved in performing the post-furosemide

scan (Taylor et al., 2008c), such finding can serve as a useful guideline for replicating experts' opinions on kidney obstruction and determining the need for the second scan in many practical settings. If the post-furosemide renogram curve is available for the patient, then the crude AUC over the entire scan period provides a firm basis for diagnosing kidney obstruction.

3.7 Discussion

In this Chapter, we propose a novel framework based on BSA that is practically useful for assessing alignment between an ordinal measurement and quantitative features that are commonly derived from functional markers. Our strategy is to adopt a general class of summary functionals that can flexibly incorporate multiple types of quantitative features in a systematic manner. Smoothing techniques, such kernel and spline methods, can be employed to account for the sampling variability and measurement error in observed functional data. In addition to estimation, we also address hypothesis testing for comparing a pair of candidate summary functionals in terms of their importance on the ordinal outcome. As suggested by the motivating example of the renal study, this research endeavor may help rigorously evaluate the usefulness of existing or novel quantitative features derived from the renogram curves for detecting kidney obstruction.

In practice, *a priori* scientific basis for generating functional data can dictate the choice of summary functions. This is indeed the case in our renal study. If a kidney operates normally, urine drains rapidly down the ureter to the bladder. With this concept, MAG3 is injected to the body to track how MAG3 travels down the ureter from the kidney to the bladder, and the renogram curve is generated by repeatedly measuring the MAG3 photon count inside the kidney over time. Hence, certain parts of the curve depict how fast the MAG3 is removed from the kidney, how long it

takes the MAG3 to produce maximum activity, etc., all of which provide a detailed account of the functional aspects of the kidney (ability to excrete, absorb, etc.). Other examples can be found in pharmacokinetic studies where the objective is to quantify the absorption, distribution, metabolism, and excretion of drug compounds over time in the body. The three common important quantitative features of a plasma drug concentration-time curve that are widely used to address this objective are: AUC (total drug exposure over time), Cmax (the peak plasma concentration) and tmax (time to reach Cmax) (Craig and Stitzel, 2004). In summary, the nature of a scientific experiment can guide the choice of summary functionals while providing sensible interpretations for them.

Chapter 4

Evaluating Quantitative Features of Functional Markers Based on Area Under the Receiver Operating Characteristic Curve

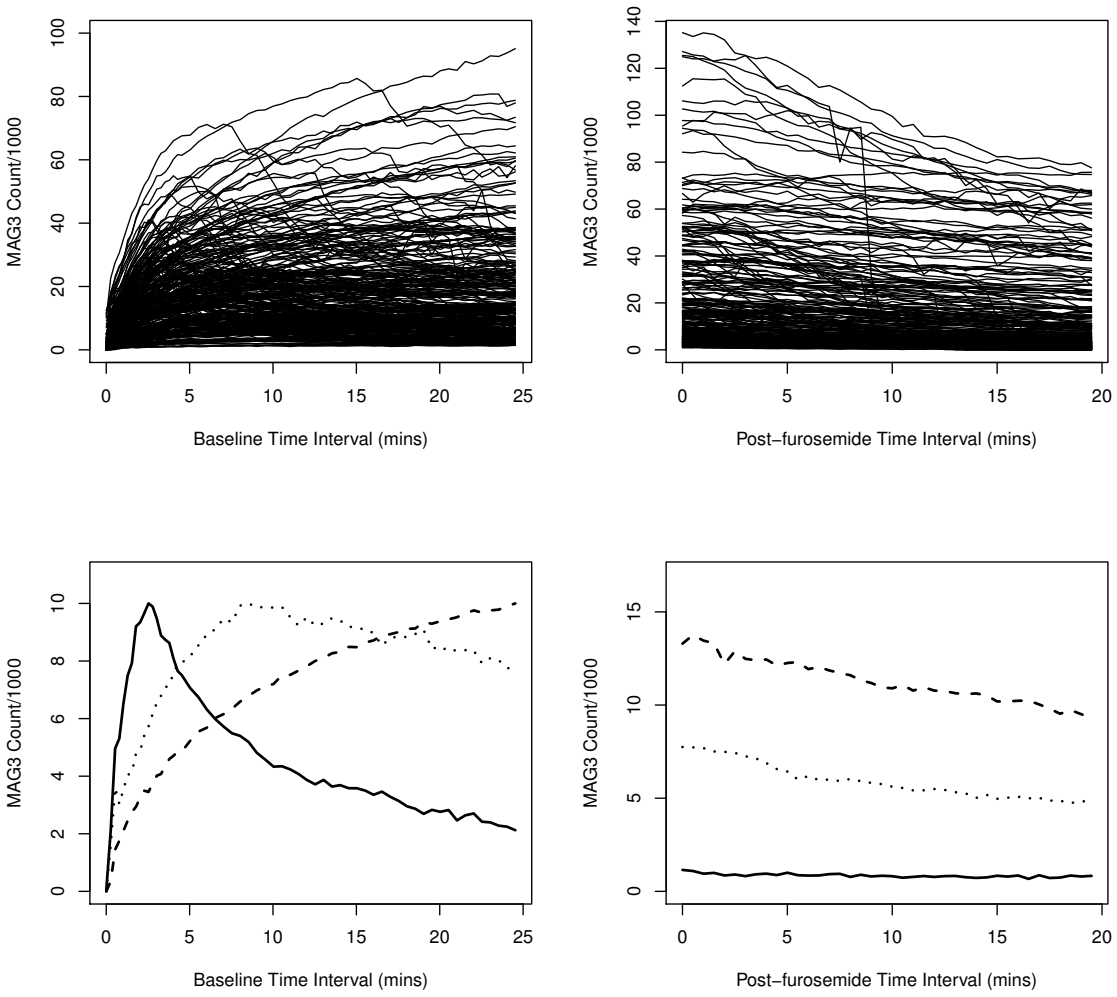
4.1 Introduction

A quality of disease management and clinical decision-making heavily depends on the availability of good diagnostic markers. With advancements in technology, more and more cutting-edge, non-invasive medical devices are being used to diagnose and monitor diseases. The increasing complexity of data they generate, however, often pose unique statistical challenges for establishing a clinically interpretative relationship between the data-derived markers and disease pathology.

In this Chapter, we specifically focus on one such type of data, namely functional markers which are increasingly being produced by modern devices. The unit of observation for each functional marker is a smooth continuous curve (function) defined on a time or space domain, and its flexible and dynamic structure is a rich source of clinical information (Ramsay and Silverman, 2005). It is thus typical in clinical research to describe and diagnose a disease using a set of "quantitative features" that characterize various dynamic, interpretative patterns of a functional marker, such as area under the curve, maximum value, time to reach maximum value and average velocity. Examples of their usage can be found in many biomedical fields, including pharmacokinetics (Craig and Stitzel, 2004), Alzheimer's disease study (Weiner et al., 2012) and cardiac safety assessment (Zhou and Sedransk, 2013).

The wide-usage of quantitative features of functional markers for diagnostic purposes calls for the need to rigorously evaluate their diagnostic utility. For instance, consider the renal study that has motivated our work. Obstruction to urine drainage from a kidney (kidney obstruction) is a serious clinical problem that can lead to irreversible loss of renal function if not properly treated. In recent years, diuresis renography has been widely used as a high-tech, non-invasive procedure to screen and diagnose of kidney obstruction (Taylor et al., 2008c). The procedure begins with an intravenous injection of a gamma emitting tracer, Technetium-99m Mercaptoacetyltriglycine (MAG3), that is rapidly removed from the blood by the kidneys

Figure 4.1: Top panel represents baseline (left) and post-furosemide (right) renogram curves of 275 kidneys. The bottom panel presents baseline (left) and post-furosemide (right) renogram curves of kidneys that are diagnosed as "non-obstructed" (solid lines), "obstructed" (dashed lines) and "equivocal" (dotted lines).



and then travels down the ureters from the kidney to the bladder. Photons emitted by tracer are then imaged and quantified in a region of interest (ROI) over each side of kidney, producing a set of renogram curves (functional markers) (Taylor et al., 2012). The first renogram curve (called baseline) represents the MAG3 photon counts detected at 59 time points during an initial period of 24 minutes. The second renogram curve (called post-furosemide) is obtained at 40 time points during an additional period of 20 minutes after an intravenous injection of furosemide, a potent diuretic. The top left and right columns of Figure 4.1 respectively depict baseline and post-furosemide renogram curves of 275 kidneys stored in Emory University Hospital's archived database.

There are several important, interpretative patterns of the renogram curves that are known to strongly related to the renal function; for example, the speed of initial MAG3 uptake in the kidney, the rate of MAG3 excretion to the bladder, etc (Mettler and Guiberteau, 2012). To illustrate, consider the baseline renogram curve of a non-obstructed kidney in the bottom left panel of Figure 4.1 (see solid lines). The curve is characterized by a quick uptake and excretion of MAG3. On the other hand, the baseline renogram curve of an obstructed kidney is characterized by a prolonged period of MAG3 accumulation with no or poor excretion (see dashed lines in the bottom left panel of Figure 4.1), a trend which persists throughout the post-furosemide renogram (see dashed lines in the bottom right panel of Figure 4.1).

A common procedure to assist physicians arrive at prompt and accurate diagnosis of kidney obstruction is to derive and study a set of quantitative features (e.g., time to maximum MAG3, maximum MAG3, etc.) that adequately reflect the aforementioned patterns (Taylor et al., 2012). For example, time to maximum MAG3 for an obstructed kidney is typically large, given its lack of absorption capacity. However, a high kidney-to-kidney variability in renogram curves is typical as seen from the top panel of Figure 4.1, and many show less distinctive patterns. For instance, the

renogram of the "equivocal" kidney in Figure 4.1 (see dotted lines) show patterns somewhat between those of non-obstructed and obstructed kidneys. Therefore, in many cases, accurate diagnosis of kidney obstruction using quantitative features of renogram curves requires substantial expertise in renal physiology and MAG3 pharmacokinetics (Taylor and Garcia, 2014).

Unfortunately, a majority of diuresis renography scans in United States are interpreted by general radiologists who have less than 4 months of training in nuclear medicine (Taylor and Garcia, 2014). The ensuing clinical problem is that these radiologists tend to select and utilize features based on ad hoc blending of intuition and past practice without proper guidance and scientific justification (Taylor et al., 2008c). Many radiologists compensate their limited experience by overrelying on a single feature such as the time to half-maximum. In fact, such naïve and uninformed reliance on a single feature is currently a leading cause of erroneous diagnosis, inappropriate patient management and unnecessary renal surgery (Taylor et al., 2008c). It is thus of interest to rigorously evaluate the diagnostic accuracy of various quantitative features of renogram curves and establish scientifically justified guidance regarding their selection and application.

In cases where a new technology is implemented and optimal thresholds are not verified, the area under the ROC curve (AUC) is often a preferred summary measure that aggregates diagnostic performance information for all thresholds (Pepe, 2003, Dodd and Pepe, 2003). It is thus sensible to apply AUC to evaluate quantitative features of newly emerging functional markers. There are three notable challenges for working with functional markers in general (Wang et al., 2016): 1) their infinite-dimensional structure renders parametric modeling too restrictive and calls for the need to adopt semi- or non-parametic approaches; (2) their true functional form should be recovered from discrete, repeated and error-prone observations; and (3) they belong to a space of square-integrable functions (non-Euclidean space), posing

additional complications in theory. Of course, one can choose to bypass these challenges by directly using observed repeated data (raw MAG3 photon counts over the scan period) to calculate quantitative features and obtain their AUC. But doing so fails to account for possible measurement error and may produce biased results. This naïve approach also ignores inherent smoothness of data, by which existence of several useful derivative-level features (e.g., average and maximum velocity) is conceptually and mathematically justified.

The goal of this Chapter is to develop a novel statistical framework for evaluating quantitative features based on AUC, while appropriately addressing the aforementioned challenges that are unique to functional markers. Our strategy is to represent different types of quantitative features by *summary functionals* (Jang et al., 2019), defined as a set of mappings from a space of square-integrable functions to a real line. This approach offers both mathematical rigor and conceptual flexibility for studying AUC of a large class of important, widely-used quantitative features, including functional area under the curve (FAUC), magnitude-specific and time-specific types. For accurate estimation of AUC, we propose a two-stage procedure in which a quantitative feature of interest (summary functional) is first appropriately estimated from discrete, error-prone measurements, and then plugged into a Mann-Whitney type statistic. We also provide an inferential framework for comparing diagnostic utility of a pair of candidate quantitative features.

Pathological mechanisms of many diseases are prone to a large population heterogeneity. As such, the use of covariate-specific AUC has been advocated to tailor the use of certain markers to specific high-benefit subpopulations (Pepe, 1998, Janes and Pepe, 2008). In our motivating renal study, a normal range of several quantitative features of renogram curves has been found to vary substantially with age (Esteves et al., 2006), suggesting potential variations in the diagnostic accuracy of these features among different age groups. Therefore, we propose a sensible adaptation of a

semi-parametric AUC regression framework introduced by Dodd and Pepe (2003) to systematically assess covariate effects on the diagnostic accuracy of various quantitative features. There are some important subtlety in formulating estimating equations for our regression model involving quantitative features. That is, outcomes are based on pairwise comparisons of features which are not directly observed but should be estimated from data. This demands extra work to establish asymptotic properties of the regression parameter estimates.

The remainder of the Chapter is organized as follows. In Section 4.2, we review the concept of summary functionals and present the estimation procedure based on appropriate kernel smoothing. In Section 4.3, we propose a framework for evaluating the diagnostic accuracy of a wide class of quantitative features (summary functionals). We study the asymptotic properties of the proposed two-stage AUC estimator and develop inferential procedures including variance estimation, confidence intervals and hypothesis testing. In Section 4.4, we introduce a semi-parametric AUC regression framework for quantitative features. We propose estimating equations that appropriately account for unobserved outcomes and carefully study their asymptotic properties. In Section 4.5, we report the results of a simulation study conducted to evaluate the performance of the proposed approaches. The application of our methods to a renal study is illustrated in Section 4.6. Finally, we conclude with some remarks in Section 4.7.

4.2 Representing Quantitative Features via Summary Functionals

4.2.1 General formulation of a summary functional

Let X_i denote the functional marker of ith subject, defined on the closed, bounded domain (time interval) $\mathcal{T} \subset \mathbb{R}$. In this notation, $X_i(t)$ denotes X_i evaluated at a given time point $t \in \mathcal{T}$. We assume that X_i belongs to a functional space $\mathcal{F}_\omega = \{f\colon \mathcal{T} \to \mathbb{R}$; for nonnegative integer ω, f is square integrable and ω-times continuously differentiable at any $t \in \mathcal{T}\}$.

Jang et al. (2019) recently introduced a concept of summary functional that can flexibly represent various quantitative features of a functional marker. Summary functional is defined as a mapping from $\mathcal{F}_\omega$ to $\mathbb{R}$, that is, $\phi : \mathcal{F}_\omega \to \mathbb{R}$. Different choice of the mapping (functional form) ϕ leads to different quantitative features; for instance, $\phi(X_i) = \int_{\mathcal{T}} X_i(t)dt$ and $\phi(X_i) = \sup_{t\in\mathcal{T}} X_i(t)$ denotes the FAUC and maximum value of the ith functional marker, respectively.

4.2.2 Three special (widely-used) cases of summary functionals

In this section, we introduce three classes of summary functionals that represent widely-used quantitative features. Denote $X_i^{(\nu)}$ as its ν^{th} derivative ($0 \le \nu \le \omega - 2$), with $X_i^{(0)} = X_i$.

FAUC-type funtionals. FAUC-type functionals summarize the functional marker over an entire ($\mathcal{T}$) or specified portion ($\mathcal{T}^* \subset \mathcal{T}$) of the time domain. They take the form

$$\phi(X_i) \coloneqq \phi^{[\nu]}_{\text{FAUC}(\mathcal{T})}(X_i) = \int_{\mathcal{T}} X_i^{(\nu)}(t)dt.$$

Setting $\nu = 0$ and $\nu = 1$ above gives the area under a crude curve (crude FAUC) and the area under the first derivative of a curve (first derivative FAUC), respectively.

Magnitude-specific functionals. Another important quantitative feature of a functional marker its magnitude at a specific point in time (milestone). Given $t^* \in \mathcal{T}$, a general form of the magnitude-specific functional (accommodating first and higher-order derivative levels) is

$$\phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{MAG}(t^*)}(X_i) = X_i^{(\nu)}(t^*).$$

A unique maximum or minimum magnitude of the functional marker often provides useful information and can be expressed as a summary functional as defined below:

$$\phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{MAX}}(X_i) = \sup_{t\in\mathcal{T}} X_i^{(\nu)}(t) \quad \text{or} \quad \phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{MIN}}(X_i) = \inf_{t\in\mathcal{T}} X_i^{(\nu)}(t).$$

Time-specific functionals. Time to attain a certain threshold value η of the functional marker is often clinically significant and can be captured by a time-specific functional that maps the space of functional markers to the time domain, i.e., $\phi : \mathcal{F}_\omega \to \mathcal{T}$. That is,

$$\phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{TIME}(\eta)}(X_i) = \inf\{t \in \mathcal{T} : X_i^{(\nu)}(t) = \eta\}.$$

The timing of a maximum or minimum value can be investigated using a time-specific functional of the form

$$\phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{tMAX}}(X_i) = \arg\sup_{t\in\mathcal{T}} X_i^{(\nu)}(t) \quad \text{or} \quad \phi(X_i) \coloneqq \phi^{[\nu]}_{\mathrm{tMIN}}(X_i) = \arg\inf_{t\in\mathcal{T}} X_i^{(\nu)}(t).$$

4.2.3 Estimation of summary functionals

In reality, the true functional marker X_i is not observed continuously in time; rather, an error-prone proxy W_i is observed at N_i discrete time points $\{(t_{i1}, t_{i2}, \ldots, t_{iN_i}) \in \mathcal{T} : t_{i1} < t_{i2} < \cdots < t_{iN_i}\}$. Accordingly, the true value of each summary functional $\phi(X_i)$ is unknown and should be estimated based on the observed data as $\phi_{N_i}(W_i)$.

The first step is to estimate the true functional marker X_i based on discrete noisy observations W_i. Several smoothing techniques, including kernel smoothing, moving average and smoothing splines, can be employed. For instance, smoothing splines (e.g., cubic B-splines) approximate the true function using the coefficients obtained via penalized least squares methods that aim to explicitly control the trade-off between fidelity to the data and roughness of the function estimate. An excellent reference on smoothing splines is Green and Silverman (1994).

In our work, we opt to a non-parametric smoothed estimate of the true underlying functional marker using the Gasser-Müller (GM) kernel estimator which has well-established asymptotic properties (Gasser and Müller, 1984, Müller, 1984, 1985). Herein, true X_i and observed W_i are related via the following functional measurement error model:

$$W_i(t_{ik}) = X_i(t_{ik}) + \epsilon_i(t_{ik}), \quad k = 1, \ldots, N_i, \tag{4.1}$$

where the random measurement error $\epsilon_i(t)$ follows an independent and identical distribution (iid) with $E(\epsilon_i) = 0$ and $\mathrm{Var}(\epsilon_i) = \sigma_e^2 < \infty$ for each t, mutually independent of X_i. Under model (4.1), the GM kernel estimator is given as

$$\widehat{W}_i^{[\nu]}(t) = \frac{1}{b_{N_i}^{\nu+1}} \sum_{k=1}^{N_i} \int_{d_{i,k-1}}^{d_{ik}} K_\nu\Big(\frac{t-u}{b_{N_i}}\Big) du \cdot W_i(t_{ij}), \tag{4.2}$$

where $d_{i,k-1} = (t_{ik} + t_{i,k-1})/2$ ($d_{i0} = 0, d_{N_i} = 1$), and b_{N_i} is a bandwidth satisfying $b_{N_i} \to 0$, $N_i b_{N_i} \to \infty$ as $N_i \to \infty$. K_ν is a kernel function of order (ν, ω) defined on a compact support $[-1, 1]$ and takes on zero values on the boundary points. This

estimator (4.2) is well known for its computational efficiency, producing accurate first or higher-order derivative estimates with a number of observed time points as small as 15 (Gasser et al., 1991).

Given a quantitative feature of interest and the corresponding summary functional $\phi(X_i)$, we propose to build its non-parametric estimator $\phi_{N_i}(W_i)$ based on the GM kernel estimator (4.2). Specifically, the estimators for the three special-case summary functionals (FAUC-type, magnitude-specific, and time-specific) are listed in the following.

FAUC-type funtionals. This type of summary functional can be estimated by the following Riemann sum of $\widehat{W}_i^{[\nu]}$ with respect to the output design points $\{t_{i1}, \ldots, t_{i,N_i}\}$:

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{FAUC}}(W_i) = \sum_{k=1}^{N_i-1} \widehat{W}_i^{[\nu]}(t_{ik})(t_{i,k+1} - t_{ik}).$$

Magnitude-specific funtionals. Given a specific time point $t^* \in \mathcal{T}$, a general magnitude-specific functional can be directly estimated by its empirical counterpart of $\widehat{W}_i^{[\nu]}$:

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{MAG}(t^*)}(W_i) = \widehat{W}_i^{[\nu]}(t^*).$$

Analogously, the unique maximum or minimum value of a functional marker and its (higher-order) derivatives can be estimated as (Gasser and Müller, 1984)

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{MAX}}(W_i) = \sup_{t\in\mathcal{T}} \widehat{W}_i^{[\nu]}(t) \quad \text{or} \quad \phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{MIN}}(W_i) = \inf_{t\in\mathcal{T}} \widehat{W}_i^{[\nu]}(t).$$

Time-specific funtionals. Assume that $X_i^{(\nu)}$ attains a certain threshold value η. Then, its timing can be non-parametrically estimated by its empirical counterpart of $\widehat{W}_i^{[\nu]}$:

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{TIME}(\eta)}(W_i) = \inf\{t \in \mathcal{T} : \widehat{W}_i^{[\nu]}(t) = \eta\}.$$

If $\widehat{W}_i^{[\nu]}$ does not attain η, we define $\phi^{[\nu]}_{N,\mathrm{TIME}(\eta)}(W_i) = 0$. Similarly, the timing of a unique maximum can be estimated by (Gasser and Müller, 1984):

$$\phi_{N_i}(W_i) \coloneqq \phi^{[\nu]}_{N_i,\mathrm{tMAX}}(W_i) = \inf\{t \in \mathcal{T} : \widehat{W}_i^{[\nu]}(t) = \sup_{u\in\mathcal{T}} \widehat{W}_i^{[\nu]}(u)\}.$$

Jang et al. (2019) showed that each of the above (three special-case) summary functional estimators $\phi_{N_i}(W_i)$ based on the GM kernel estimator (4.2) converges in probability to the true values $\phi(X_i)$ as $N_i \to \infty$. Specifically, given $b_{N_i} = \mathcal{O}\{(\log N_i/N_i)^{1/(2\omega+1)}\}$, the convergence rates for FAUC and magnitude summary functionals are of the order $\mathcal{O}_p(B_{N_i})$, and the rate for time-specific summary functionals is $\mathcal{O}_p(B^{\theta}_{N_i})$, where $B_{N_i} = (\log N_i/N_i)^{(\omega-\nu)/(2\omega+1)}$ and $0 < \theta \leq 1$, suggesting that the latter generally have slower convergence rates.

4.3 AUC Analysis of Quantitative Features

4.3.1 Formulation and estimation

In this section, we aim to evaluate the diagnostic utility of quantitative features of a functional marker for classifying "diseased" and "non-diseased" states (groups), denoted as D and $\overline{D}$, respectively. Let X_i^D and $X_j^{\overline{D}}$ ($i = 1,\ldots,n_D$; $j = 1,\ldots,n_{\overline{D}}$; $n = n_D + n_{\overline{D}}$) respectively denote functional markers of subjects from D and $\overline{D}$. Functional markers are assumed to be iid within groups and mutually independent between groups. Given a quantitative feature of interest, we directly consider the corresponding summary functionals $\phi(X_i^D)$ and $\phi(X_j^{\overline{D}})$, and assume that larger summary functional values indicate greater likelihood of disease ("positive" if summary functional value exceeds some given cutoff value c).

The ROC curve plots false-positives (1-specificity) versus true-positives (sensitivity) for varying cutoff values c: $\{\Pr(\phi(X_j^{\overline{D}}) > c), \Pr(\phi(X_i^D) > c)\}$. The area under

the ROC curve (AUC) summarizes performance information across all cutoff values and is equal to

$$\text{AUC}(\phi) \coloneqq \theta(\phi) = \Pr(\phi(X_i^D) > \phi(X_j^{\overline{D}})), \tag{4.3}$$

which represents a probability that a summary functional value of a randomly selected diseased subject is greater than that of a randomly selected non-diseased subject (Bamber, 1975). Note that we are assuming $\Pr(\phi(X_i^D) = \phi(X_j^{\overline{D}})) = 0$ as summary functional values (quantitative features) are continuous. $\text{AUC}(\phi) = 1$ corresponds to a quantitative feature (captured by ϕ) that perfectly classifies subjects into the two states, while $\text{AUC}(\phi) = 0.5$ denotes a feature that performs no better than chance. The closer $\text{AUC}(\phi)$ is to 1, the better the overall diagnostic performance of the feature of interest.

If true functional markers (X_i^D and $X_j^{\overline{D}}$'s) were completely observable, a non-parametric Mann-Whitney type statistic that compares the true summary functional value of each diseased subject to that of every other non-diseased subject in pairs would estimate $\theta(\phi)$:

$$\widehat{\theta}(\phi) = \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} I\{\phi(X_i^D) > \phi(X_j^{\overline{D}})\}, \tag{4.4}$$

where indicator function $I(\cdot)$ equals 1 if the bracketed expression is true and 0 otherwise.

In practice, however, functional markers from both groups are observed at discrete time points with measurement error as $\{W_i^D(t_{ik}),\ k = 1, \ldots, N_i\}$ and $\{W_j^{\overline{D}}(t_{jk}),\ k = 1, \ldots, N_j\}$. Accordingly, the estimated summary functionals (see Section 4.2.3 for examples) should replace the true ones in (4.4) as:

$$\widehat{\theta}(\phi_N) = \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} I\{\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})\}. \tag{4.5}$$

Ties can be handled by adding $0.5 \cdot I\{\phi_{N_i}(W_i^D) = \phi_{N_j}(W_j^{\overline{D}})\}$ to the kernel of (4.5); we omit the term without loss of generality. There are two potential sources of sampling error in the estimation of $\theta(\phi)$ using $\widehat{\theta}(\phi_N)$. The first source of error stems from using a two-sample Mann-Whitney statistic (4.4) to estimate the probability in (4.3). The second source of error comes from replacing the true summary functionals with their estimates and is specific to our situation involving functional markers. It is important that both sources of error are taken into consideration when studying the (asymptotic) properties of the proposed estimator (4.5).

4.3.2 Asymptotic properties

In Theorem 4.3.1, we state that the proposed estimator $\widehat{\theta}(\phi_N)$ given by (4.5) is a consistent estimator of $\theta(\phi)$ and asymptotically normal, as long as we can estimate the summary functional value of each subject using observed data.

Theorem 4.3.1. *Suppose* $\sup_{1\le i\le n_D} |\phi_{N_i}(W_i^D) - \phi(X_i^D)| \le \mathcal{O}_p(P_N^D)$ *and* $\sup_{1\le i\le n_{\overline{D}}} |\phi_{N_j}(W_j^{\overline{D}}) - \phi(X_j^{\overline{D}})| \le \mathcal{O}_p(P_N^{\overline{D}})$*, where* P_N^D *and* $P_N^{\overline{D}}$ *are nonnegative sequences. Let* $P_N = \max(P_N^D, P_N^{\overline{D}})$ *and* $n = n_D + n_{\overline{D}}$*. (i) If* $P_N \to 0$ *as* $n \to \infty$ *and the regularity conditions (A1)–(A3) provided in Appendix C.1 hold,* $\widehat{\theta}(\phi_N)$ *is a consistent estimator for* $\theta(\phi)$*. (ii) If* $\sqrt{n}P_N \to 0$ *as* $n \to \infty$ *and the regularity conditions (A1), (A4) and (A5) provided in Appendix C.1 hold,* $\sqrt{n}\{\widehat{\theta}(\phi_N) - \theta(\phi)\}$ *converges to mean zero normal distribution with variance* $\sigma^2_{\theta(\phi)}$*.*

The explicit formula for $\sigma^2_{\theta(\phi)}$ as well as the proof of the above theorem is given in Appendix C.1. The key idea of the proof is to consider the decomposition $\widehat{\theta}(\phi_N) - \theta(\phi) = T_1 + T_2$, where $T_1 = \widehat{\theta}(\phi_N) - \theta(\phi_N)$, $T_2 = \theta(\phi_N) - \theta(\phi)$ and $\theta(\phi_N) = \Pr(\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}}))$. Given sufficiently large fixed N_i and N_k values, the consistency and asymptotic normality of the first term T_1 can be established based on generalized U-statistics theory; T_2 can be shown negligible as $n = n_D + n_{\overline{D}} \to \infty$.

Note that since FAUC-type, magnitude-specific and time-specific summary functionals can be consistently estimated using observed data (via GM kernel estimator; see Section 4.2.3), the consistency and asymptotic normality of their AUC estimators are guaranteed by Theorem 4.3.1. To see this, set $\sup_i B^{\theta}_{N_i} = P^{D}_{N}$ and $\sup_j B^{\theta}_{N_j} = P^{\overline{D}}_{N}$.

4.3.3 Statistical inference

Given the rather complicated analytic form of $\sigma^2_{\theta(\phi)}$, we recommend a bootstrap approach for its estimation. Note that here the bootstrap sample is drawn separately from each group (D and $\overline{D}$) with replacement. Other non-parametric methods such as the jackknife, half-sampling or subsampling can also be used; see Efron (1981) for details of other applicable methods. The validity of the non-parametric approaches to asymptotic variance estimation is due to the fact that both $\mathrm{Var}(T_2)$ and $\mathrm{Cov}(T_1, T_2)$ (T_1 and T_2 are defined in Section 4.3.2 under Theorem 1) are asymptotically negligible given that summary functionals can be consistently estimated. Note that the consistency of the jackknife estimator is guaranteed by the fact that $\sqrt{n}\{\widehat{\theta}(\phi_N) - \theta(\phi)\}$ is, asymptotically, a U-statistic (Arvesen, 1969).

One can use normal approximation to construct confidence intervals (CIs) of AUC(ϕ). Since the scale for the AUC is restricted to $(0, 1)$, adopting a logit transformation may accelerate the convergence of $\widehat{\theta}(\phi_N)$ to asymptotic normality, especially when it is close to the boundary. Define $l(x) = \ln\{(x/(1-x)\}$, $l'(x) = dl(x)/dx$ and $l^{-1}(\cdot)$ as the inverse function of $l(\cdot)$. Using the delta method, the $100(1-\alpha)\%$ CI for AUC(ϕ) is constructed as

$$\Big[l^{-1}\big(\tilde{l} - z_{1-\alpha/2} \cdot \tilde{l}'\hat{s}\big), l^{-1}\big(\tilde{l} + z_{1-\alpha/2} \cdot \tilde{l}'\hat{s}\big)\Big],$$

where $\tilde{l} \equiv l\{\widehat{\theta}(\phi_N)\}$, $\tilde{l}' \equiv l'\{\widehat{\theta}(\phi_N)\}$, and $z_{1-\alpha/2}$ is the $100(1-\alpha/2)^{\text{th}}$ percentile of $N(0, 1)$.

Suppose we are interested in selecting a quantitative feature that is more likely to provide useful information about a patients disease state relative to others. To address this objective, we develop a hypothesis testing procedure that can determine whether a particular feature, captured by the summary functional ϕ_1, leads to a significantly better AUC than that of a competing feature captured by ϕ_2. The null and alternative hypotheses are

$$H_0 : \text{AUC}(\phi_1) = \text{AUC}(\phi_2) \quad \text{vs.} \quad H_1 : \text{AUC}(\phi_1) > \text{AUC}(\phi_2),$$

respectively. Let ϕ_{N1} and ϕ_{N2} respectively denote the estimated versions of ϕ_1 and ϕ_2. The hypothesis can be tested based on the following Wald test statistic:

$$T_{N,n} = \frac{\sqrt{n}\big[l\{\widehat{\theta}(\phi_{N1})\} - l\{\widehat{\theta}(\phi_{N2})\}\big]}{\widehat{V}_J} \xrightarrow{d} N(0,1),$$

where the denominator term $\widehat{V}_J$ represents the estimated asymptotic standard error of the numerator. $\widehat{V}_J$ can be obtained by bootstrapping the observations at the subject level. Given the significance level of α, the null hypothesis is rejected when $|T_{N,n}| > z_{1-\alpha}$.

4.4 Covariate-specific AUC Analysis of Quantitative Features

The diagnostic accuracy of a quantitative feature may depend on patient characteristics. In this section, we extend our framework to adjust for covariates and further investigate population heterogeneity in AUC of quantitative features.

4.4.1 Model Formulation

Suppose both functional marker and covariate data, $(X_i^D, \mathbf{Z}_i^D)$ and $(X_j^{\overline{D}}, \mathbf{Z}_j^{\overline{D}})$, are available for study subjects from D and $\overline{D}$. The covariate-specific AUC of a quantitative feature ϕ of interest, $\text{AUC}_{ij}(\phi)$, is defined as the probability that the ϕ value of a randomly selected diseased subject with covariate value $\mathbf{Z}_i^D = \mathbf{z}_i^D$ is greater than that of a randomly selected non-diseased subject with covariate value $\mathbf{Z}_j^{\overline{D}} = \mathbf{z}_j^{\overline{D}}$. That is,

$$\text{AUC}_{ij}(\phi) \coloneqq \theta_{ij}(\phi) = \Pr(\phi(X_i^D) > \phi(X_j^{\overline{D}}) \mid \mathbf{Z}_i^D = \mathbf{z}_i^D, \mathbf{Z}_j^{\overline{D}} = \mathbf{z}_j^{\overline{D}}). \tag{4.6}$$

In many cases, investigating the AUC between diseased and non-diseased subjects with a common covariate value $\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}} = \mathbf{z}$ is of subtantiative scientific interest. Under this setting, the covariate-specific AUC (4.6) becomes:

$$\text{AUC}_{\mathbf{z}}(\phi) \coloneqq \theta_{\mathbf{z}}(\phi) = \Pr(\phi(X_i^D) > \phi(X_j^{\overline{D}}) \mid \mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}} = \mathbf{z}). \tag{4.7}$$

Given definition (4.7), we propose to conduct a covariate-adjusted AUC analysis of quantitative features based on an appropriate adaptation of the semiparametric regression model introduced by Dodd and Pepe (2003):

$$\text{AUC}_{\mathbf{Z}}(\phi) \coloneqq \theta_{\mathbf{Z}}(\phi) = g^{-1}(\mathbf{Z}^T\boldsymbol{\beta}), \tag{4.8}$$

where $g(\cdot)$ is a monotone link function (e.g., logit and probit). $\boldsymbol{\beta}$ is a p-vector of unknown parameters that quantify covariate effects on the AUC. To illustrate, consider the model that uses a logit link function with a single covariate Z: $logit\{\theta_z(\phi)\} = \beta_0 + \beta_1 Z$. Here, $\exp(\beta_0 + \beta_1 z) = \text{AUC}_z(\phi)/\{1 - \text{AUC}_z(\phi)\}$ denotes the AUC odds in a subpopulation defined by covariate value $Z = z$, and accordingly $\exp(\beta_1)$ represents the AUC odds ratio between subpopulations corresponding to $Z = z + 1$ and $Z = z$. For instance, if $Z = 0$ and $Z = 1$ respectively denote males and females, $\exp(\beta_1)$ is

the AUC odds ratio (OR) for the quantitative feature ϕ in females versus males. If $\beta_1 > 0$, then ϕ is better at classifying diseased and non-diseased females than between diseased and non-diseased males.

4.4.2 Estimated estimating equations

Define $U_{ij} = I\{\phi(X_i^D) > \phi(X_j^{\overline{D}})\}$ $(i = 1, \ldots, n_D;\ j = 1, \ldots, n_{\overline{D}};\ n = n_D + n_{\overline{D}})$ such that

$$E(U_{ij} \mid \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}) = \Pr(\phi(X_i^D) > \phi(X_j^{\overline{D}}) \mid \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}) = g^{-1}\{(\mathbf{Z}_i^D)^T\boldsymbol{\beta}_0^D + (\mathbf{Z}_j^{\overline{D}})^T\boldsymbol{\beta}_0^{\overline{D}}\}$$
$$= g^{-1}(\mathbf{Z}\boldsymbol{\beta}_0), \quad \text{if} \quad \mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}} = \mathbf{Z}.$$

This suggests that if the true functional markers were available for all subjects, our model (4.8) would correspond to a generalized linear regression model for the binary variables U_{ij} restricted to (i, j) pairs with $\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}} = \mathbf{Z}$. Accordingly, the estimating equations for the regression parameters $\boldsymbol{\beta}$ in (4.8) are given by

$$S_n(\boldsymbol{\beta}) = \sum_{i=1}^{n_D}\sum_{j=1}^{n_{\overline{D}}} \frac{\partial\theta_{ij}}{\partial\boldsymbol{\beta}}\Omega_{ij}^{-1}(U_{ij} - \theta_{ij})I(\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}}) = \sum_{i=1}^{n_D}\sum_{j=1}^{n_{\overline{D}}} S_{ij}(\boldsymbol{\beta}), \qquad (4.9)$$

where $\theta_{ij} = g^{-1}\{(\mathbf{Z}_i^D)^T\boldsymbol{\beta}^D + (\mathbf{Z}_j^{\overline{D}})^T\boldsymbol{\beta}^{\overline{D}}\} = g^{-1}(\mathbf{Z}^T\boldsymbol{\beta})$ (for $\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}} = \mathbf{Z}$ pairs), and $\Omega_{ij} = \theta_{ij}(1 - \theta_{ij})$ is a variance function. Since $E\{S_n(\boldsymbol{\beta}_0)\} = \mathbf{0}$, the estimating equations (4.9) are the classic score equations for binary regression, except that U_{ij}'s are cross-correlated. Under some moderate conditions, Dodd and Pepe (2003) showed that the estimators $\widehat{\boldsymbol{\beta}}$ obtained from solving $S_n(\boldsymbol{\beta}) = \mathbf{0}$ are consistent and asymptotically normal.

The estimating equations (4.9), however, cannot be used as the true functional markers are unavailable in practice. Accordingly, we propose to replace the components in (4.9) by their respective proxy versions based on estimated summary func-

tionals (see Section 4.2.3) and obtain the parameter estimates based on the following *estimated* estimating equations:

$$S_{Nn}(\boldsymbol{\beta}) = \sum_{i=1}^{n_D}\sum_{j=1}^{n_{\overline{D}}} \frac{\partial \theta_{ij}}{\partial \boldsymbol{\beta}} \Omega_{ij}^{-1}(U_{Nij} - \theta_{ij}) I(\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}}) = \sum_{i=1}^{n_D}\sum_{j=1}^{n_{\overline{D}}} S_{Nij}(\boldsymbol{\beta}), \qquad (4.10)$$

where $U_{Nij} = I\{\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})\}$.

4.4.3 Estimation with continuous covariate

When the covariates are continuous or data are too sparse within covariate stata, the estimated estimating equations (4.10) may not be efficient or feasible as there may be few or no pairs from D and $\overline{D}$ with the identical covariate value. In such a case, the strategy is to temporarily consider estimated estimating equations (4.10) derived from a model that takes the form of $g\{\theta_{ij}(\phi)\} = \beta_0 + \beta_1 Z_i^D + \beta_2(Z_i^D - Z_j^{\overline{D}})$, and replace $I(\mathbf{Z}_i^D = \mathbf{Z}_j^{\overline{D}})$ with $I(||\mathbf{Z}_i^D - \mathbf{Z}_j^{\overline{D}}|| \leq \eta)$ to consider additional pairwise comparisons U_{ij} with covariates that are sufficiently close to each other (Dodd and Pepe, 2003, Liu and Zhou, 2013). Note that this model reduces to our target model (4.8): $g\{\theta_z(\phi)\} = \beta_0 + \beta_1 z$ for $Z_i^D = Z_j^{\overline{D}} = z$. Thus, our goal remains the same: to estimate β_1, which describes how the AUC for diseased and-nondiseased subjects at the same covariate level changes as that covariate varies.

There is a trade-off between bias and efficiency as η varies (Dodd and Pepe, 2003). By specifying large η, pairs with further-apart covariate values can be used for estimation to achieve high efficiency. But doing so imposes more structure in the model and may introduce some bias if the structure is misspecified. Small η imposes a less restrictive structure in the model but may be less efficient. One can choose the value of η manually or in a data-driven manner. With regards to the latter, we first note that conventional cross-validation approaches to select optimal bandwidth for smoothing in usual regression models cannot be applied as pairwise comparison data

U_{Nij} with $Z_i^D = Z_j^{\overline{D}} = z$ are not always available. Hence, for our case, we propose to utilize the fact that $\theta(\phi) = E\{\theta_{\mathbf{Z}}(\phi)\}$ and choose η as

$$\eta_{opt} = \arg\min_{\eta} |\widehat{\theta}(\phi_N) - \widehat{\theta}_*(\phi_N)|, \tag{4.11}$$

where $\widehat{\theta}_*(\phi_N) = n^{-1}\sum_{k=1}^{n} g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 z_k)$, with z_k coming from the pooled sample of diseased and non-diseased subjects. In practice, a grid search can be performed to compute $\widehat{\theta}_*(\phi_N)$ over candidate η values and select η_{opt}.

4.4.4 Asymptotic properties

Denote $\widehat{\boldsymbol{\beta}}_N$ as the solution to $S_{Nn}(\boldsymbol{\beta}) = \mathbf{0}$. Theorem 2 summarizes the large sample distribution properties of $\widehat{\boldsymbol{\beta}}_N$.

Theorem 4.4.1. *Suppose* $\phi_{N_i}(W_i^D) \xrightarrow{p} \phi(X_i^D)$ *and* $\phi_{N_j}(W_j^{\overline{D}}) \xrightarrow{p} \phi(X_j^{\overline{D}})$ *as* $N_i, N_j \to \infty$ *for all* $i = 1, \ldots, n_D$ *and* $j = 1, \ldots, n_{\overline{D}}$. *Then under the regularity conditions (B1)–(B8) and Lemmas C.2.1–C.2.4 provided in Appendix C.2,* $\widehat{\boldsymbol{\beta}}_N \xrightarrow{p} \boldsymbol{\beta}_0$ *and* $\sqrt{\frac{n_D n_{\overline{D}}}{n}}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\beta)$ *as* $n \to \infty$, *where* $\boldsymbol{\Sigma}_\beta = \mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-1}$. *The forms of* $\mathbf{Q}$ *and* $\boldsymbol{\Sigma}$ *are given in Appendix B.*

The proofs of Lemmas C.2.1–C.2.4 are given in Appendix C.2. Consistency is established by demonstrating that the original estimating equations $(n_D n_{\overline{D}})^{-1} S_n(\boldsymbol{\beta})$, whose solution is consistent, and the estimated estimating equations $(n_D n_{\overline{D}})^{-1} S_{Nn}(\boldsymbol{\beta})$ converge in probability to the same limit uniformly for $\boldsymbol{\beta}$ within the neighborhood of $\boldsymbol{\beta}_0$. Asymptotic normality is proved by applying the Projection Theorem of U-statistics (Serfling, 1980) to express $(n_D n_{\overline{D}})^{-1} S_{Nn}(\boldsymbol{\beta})$ as the sum of independent random variables and considering its Taylor expansion about $\boldsymbol{\beta}_0$. Note that although the explicit forms of $\boldsymbol{\Sigma}$ and $\mathbf{Q}$ are available in Appendix B, a bootstrap or jackknife variance estimate can be used for convenience. For instance, the jackknife estimator

is given by (Shao, 1992)

$$\widehat{\boldsymbol{\Sigma}}_{\beta} = \frac{n-p}{n}\sum_{k=1}^{n}(\widehat{\boldsymbol{\beta}}_N^{(-k)} - \widehat{\boldsymbol{\beta}}_N)(\widehat{\boldsymbol{\beta}}_N^{(-k)} - \widehat{\boldsymbol{\beta}}_N)^T,$$

where $\widehat{\boldsymbol{\beta}}_N^{(-k)}$ is the estimate of $\boldsymbol{\beta}$ based on the data with kth observation, $\{\phi_{N_k}(w_k), \mathbf{z}_k\}$, from the pooled sample of diseased and non-diseased subjects removed.

4.5 Simulations

We conducted simulation studies to evaluate the finite sample performance of the proposed methods. Firstly, the performance of the AUC estimator (4.5) and its inference procedures described in Section 4.3 was assessed. We considered three widely-used summary functionals (FAUC-type, magnitude-specific, time-specific) introduced in Section 4.2.2.

We first generated the disease status of each subject using a Bernoulli distribution, $Bernoulli(0.5)$. The true functional markers X were generated over a time interval $\mathcal{T} = [0, 1]$ under four different scenarios. In Scenario 1, we generated $X(t)$ by a Gaussian process with mean functions $\mu^{\overline{D}}(t) = 1$ and $\mu^{D}(t) = 2$ for the non-diseased and diseased subjects, respectively. Likewise, in Scenario 2, X were generated as a Gaussian process, but this time with time-varying mean functions $\mu^{\overline{D}}(t) = t$ and $\mu^{D}(t) = 2t$. In Scenario 3, a Gaussian processes with periodical mean functions $\mu^{\overline{D}}(t) = \sin(\pi t)$ and $\mu^{D}(t) = 3\sin(\pi t)$ were used to generate X. Note that a covariance function $\mathrm{Cov}(X(s), X(t)) = \exp\{-(s-t)^2\}$, $s, t \in \mathcal{T}$ was used for the first three scenarios. In Scenario 4, we generated $X(t) = \sin(2\pi t)$ with probability $2/3$ and $X(t) = \sin(\frac{2}{3}\pi t)$ with probability $1/3$ for the non-diseased, and generated $X(t) = \sin(\pi t)$ with probability 1 for the diseased. In all four scenarios, we obtained error-prone proxy data W by further contaminating the generated X under model

(4.1); the measurement errors ϵ were iid generated from $N(0, 0.1^2)$. Different summary functional types was considered under different scenarios: ϕ_{FAUC} under Scenarios 1 and 2; $\phi_{\text{MAG}(0.5)}$ and ϕ_{MAX} under Scenario 3; and ϕ_{tMAX} under Scenario 4.

For each scenario, we considered the following five study designs to assess the sensitivity of the proposed framework to varying density of observed time points: (20^U) unbalanced design with N_i (and N_j) following a Poisson distribution with mean 20; (40^U) unbalanced design with N_i following a Poisson distribution with mean 40; (20^B) balanced design with $N_i = 20$; (40^B) balanced design with $N_i = 40$; and (60^B) balanced design with $N_i = 60$. Except for the two endpoints (0 and 1), the N_i observation times in all these study designs were randomly drawn from a uniformly distributed grid $\mathcal{T}_{\text{grid}} = \{(u-1)/59,\ u = 2, \ldots, 59\}$ separately for each subject.

We estimated the true markers X using GM kernel estimators (4.2) with a polynomial kernel of degree 2 (Müller, 1984) and an automatically adapted global "plug-in" bandwidth that is asymptotically optimal with respect to the mean integrated square error (MISE) (Gasser et al., 1991). Standard errors were estimated via bootstrap with 1000 resamples, and 95% CIs were computed based on the logit transformation as described in Section 4.3.3.

Table 4.1 presents the simulation results based on 1000 replicates for each scenario with sample size $n = 40$ and 100. Our proposed estimation approach provides virtually unbiased estimates even for a small sample size ($n = 40$) and number of observed time points ($N_i \approx 20$). The bootstrap standard error estimation is generally an accurate estimator of the empirical standard deviation of AUC estimates regardless of the study design and the choice of the summary functional. In all configurations, the empirical coverage probabilities rapidly approach the nominal level 0.95 as the sample size increases.

Next, we considered the semiparametric AUC regression model of the form $\text{AUC}_z(\phi) = g^{-1}(\beta_0 + \beta_1 z)$, where g is a probit or logit link, and examined the fi-

Table 4.1: Simulation results for proposed AUC measures AUC(ϕ): mean of 1000 biases (EmpBias), standard deviation of 1000 AUC estimates(EmpSD), mean of 1000 standard error estimates (EstSE) and proportion of 95% CIs containing the true AUC value (Cov95). D denotes the five study designs for the observed time domain.

			$n = 40$				$n = 100$			
Scenario	True AUC	D	EmpBias	EmpSD	EstSE	Cov95	EmpBias	EmpSD	EstSE	Cov95
1	AUC(ϕ_{FAUC})	(20^U)	0.003	0.072	0.074	0.959	0.002	0.047	0.046	0.947
	$= 0.777$	(40^U)	-0.003	0.071	0.074	0.960	-0.003	0.047	0.047	0.950
		(20^B)	0.001	0.074	0.074	0.962	0.002	0.044	0.046	0.961
		(40^B)	0.002	0.073	0.073	0.962	0.000	0.045	0.046	0.957
		(60^B)	0.002	0.072	0.074	0.964	-0.001	0.048	0.046	0.951
2	AUC(ϕ_{FAUC})	(20^U)	0.003	0.087	0.088	0.964	0.002	0.056	0.055	0.945
	$= 0.648$	(40^U)	-0.004	0.085	0.088	0.959	-0.003	0.056	0.055	0.937
		(20^B)	0.001	0.088	0.088	0.955	0.001	0.052	0.055	0.957
		(40^B)	0.003	0.088	0.088	0.955	-0.001	0.055	0.055	0.955
		(60^B)	0.001	0.086	0.088	0.964	-0.001	0.057	0.055	0.945
3	AUC($\phi_{\text{MAG}(0.5)}$)	(20^U)	-0.001	0.042	0.042	0.967	-0.001	0.027	0.027	0.945
	$= 0.921$	(40^U)	-0.002	0.041	0.042	0.960	-0.003	0.027	0.027	0.950
		(20^B)	-0.001	0.043	0.042	0.958	-0.001	0.026	0.026	0.961
		(40^B)	0.000	0.040	0.042	0.962	-0.001	0.025	0.026	0.967
		(60^B)	0.002	0.042	0.041	0.962	-0.002	0.027	0.027	0.940
3	AUC(ϕ_{MAX})	(20^U)	0.000	0.045	0.044	0.959	0.000	0.029	0.028	0.951
	$= 0.914$	(40^U)	-0.001	0.042	0.044	0.960	-0.002	0.028	0.028	0.950
		(20^B)	0.000	0.045	0.044	0.957	0.000	0.027	0.028	0.962
		(40^B)	0.001	0.043	0.044	0.969	0.000	0.027	0.028	0.966
		(60^B)	0.003	0.044	0.043	0.965	-0.001	0.028	0.028	0.945
4	AUC(ϕ_{tMAX})	(20^U)	0.006	0.100	0.102	0.972	0.003	0.066	0.066	0.955
	$= 0.667$	(40^U)	0.006	0.106	0.102	0.965	0.002	0.066	0.066	0.954
		(20^B)	0.003	0.105	0.102	0.968	0.001	0.069	0.066	0.951
		(40^B)	0.004	0.102	0.103	0.972	0.002	0.068	0.066	0.955
		(60^B)	0.003	0.107	0.102	0.952	0.002	0.067	0.066	0.954

nite sample performance of the regression coefficient (slope) estimate $\widehat{\beta}_1$ obtained by solving the estimated estimating equations (4.10). The disease status of each subject was first generated from $Bernoulli(0.5)$. In the first three scenarios, we consider a binary covariate Z (1 or 0), which is generated from $Bernoulli(0.5)$. In Scenario A, we generated X given $Z = z$ on $\mathcal{T} = [0, 1]$ as a Gaussian process with mean functions $\mu^D(t) = 0.5 + 2z$ and $\mu^{\overline{D}}(t) = z$. In Scenario B, X given $Z = z$ was generated by a Gaussian process with $\mu^D(t) = 0.5+2z+\sin(\pi t)$ and $\mu^{\overline{D}}(t) = z+\sin(\pi t)$. In Scenario C, if non-diseased, $X(t) = \sin(2\pi t)$ with probability $p = \{1+\exp(-1-0.68z)\}^{-1}$ and $X(t) = \sin(\frac{2}{3}\pi t)$ with probability $1 - p$; if diseased, $X(t) = \sin(\pi t)$ with probability 1.

In the next three scenarios, we considered a continuous covariate Z, which we draw from $Uniform(0, 10)$. X given $Z = z$ was generated as a Gaussian process with $\mu^D(t) = 0.5+0.2z$ and $\mu^{\overline{D}}(t) = 0.1z$ in Scenario D, and $\mu^D(t) = 0.5+0.2z+\sin(\pi t)$ and $\mu^{\overline{D}}(t) = 0.1z + \sin(\pi t)$ in Scenario E. In Scenario F, if non-diseased, $X(t) = \sin(2\pi t)$ with probability $p = \{1 + \exp(-1 - 0.068z)\}^{-1}$ and $X(t) = \sin(\frac{2}{3}\pi t)$ with probability $1 - p$; if diseased, $X(t) = \sin(\pi t)$ with probability 1.

Under such settings, the true model is: the probit regression model in Scenarios A, B, D and E; and the logistic regression model in Scenarios C and F. In every scenario involving Gaussian processes, we set $\text{Cov}(X(s), X(t)) = \exp\{-(s - t)^2\}$, $s, t \in \mathcal{T}$. All X's were further contaminated by measurement errors generated from $N(0, 0.1^2)$. The types of summary functionals we considered were: ϕ_{FAUC} under Scenarios 1 and 4; $\phi_{\text{MAG}(0.5)}$ under Scenarios 2 and 5; ϕ_{tMAX} under Scenarios 3 and 6. The same five study designs (20^U)–(60^B) utilized in the above no-covariate case. For continuous covariates, the estimated estimating equations (4.10) utilized pairwise comparisons U_{ij} with covariates less than 5 units apart, that is, $\eta = 5$. GM kernel estimator based on polynomial kernel of degree 2 and automatically adapted bandwidth that optimizes MISE (Müller, 1984, Gasser et al., 1991) was used. Standard errors were

estimated via jackknife.

Simulation results based on 1000 replicates with sample size $n = 200$ and 300 are presented in Table 4.2. Relative biases tend to be slightly larger when data are highly sparse (20^U and 20^B), but rapidly diminish with increasing number of time points. Relative biases in other cases are all smaller than or close to 3%, suggesting that the regression slope estimates obtained from our proposed approach are reasonably unbiased in a finite sample setting. For ϕ_{FAUC} and $\phi_{\text{MAG}(0.5)}$, the estimated standard errors based on the jackknife method generally agree well with the empirical standard deviations. The 95% confidence intervals have coverage probabilities close to the nominal level. For ϕ_{tMAX}, a coverage of 95% or 96% is generally achieved in the continuous covariate case. In the binary covariate case, coverage probabilities are close to 97% with $N_i \leq 40$ and $n = 200$, but they rapidly approach the nominal level as N_i and n increase. In summary, our practical recommendation is to perform semi-parametric regression analysis using functional markers that are, on average, collected on at least 20 time points (preferably 40 time points for time-specific features) to ensure accurate estimation and inference.

Finally, we evaluated and compared the finite-sample performance of the manual and proposed data-driven approaches for selecting η when continuous covariates are involved. We first followed the data generation scheme given by Scenario D above. The primary model is $\text{AUC}_z(\phi_{\text{FAUC}}) = \Phi(\beta_0 + \beta_1 z) = \Phi(0.381 + 0.076z)$, where β_1 describes the effect of one unit increase in z on the AUC between diseased and non-diseased subjects of the same covariate value ($z^D = z^{\overline{D}} = z$). As described in Section 4.4.3, β_1 was estimated using the estimated estimating equations 4.10 derived from a temporary model $\text{AUC}_z(\phi_{\text{FAUC}}) = \Phi\{\beta_0 + \beta_1 z^D + \beta_2(z^D - z^{\overline{D}})\}$, which can accommodate diseased and non-diseased pairs with covariate values within η units apart. The top panel of Table 4.3 reports the mean of 1000 mean squared errors (MSEs) of β_1 based on manually chosen $\eta = 2$ and $\eta = \infty$, and data-driven η that

Table 4.2: Simulation results for regression coefficient (slope) estimates $\widehat{\beta}_1$ from the semiparametric regression model $\text{AUC}_z(\phi) = g^{-1}(\beta_0 + \beta_1 z)$, obtained as the solution to (4.10). $\Phi(\cdot)$ and $l^{-1}(\cdot)$ respectively denote cumulative standard normal distribution function and inverse logit function. Mean of 1000 relative biases (RBias), standard deviation of 1000 AUC estimates (EmpSD), mean of 1000 standard error estimates (EstSE) and proportion of 95% CIs containing the true AUC value (Cov95). D denotes the five study designs for the observed time domain

			$n = 200$				$n = 300$			
Scenario	True Model	D	RBias (%)	EmpSD	EstSE	Cov95	RBias (%)	EmpSD	EstSE	Cov95
A	$\text{AUC}_z(\phi_{\text{FAUC}}) =$	(20^U)	4.471	0.235	0.233	0.949	1.624	0.178	0.187	0.952
z: binary	$\Phi(0.381 + 0.762z)$	(40^U)	2.598	0.234	0.233	0.955	-0.379	0.189	0.186	0.950
		(20^B)	1.936	0.231	0.233	0.952	2.390	0.180	0.186	0.955
		(40^B)	1.659	0.224	0.232	0.958	0.768	0.187	0.186	0.955
		(60^B)	0.906	0.235	0.233	0.948	-0.706	0.181	0.185	0.955
B	$\text{AUC}_z(\phi_{\text{MAG}(0.5)}) =$	(20^U)	2.501	0.233	0.229	0.952	1.583	0.176	0.184	0.957
z: binary	$\Phi(0.354 + 0.707z)$	(40^U)	1.237	0.230	0.229	0.941	-0.268	0.182	0.184	0.961
		(20^B)	1.753	0.222	0.229	0.962	1.408	0.182	0.184	0.943
		(40^B)	1.485	0.222	0.229	0.956	0.901	0.185	0.184	0.958
		(60^B)	0.841	0.223	0.228	0.954	0.079	0.178	0.183	0.958
C	$\text{AUC}_z(\phi_{\text{tMAX}}) =$	(20^U)	5.410	0.550	0.589	0.969	2.784	0.444	0.435	0.956
z: binary	$l^{-1}(1 + 0.68z)$	(40^U)	1.781	0.552	0.622	0.973	2.848	0.426	0.435	0.964
		(20^B)	4.271	0.567	0.618	0.971	5.116	0.443	0.454	0.959
		(40^B)	3.236	0.542	0.586	0.974	2.816	0.444	0.440	0.960
		(60^B)	2.025	0.552	0.571	0.964	1.330	0.430	0.436	0.961
D	$\text{AUC}_z(\phi_{\text{FAUC}}) =$	(20^U)	3.912	0.042	0.041	0.943	2.655	0.032	0.033	0.963
z: continuous	$\Phi(0.381 + 0.076z)$	(40^U)	1.900	0.040	0.041	0.951	3.383	0.032	0.033	0.955
		(20^B)	1.654	0.040	0.041	0.953	1.511	0.033	0.033	0.951
		(40^B)	1.167	0.039	0.041	0.957	1.034	0.034	0.033	0.943
		(60^B)	1.628	0.040	0.041	0.956	-0.078	0.032	0.033	0.957
E	$\text{AUC}_z(\phi_{\text{MAG}(0.5)}) =$	(20^U)	3.221	0.039	0.040	0.954	2.896	0.031	0.032	0.958
z: continuous	$\Phi(0.354 + 0.071z)$	(40^U)	2.379	0.039	0.040	0.958	-0.441	0.032	0.032	0.959
		(20^B)	1.876	0.040	0.040	0.950	1.400	0.031	0.032	0.954
		(40^B)	1.645	0.040	0.040	0.947	1.322	0.032	0.032	0.953
		(60^B)	0.538	0.040	0.040	0.954	0.195	0.032	0.032	0.953
F	$\text{AUC}_z(\phi_{\text{tMAX}}) =$	(20^U)	4.985	0.093	0.093	0.959	2.562	0.073	0.075	0.962
z: continuous	$l^{-1}(1 + 0.068z)$	(40^U)	0.658	0.093	0.094	0.961	2.707	0.076	0.075	0.952
		(20^B)	2.226	0.092	0.093	0.969	-1.248	0.070	0.074	0.965
		(40^B)	-0.245	0.093	0.093	0.957	-2.218	0.074	0.074	0.952
		(60^B)	1.606	0.095	0.094	0.961	-0.082	0.073	0.074	0.961

Table 4.3: Mean of 100×MSEs of β_1 from the primary model $\text{AUC}_z(\phi_{\text{FAUC}}) = \Phi(\beta_0 + \beta_1 z) = \Phi(0.381 + 0.076z)$ computed for 1,000 simulated datasets, given correctly (top-panel) and incorrectly (bottom-panel) specified structure of the temporary model $\text{AUC}_z(\phi_{\text{FAUC}}) = \Phi\{\beta_0 + \beta_1 z^D + \beta_2(z^D - z^{\overline{D}})\}$. η values were chosen either manually ($\eta = 2$ and $\eta = \infty$) or by a data-driven (D-D) approach that minimizes (4.11) of each generated dataset. D denotes the five study designs (20^U)–(60^B) for the observed time domain.

	$n = 200$			$n = 300$		
D	$\eta = 2$	$\eta = \infty$	D-D	$\eta = 2$	$\eta = \infty$	D-D
Correct model structure						
(20^U)	0.181	0.169	0.174	0.103	0.096	0.101
(40^U)	0.163	0.150	0.158	0.108	0.096	0.100
(20^B)	0.167	0.154	0.162	0.112	0.104	0.110
(40^B)	0.157	0.148	0.152	0.116	0.112	0.115
(60^B)	0.159	0.154	0.158	0.104	0.094	0.100
Incorrect model structure						
(20^U)	1.125	0.848	0.214	0.745	0.619	0.120
(40^U)	0.914	0.764	0.192	0.700	0.607	0.123
(20^B)	1.016	0.801	0.202	0.730	0.624	0.123
(40^B)	0.964	0.745	0.189	0.697	0.608	0.131
(60^B)	0.945	0.785	0.188	0.771	0.633	0.131

minimizes (4.11) of each generated dataset. MSEs are largest for $\eta = 2$ due to low efficiency. On the other hand, setting $\eta = \infty$ results in the smallest MSEs as this not only ensures maximum efficiency but also produces small bias under the correct specification of the model structure. The performance of data-driven η falls in between.

We additionally considered the case where the model structure is misspecified. Specifically, we now assume that covariates depend on the disease status and generated them from $N(5, 3)$ and $N(5.1, 3)$ for non-diseased and diseased subjects, respectively. In this case, β_1 from the temporary model no longer accurately represents β_1 from the primary model as the former characterizes the effect of z^D whose expecta-

tion is now different from that of the common covariate z. Thus, the estimator is asymptotically biased, and the corresponding MSEs for pre-fixed $\eta = 2$ and $\eta = \infty$ are much larger than the previous case (see the bottom panel of Table 4.3). On the other hand, our proposed data-driven approach, which fully exploits information from each pooled sample in selecting η, produces robust estimates with much smaller MSEs, suggesting its utility in various practical situations.

4.6 Application to Renal Study

In this section, we apply the proposed method to the renal study described in Section 4.1. Diagnosing kidney obstruction with diuresis renography requires a thorough understanding of renal physiology and MAG3 pharmacokinetics (Taylor and Garcia, 2014). Due to limited resources, however, a vast majority of diuretic renography scans in United States are interpreted by general radiologists who have less than 4 months of training in nuclear medicine, resulting in increased erroneous diagnoses (Taylor et al., 2008c). A common practice to assist radiologists arrive at correct diagnosis is to compute and analyze simple interpretative features of the baseline and post-furosemide renogram curves, such as time to reach half-maximum MAG3 photon count, maximum MAG3 photon count, etc. (Taylor et al., 2008c, Bao et al., 2011). Under such circumstances, the scientific goal of the renal study is three-fold: 1) to evaluate diagnostic utility (AUC) of quantitative features that are currently widely used in practice; 2) to identify and evaluate new features that can be equally or even more important; 3) and identify subpopulation for whom certain features are more useful.

To study these goals, diuresis renography data of 275 kidneys from 145 patients (75 men [52%], 70 women [48%]; mean age, 58 years; SD, 16 years; range, 18-87 years) were randomly selected from the Emory University Hospital's archived database.

All kidneys have complete baseline and post-furosemide renogram curve data. The obstruction status of each kidney was determined based on diuretic renography scan interpretation provided by Dr. Andrew Taylor from Emory University who has more than 20 years of experience in nuclear medicine. 200 kidneys were diagnosed as non-obstructed ($\overline{D}$), and 75 kidneys were diagnosed as obstructed (D) by this method.

Several quantitative features of renogram curves reflective of obstruction severity were considered. For baseline renogram curves, we selected features that characterize the speed of initial MAG3 uptake or the rate of its eventual excretion into the bladder, both of which are known to strongly relate with the obstruction status (Mettler and Guiberteau, 2012). Specifically, we considered time to reach half-maximum MAG3 ($\phi_{\frac{1}{2}\text{tMAX}}$; which is widely used in practice), time to reach maximum MAG3 (ϕ_{tMAX}) and minimum velocity ($\phi^{[1]}_{\text{MIN}}$). For post-furosemide renogram curves, two features that characterize the overall MAG3 intensity were considered: functional area under the curve (ϕ_{FAUC}) and maximum MAG3 (ϕ_{MAX}). All summary functionals were estimated based on the Gasser-Müller kernel estimates (4.2) of the crude ($\nu = 0$) renogram curves and their first derivatives ($\nu = 1$) using a polynomial kernel of degree 2 and 3, respectively (Müller, 1984), and an automatically adapted global "plug-in" bandwidth that optimizes MISE (Gasser et al., 1991). Standard errors were estimated by bootstrap resampling (1,000 samples).

Table 4.4 presents the AUC estimates of the selected summary functions of the baseline and post-furosemide renogram curves. For the baseline renogram, AUC of $\phi_{\frac{1}{2}\text{tMAX}}$ is 0.797 (95% CI: 0.723–0.856), suggesting its moderate diagnostic utility. AUCs of the two newly identified summary functionals, ϕ_{tMAX} and $\phi^{[1]}_{\text{MIN}}$ are 0.855 (95% CI: 0.801–0.897) and 0.812 (95% CI: 0.742–0.866), respectively. Especially, the AUC of ϕ_{tMAX} is statistically significantly higher than that of $\phi_{\frac{1}{2}\text{tMAX}}$ (P-value = 0.045). Taylor et al. (2008c) noted that hospitals can save time and medical costs required to perform a furosemide administration if the baseline renogram alone can

Table 4.4: Estimated AUCs of the summary functionals (SFs) of the baseline and post-furosemide renogram curves. SE: standard error, CI: confidence interval.

SF	AUC (SE)	95% CI
Baseline Renogram		
$\phi_{\frac{1}{2}\text{tMAX}}$	0.797 (0.034)	[0.723, 0.856]
ϕ_{tMAX}	0.855 (0.024)	[0.801, 0.897]
$\phi^{[1]}_{\text{MIN}}$	0.812 (0.031)	[0.742, 0.866]
Post-furosemide Renogram		
ϕ_{FAUC}	0.892 (0.023)	[0.838, 0.930]
ϕ_{MAX}	0.856 (0.027)	[0.794, 0.902]

exclude kidney obstruction in practice. Our findings thus have potential clinical implications for more prompt, accurate and economical detection of kidney obstruction in many practical settings. For the post-furosemide renogram, both ϕ_{FAUC} (AUC: 0.892; 95% CI: 0.838–0.930) and ϕ_{MAX} (AUC: 0.856; 95% CI: 0.794–0.902) have good diagnostic utility for discrimination between obstructed cases and non-obstructed controls.

Next, we conducted a semiparametric regression analysis (4.8) to identify certain age and gender groups for whom a given quantitative feature is more useful. The primary model of interest is $logit\{\theta(\phi)\} = \beta_0 + \beta_1 age + \beta_2 sex$, where age is binary age group (65 years & older vs. younger) and sex denotes gender (male vs. female). We also considered a three-category age group: under 50 (50-), 50-64 and 65 & older (65+); specifically, the model is $logit\{\theta(\phi)\} = \beta_0 + \beta_1 age_1 + \beta_2 age_2 + \beta_3 sex$, where age_1 and age_2 are dummy variables that represent "50- years" and " 65+ years" groups, respectively. The regression parameters were estimated as the solution to the

estimated estimating equations (4.10), and the standard errors were obtained by the jackknife method.

Table 4.5: Estimated AUC odds ratios (OR) of $\phi^{[1]}_{\text{MIN}}$ derived from the baseline renogram. Model 1 included binary age group (65+ years vs. younger) and gender (male vs. females). Model 2 included categorical age group (50- years, 50-64 years and 65+ years) and gender. SF: summary functional, CI: confidence interval.

SF	Covariate	AUC OR (95% CI)	P-value
Model 1: $logit\{\theta(\phi)\} = \beta_0 + \beta_1 age + \beta_2 sex$			
$\phi^{[1]}_{\text{MIN}}$	Age (65+ years/younger)	2.95 (1.10, 7.87)	0.03
	Gender (male/female)	0.66 (0.27, 1.62)	0.36
Model 2: $logit\{\theta(\phi)\} = \beta_0 + \beta_1 age_1 + \beta_2 age_2 + \beta_3 sex$			
$\phi^{[1]}_{\text{MIN}}$	Age 50- years	2.30 (0.74, 7.20)	0.15
	Age 65+ years	4.76 (1.49, 15.21)	0.01
	Age 50-65 years (reference)	-	-
	Gender (male/female)	0.59 (0.23, 1.51)	0.27

Estimated AUC odds ratios of $\phi^{[1]}_{\text{MIN}}$ derived from the baseline renogram curve is are listed in Table 4.5. For the first model that includes binary age group and gender (Model 1), the estimates indicate that the AUC odds of $\phi^{[1]}_{\text{MIN}}$ are 3 times higher for 65+ year-old patients than for younger patients (AUC odds: 2.95; 95% CI: 1.10–7.87), holding gender fixed. The estimates of the second model (Model 2) indicate that AUC odds of $\phi^{[1]}_{\text{MIN}}$ for 65+ year-old patients are 4.8 times higher than those of 50-64 year-old patients (AUC odds: 4.76; 95% CI: 1.49–15.21). These findings suggest that $\phi^{[1]}_{\text{MIN}}$ of the baseline renogram has superior diagnostic utility for detecting kidney obstruction in elderly patients, especially compared to those who are middle-aged (50-65 years).

We also fit the model with a continuous age covariate, but the result is inconclusive at the 5% significant level, perhaps due to no significant difference in AUC between 50-year-old patients and other age groups. The classification ability of other summary functionals listed in Table 4.4 does not depend on patient characteristics.

4.7 Discussion

Functional markers are increasingly being collected in biomedical studies to better understand complex diseases. In many medical practices, their various quantitative features are derived and studied as they represent important interpretative and pathological information, but are often naively relied upon without appropriate scientific justification. As such, we have developed a much-needed framework that can rigorously evaluate quantitative features based on AUC and appropriately guide their selection and application in practice. We adopted a concept of a summary functional that provides mathematical rigor and flexibility in representing a wide class of quantitative features. We proposed a two-stage AUC estimator that appropriately addresses discreteness and measurement error in observed data and established its asymptotic properties. To systematically describe the heterogeneity of AUC of quantitative features in different subpopulations, we proposed a sensible adaptation of a semi-parametric regression model, whose parameters can be estimated and inferred by our estimated estimating equations.

One main contribution of the proposed framework is the provision of a systematic tool to examine new quantitative features that are potentially highly informative, but have not been used in previous clinical settings. For instance, application of our framework to a renal study in Section 4.6 shows that AUCs of time to maximum and functional area under the curve, which are not used by radiologists in practice, are over 0.85. This shows the potential for identifying new quantitative features of

renogram curves that allow better detection of kidney obstruction, and our proposed method provides a new promising way to facilitate and justify such findings.

Although the proposed framework is illustrated by a specific example, its application could be extended to other clinical studies where functional markers are frequently collected. For example, a plasma drug concentration-time curve, which consists of drug concentration in blood plasma densely measured over an active drug exposure time period, is frequently collected in the field of pharmacokinetics. (Craig and Stitzel, 2004). Commonly derived quantitative features from this curve to study the way the body deals with the drugs include: AUC (total drug exposure over time), $C_{\max}$ (the peak plasma concentration of a drug after administration) and $t_{\max}$ (time to reach $C_{\max}$), which can all be captured and consistently estimated using appropriate summary functionals. Therefore, the proposed framework can be readily applied to evaluate the diagnostic utility of these quantitative features and shed useful scientific insight in many drug-related studies.

Chapter 5

A Novel Statistical Approach to Evaluate Functional Markers Without a Gold Standard

5.1 Introduction

Researchers in public health and biomedical fields are often interested in evaluating diagnostic and prognostic accuracy of diagnostic markers (Pepe, 2003). Usefulness of binary marker tests is generally assessed based on its sensitivity (probability of a positive test given disease is present) and specificity (probability of a negative test given disease is absence). The diagnostic accuracy of continuous markers is evaluated by receiver operating characteristic (ROC) analysis. Specifically, the ROC curve plots the estimated sensitivity versus specificity probabilities evaluated at all possible cutoffs. The area under the ROC curve (AUC) is a measure of the average accuracy; an AUC of 1 represents a perfect marker, while AUC of 0.5 represents a worthless marker (e.g., coin flip).

The estimation of aforementioned performance metrics is straightforward when the true disease status or gold standard test is available. For many diseases, however, it is difficult or impossible to establish definitive diagnosis due to complex clinical conditions, or a gold standard test may be too invasive or expensive to administer. To estimate diagnostic accuracy without a gold standard, a latent class modeling approach that treat the true disease status as a latent variable has been proposed (Hui and Walter, 1980, Hui and Zhou, 1998, Collins and Huynh, 2014). Hui and Walter (1980) built a two-component latent class model to estimate sensitivity and specificity of two binary tests, assuming that they are conditionally independent given disease status. Qu et al. (1996) proposed a random effect latent class model with normally distributed random effects to introduce conditional dependence between tests. More recently, Xu and Craig (2009) proposed a probit latent class model that allows a general correlation structure between tests. For continuous markers, a multivariate latent binormal model has been widely used to estimate the ROC curve and AUC without a gold standard (Choi et al., 2006a, Yu et al., 2011, Jafarzadeh et al., 2016).

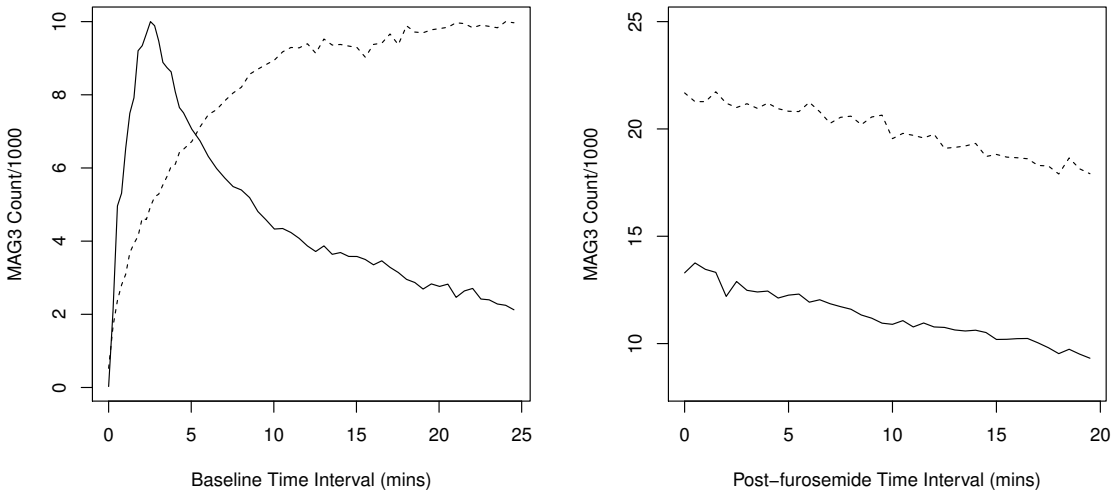
With the advancement in data collection technology, more and more clinically

applicable markers are being collected as functional data (functional markers). Their units of observation are smooth continuous curves (or functions) defined on a continuum (e.g., time or space) but sampled at discrete grids (Ramsay and Silverman, 2005). It is typical in clinical research to use a set of scalar metrics that summarize certain characteristics of a functional marker, such as area under the curve, maximum value and time to reach maximum value, to describe a disease or biological phenomenon; examples of their usage can be found in pharmacokinetics (Craig and Stitzel, 2004), Alzheimer's disease study (Taylor and Garcia, 2014), cardiac safety assessment (Zhou and Sedransk, 2013) and so on. The selection and application of these metrics, however, are mostly based on ad hoc blending of intuition and past practice without rigorous justification. Moreover, enormous information loss may result when aggregating the high-dimensional functional data into a scalar metric.

A more sensible approach to evaluate and utilize functional markers will be to incorporate their inherent dynamic nature that is not fully characterized by some simple scalar metrics. That is, various changing patterns of functional markers that relate to their overall intensity, maximum value, rate of change and many more all constitute valuable information about the curve and may better explain and predict the disease progression. Yan et al. (2017) proposed a functional principal component analysis (FPCA) approach to extract changing patterns of functional markers, and then used them as covariates in a Cox proportional hazards model to make dynamic prediction of disease progression. A similar FPCA-based approach was used by Li and Luo (2017) to incorporate functional markers as covariates in a joint model of longitudinal and time-to-event data. These existing approaches, however, focus entirely on prediction or association, and to our knowledge, no systematic research exists on how to effectively evaluate diagnostic and prognostic accuracy of functional markers under no gold standard.

Our work is motivated by data collected in the renal study. Obstruction to urine

Figure 5.1: Representative baseline and post-furosemide renogram curves for two kidneys. The solid lines are from a kidney interpreted as non-obstructed, and the dotted lines are from a kidney interpreted as obstructed by a nuclear medicine expert.



drainage from a kidney (kidney obstruction) is a serious clinical problem that can lead to irreversible loss of renal function if not properly treated. The diagnosis of kidney obstruction is not straightforward, mostly because there is no consensus in the field on what constitutes the gold standard (Taylor et al., 2008c). In recent years, nuclear medicine renal scans have been widely adopted as a cost-effective and non-invasive approach to detect kidney obstruction. Renal scans start with an intravenous injection of ^{99m}Tc-Mercaptoacetyltriglycine (MAG3) into a kidney to monitor how MAG3 travels down the ureter from the kidney to the bladder. Then, a set of renogram curves is generated by repeatedly measuring the MAG3 photon count inside the kidney over time (Bao et al., 2011). The first renogram curve (called baseline) represents the MAG3 photon counts detected at 59 time points during an initial period of 24 minutes (see the left panel in Figure 5.1). The second renogram curve (called post-furosemide) is obtained at 40 time points during an additional period of 20 minutes after an intravenous injection of furosemide, a potent diuretic (see the right panel in

Figure 5.1).

Current diagnostic practice focuses on several changing patterns of the renogram curves that depict how fast the MAG3 exits a kidney, how long it takes the MAG3 to produce maximum activity, etc., all of which are strongly related to the functional aspects of the kidney (ability to excrete, absorb, etc.). Accordingly, interpretation of renogram curves requires a thorough understanding of renal physiology and MAG3 pharmacokinetics (Taylor and Garcia, 2014). However, a majority of renal scans in United States are interpreted by general radiologists who have less than 4 months of training in those fields, resulting in increased erroneous diagnoses (Taylor et al., 2008c). It is thus of interest to help radiologist improve their diagnosis by evaluating and increasing the diagnostic utility of renogram curves.

In this Chapter, we develop an integrative framework consisting of the following three steps: (1) systematically extract important changing patterns of functional markers (e.g., renogram curves); (2) rigorously evaluate their usefulness for detecting the disease (e.g., kidney obstruction) without a gold standard; and (3) predict the disease status of a future subject (not in the original dataset) using their functional marker data. In the first step, we propose using FPCA to extract the changing patterns of each subject's functional marker. FPCA is an extension of multivariate principal components analysis which examines the variability of a sample of curves and characterizes each of their changing patterns (Rice and Silverman, 1991, Yao et al., 2003, Ramsay and Silverman, 2005, Yao et al., 2005, Yao and Lee, 2006). Advantage of using FPCA is two-fold. Firstly, FPCA provides a systematic approach to capture the changing patterns by first extracting those that explain most variability in the data. Secondly, FPCA provides a natural way of overcoming the intrinsic infinite-dimensionality of functional markers, which presents both conceptual and mathematical difficulties in their use. Specifically, a finite-dimensional representation (FPCA scores) of each subject's functional marker can be obtained by projecting the

function onto a subspace spanned by a finite set of orthonormal eigenfunctions (FPCA basis functions) determined by the observed data. Herein, each FPCA basis function traces a particular changing pattern of a curve, and the corresponding FPCA score describes how strongly the functional marker follows this pattern.

It is possible that some clinical studies collect several functional markers (possibly with different domains) per observation unit. For example, in our renal study, the renogram data for each kidney consist of baseline and post-furosemide renogram curves. Given so-called multivariate functional marker data, we propose an approach based on multivariate functional principal component analysis (MFPCA), recently introduced by Happ and Greven (2018), to extract several joint changing patterns of the multivariate functional markers that may be predictive of a disease outcome. By using MFPCA, the aforementioned advantages of the FPCA approach for univariate functional markers carry over to the context of multivariate functional marker data. That is, the joint changing patterns (MFPCA basis functions) and the corresponding MFPCA scores can be obtained in a systematical manner.

In the second step, we propose using a multivariate binormal latent model to estimate ROC curves and their areas of the obtained FPCA or MFPCA scores in the absence of a gold standard. This amounts to identifying and evaluating changing patterns of a functional marker that are important for understanding and predicting the latent disease status. In the third step, we propose to use marker information gained from the original data to first compute FPCA scores of a new subject and combine these scores in a way that produces an optimal composite test with maximum predictive power under the binormal model (Su and Liu, 1993). Then, given the composite test value, a prediction rule based on the predictive probability of disease can be established.

Although a gold standard is absent, there are situations where information from an imperfect reference test that is highly accurate but subject to small error, such as

diagnostic results from an expert or values of a well-established marker, is available on the same subjects. In the motivating renal study example, there are diagnostic results (scores on the severity of kidney obstruction) on each kidney from three nuclear medicine experts. These experts all had more than 20 years of experience in full-time nuclear medicine, published multiple articles on renal nuclear medicine, were invited to present renal nuclear medicine educational sessions at national radiology meetings, and chaired a panel responsible for the development of guidelines dealing with aspects of renal nuclear medicine. In past marker studies, diagnostic results from imperfect reference tests have been employed to robustify the estimation of diagnostic accuracy of new tests (Albert, 2009, Zhang et al., 2012). In this Chapter, we propose to exploit the imperfect reference test, if available, using functional partial least squares (FPLS) (Delaigle and Hall, 2012) to more efficiently extract changing patterns that are related to the disease mechanism and ultimately achieve superior prediction performance compared to the FPCA approach.

The remainder of the Chapter is organized as follows. In Section 5.2, we briefly review FPCA (and MFPCA) for functional markers and obtain FPCA scores that characterize several changing patterns of functional markers that are potentially important for describing and predicting the underlying disease. Then, we build a multivariate binormal model for estimating the diagnostic accuracy of FPCA scores without a gold standard. A method for predicting the disease status of a new subject based on a composite test is also introduced in this section. In Section 5.3, we propose a FPLS approach for incorporating an imperfect reference standard test to improve the efficiency of prediction. In Section 5.4, we conduct extensive simulation studies to evaluate the finite-sample performance of the proposed estimation, inference and prediction procedures. The application of the proposed framework to a renal study is illustrated in Section 5.5. A discussion follows in Section 5.6.

5.2 A FPCA Approach for Evaluating Functional Markers Without a Gold Standard

5.2.1 FPCA for univariate functional markers

Let X_i $(i = 1, \ldots, n)$ denote the ith subject's functional marker, which is assumed to be an independent random realization of a square-integrable process defined on a compact time interval $\mathcal{T} \subset \mathbb{R}$; that is, $X_i : \mathcal{T} \to \mathbb{R}$ is assumed to be in $L^2(\mathcal{T})$. In this notation, $X_i(t)$ represents a value of the function X_i evaluated at a given time point $t \in \mathcal{T}$ for all i. Let $\mu(t) = E\{X_i(t)\}$ denote the mean function of $X_i(t)$ and $V(s,t) = \text{cov}\{X_i(s), X_i(t)\}$ be its covariance function between two time points $s, t \in \mathcal{T}$. For notational convenience, we will assume that $\mu(t) = 0$ for all $t \in \mathcal{T}$.

Given that $V(s,t)$ is symmetric and non-negative definite, Mercers theorem (Hall and Wang, 2006) implies a spectral decomposition

$$V(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t),$$

where ϕ_k are orthonormal eigenfunctions in $L^2(\mathcal{T})$ corresponding to the eigenvalues λ_k for $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. Note that ϕ_k is orthonormal with respect to the space $L^2(\mathcal{T})$, that is, $\int_{\mathcal{T}} \{\phi_k(t)\}^2 dt = 0$ and $\int_{\mathcal{T}} \phi_r(t)\phi_k(t)dt = 1$ for $r \neq k$. These eigenfunctions form an orthonormal basis of the space $L^2(\mathcal{T})$, and so we may use the Karhunen-Loeve decomposition (Yao et al., 2005) to represent each random function X_i as

$$X_i(t) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \tag{5.1}$$

where ϕ_k is referred to as kth functional principal component (kth FPCA basis function), and $\xi_{ik} = \int_{\mathcal{T}} X_i(t)\phi_k(t)dt$ is the kth FPCA score of X_i. The fact that ϕ_j and ϕ_k are orthogonal for $j \neq k$ implies that the random variables $\xi_{ik}, 1 \leq k < \infty$, are un-

correlated with mean zero and variance λ_k. The eigenvalue λ_k represents the amount of variability in functional marker data explained by the kth FPCA basis function ϕ_k.

In practice, truncated version of the expansion (5.1), that is, optimal K-dimensional approximations to X_i (Yao et al., 2005, Kokoszka and Reimherr, 2017)

$$X_i(t) \approx \sum_{k=1}^{K} \xi_{ik}\phi_k(t), \tag{5.2}$$

are used. The number of components K can be determined based on the proportion of explained variance, Akaike information criterion (AIC) or cross-validation. Details of how to choose K can be found in Rice and Silverman (1991), Yao et al. (2005) and Yan et al. (2017).

5.2.2 FPCA for multivariate functional markers (MFPCA)

Consider multivariate functional marker data where $p \geq 2$ univariate functional markers $X_i^{(1)}, \ldots, X_i^{(p)}$ such that $X_i^{(m)} \in L^2(\mathcal{T}_m)$ $(m = 1, \ldots, p)$ are collected for each subject $i = 1, \ldots, n$. The simplest approach to evaluating multivariate functional markers is to apply univariate FPCA (see Section 5.2.1) to each functional element $X^{(m)}$, $m = 1, \ldots, p$. However, there often exists a non-negligible correlation between the univariate FPCA scores extracted from different functional elements, capturing joint variation among multivariate functional data only indirectly and making interpretation of the results difficult (Happ and Greven, 2018).

To directly address potential covariation among different functional elements, several authors have proposed an approach based on multivariate FPCA (MFPCA) (Berrendero et al., 2011, Chiou et al., 2014, Jacques and Preda, 2014, Happ and Greven, 2018). Among different approaches for MFPCA, we use the approach proposed by Happ and Greven (2018) where a multivariate functional object $\mathbf{X}_i$ that

combines p different functional markers $X_i^{(1)}, \ldots, X_i^{(p)}$ defined on respective time intervals $\mathcal{T}_1, \ldots, \mathcal{T}_p$ is considered, that is,

$$\mathbf{X}_i(\mathbf{t}) = [X_i^{(1)}(t^{(1)}), \ldots, X_i^{(p)}(t^{(p)})]^T \in \mathbb{R}^p,$$

where $\mathbf{t} = [t^{(1)}, \ldots, t^{(p)}]^T \in \mathcal{T}^* = \mathcal{T}_1 \times \cdots \times \mathcal{T}_p$ and $X_i \in \mathcal{H} = L^2(\mathcal{T}_1) \times \cdots \times L^2(\mathcal{T}_p)$. Accordingly, the mean vector $\mu(\mathbf{t}) = E\{\mathbf{X}_i(\mathbf{t})\} = [E\{X_i^{(1)}(t^{(1)})\}, \ldots, E\{X_i^{(p)}(t^{(p)})\}]^T$ and the matrix of covariances $C(\mathbf{s}, \mathbf{t}) = E\{\mathbf{X}_i(\mathbf{s}) \otimes \mathbf{X}_i(\mathbf{t})\}$, $\mathbf{s}, \mathbf{t} \in \mathcal{T}^*$ with elements $C_{lm}(s^{(l)}, t^{(m)}) = \text{cov}\{X^{(l)}(s^{(l)}), X^{(m)}(t^{(m)})\}$, $l, m = 1, \ldots, p$, can be defined (Happ and Greven, 2018). For simplicity of notation, assume that $\mu(\mathbf{t}) = \mathbf{0}$ for all $\mathbf{t} \in \mathcal{T}^*$.

After establishing that $\mathcal{H}$ is a Hilbert space, and $C_{lm}(s^{(l)}, t^{(m)})$ is symmetric and non-negative definite, Happ and Greven (2018) provides a multivariate version of Mercer's theorem that implies a spectral decomposition

$$C_{mm}(s^{(m)}, t^{(m)}) = \sum_{k=1}^{\infty} \upsilon_k \psi_k^{(m)}(s^{(m)}) \psi_k^{(m)}(t^{(m)}), \quad s_m, t_m \in \mathcal{T}_m, \tag{5.3}$$

where functions $\{\psi_k^{(m)}, m = 1, \ldots, p\}$ altogether form an orthonormal eigenfunction $\boldsymbol{\psi}_k = [\psi_k^{(1)}, \ldots, \psi_k^{(p)}]^T \in \mathcal{H}$ corresponding to eigenvalues υ_k for $\upsilon_1 \geq \upsilon_2 \geq \cdots \geq 0$. Note that ψ_k is orthonormal with respect to the space $\mathcal{H}$, that is,
$\sum_{m=1}^p \int_{\mathcal{T}_m} \{\psi_k^{(m)}(t^{(m)})\}^2 dt_m = 0$ and $\sum_{m=1}^p \int_{\mathcal{T}_m} \psi_r^{(m)}(t^{(m)}) \psi_k^{(m)}(t^{(m)}) dt_m = 1$ for $r \neq k$. The decomposition (5.3) provides a basic tool to prove the following multivariate Karhunen-Loeve decomposition (Happ and Greven, 2018):

$$\mathbf{X}_i(\mathbf{t}) = \sum_{k=1}^{\infty} \gamma_{ik} \boldsymbol{\psi}_k(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T}^*, \tag{5.4}$$

where ψ_k is referred to as kth multivariate functional principal component (kth MFPCA basis function), and the uncorrelated random variable
$\gamma_{ik} = \sum_{m=1}^p \int_{\mathcal{T}_m} X_i^{(m)}(t^{(m)}) \psi_k^{(m)}(t^{(m)}) dt_m$ with mean zero and variance υ_k is referred to

as the kth MFPCA score of $\mathbf{X}_i$. The eigenvalue υ_k represents the amount of variability in multivariate functional marker data explained by the MFPCA basis function ψ_k.

Analogous to the univariate case, we adopt a truncated approximation for $\mathbf{X}_i$ given as

$$\mathbf{X}_i(\mathbf{t}) \approx \sum_{k=1}^{K} \gamma_{ik} \boldsymbol{\psi}_k(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T}^*, \tag{5.5}$$

where the appropriate truncation lag K can be chosen based on the proportion of explained variance (Ramsay and Silverman, 2005, Yao et al., 2005).

We end this section by discussing the relationship between univariate and multivariate Karhunen-Loeve decompositions (Happ and Greven, 2018). Consider the univariate Karhunen-Loeve decomposition (5.2) for each mth functional element $X_i^{(m)}$ of $\mathbf{X}_i$: $X_i^{(m)}(t^{(m)}) = \sum_{k=1}^{K_m} \xi_{ik}^{(m)} \phi_m(t^{(m)})$. For $K \leq \sum_{m=1}^{p} K_m = K_+$, let $\mathbf{Z}$ denote a $K_+ \times K_+$ matrix consisting of $K_l \times K_m$ blocks $\mathbf{Z}^{(lm)}$ for entries $Z_{rk}^{(lm)} = \text{cov}(\xi_{ir}^{(l)}, \xi_{ik}^{(m)})$ with $r = 1, \ldots, K_l$, $k = 1, \ldots, K_m$ and $l, m = 1, \ldots p$. Then the positive eigenvalues of $\mathbf{Z}$ correspond to the positive eigenvalues $\upsilon_1 \geq \upsilon_2 \geq \cdots \geq 0$ in (5.3). The eigenfunctions in (5.3) are determined by their univariate counterparts:

$$\psi_k^{(m)}(t^{(m)}) = \sum_{r=1}^{K_m} [\mathbf{c}_k]_r^{(m)} \phi_r^{(m)}(t^{(m)}), \tag{5.6}$$

where $[\mathbf{c}_k]^{(m)} \in \mathbb{R}^{K_m}$ denotes the mth block of an orthonormal eigenvector $\mathbf{c}_k$ corresponding to eigenvalue υ_k of $\mathbf{Z}$. The MFPC scores are given by

$$\gamma_{ik} = \sum_{m=1}^{p} \sum_{r=1}^{K_m} [\mathbf{c}_k]_r^{(m)} \xi_{ir}^{(m)}. \tag{5.7}$$

Proofs for above representations are given in Happ and Greven (2018).

5.2.3 Estimated FPCA scores: a lower dimensional representation of a functional marker

In practice, each functional marker X_i is not observed continuously in time; instead, it is observed at N discrete time points $\{X_i(t_j),\ t_j \in \mathcal{T},\ i = 1, \ldots, n,\ j = 1, \ldots, N\}$. The mean function $\mu(t)$ and convariance function $V(s,t)$ can be consistently estimated with the observed data by $\hat{\mu}(t) = \sum_{i=1}^{n} X_i(t)/n.$ and $\hat{V}(s,t) = \sum_{i=1}^{n}\{X_i(s) - \hat{\mu}(s)\}\{X_i(t) - \hat{\mu}(t)\}/(n-1)$, respectively (Kokoszka and Reimherr, 2017). Then, the estimated eigenvalues $\hat{\lambda}_k$ and estimated FPCA basis functions (eigenfunctions) $\hat{\phi}_k$ are solutions to the functional eigenequation (Castro et al., 1986)

$$\int_{\mathcal{T}} \hat{V}(s,t)\hat{\phi}_k(t)dt = \hat{\lambda}_k\hat{\phi}_k(s),$$

where $\hat{\phi}_k$ are restricted to be orthonormal with respect to the space $L^2(\mathcal{T})$, and the integral can be approximated by a quadrature rule. Finally, a numerical integration can be used to estimate the corresponding scores as

$$\hat{\xi}_{ik} = \int_{\mathcal{T}} \{X_i(t) - \hat{\mu}(t)\}\hat{\phi}_k(t)dt. \tag{5.8}$$

For sparse ($N < 20$) and potentially irregularly sampled functional markers, the principal analysis by conditional estimation (PACE) algorithm can be used to estimate the mean function, covariance function, FPCA basis functions and FPCA scores (Yao et al., 2005). Specifically, the PACE algorithm uses one-dimensional and two-dimensional kernel smoothers to estimate the mean function and covariance function (using off-diagnoal elements), respectively, and performs eigenanalysis on the smoothed covariance to estimate the FPCA basis functions and corresponding eigenvalues. Then, after projecting smoothed covariance on a positive semi-definite surface (Hall et al., 2008), FPCA scores can be estimated based on its conditional

expectation given observed data. Please see Yao et al. (2005) and Liu and Müller (2009) for more information.

$\hat{\phi}_k$ and $\hat{\xi}_{ik}$ are referred to as the *optimal* empirical orthonormal basis functions and coefficients in the sense of minimizing $\sum_{i=1}^{n} ||x_i - \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\phi}_k||^2$, that is, the distance between infinite-dimensional functional objects x_i and their projections onto a K-dimensional space spanned by $\{\hat{\phi}_k, k = 1, \ldots K\}$ (Kokoszka and Reimherr, 2017). Thus, the first K estimated FPCA scores $\hat{\boldsymbol{\xi}}_i = [\hat{\xi}_{i1}, \ldots, \hat{\xi}_{iK}]$ are often used as the lower-dimensional representation of a functional marker X_i, where K is chosen based on the proportion of explained variance, Akaike information criterion (AIC) or cross-validation (Rice and Silverman, 1991, Yao et al., 2005, Yan et al., 2017). Each $\hat{\phi}_k$ traces a unique changing pattern of a curve on $\mathcal{T}$, and $\hat{\xi}_{ik}$ describes how strongly X_i follows this pattern; thus, $\hat{\boldsymbol{\xi}}_i$ can be used as a collection of scalar markers for disease in subsequent ROC analysis.

To estimate the MFPCA, we exploit the relationship between univariate and multivariate FPCA for Karhunen-Loeve decompositions outlined in the last paragraph of Section 5.2.2 (Happ and Greven, 2018). First, for each functional element $X_i^{(m)}$ of a multivariate functional marker $\mathbf{X}_i$, apply the aforementioned method for estimation of univariate FPCA to obtain FPCA basis functions $\hat{\phi}_k^{(m)}$ and the corresponding scores $\hat{\xi}_{ik}$, $i = 1, \ldots, n$, $m = 1, \ldots, p$, $k = 1, \ldots, K_m$ for suitably chosen truncation lags K_m. Second, the block matrix $\mathbf{Z}$ is estimated by $\hat{\mathbf{Z}} = (n-1)^{-1}\boldsymbol{\Xi}^T\boldsymbol{\Xi}$, where $\boldsymbol{\Xi}$ is a $n \times K_+$ matrix with each row consisting of $(\hat{\xi}_{i1}^{(1)}, \ldots, \hat{\xi}_{iK_1}^{(1)}, \ldots, \hat{\xi}_{i1}^{(p)}, \ldots, \hat{\xi}_{iK_p}^{(p)})$. Third, eigenanalysis of $\hat{\mathbf{Z}}$ gives estimated eigenvalues $\hat{\upsilon}_k$ and estimated orthonormal eigenvectors $\hat{\mathbf{c}}_m$. Finally, we plug in above estimates to the equations (5.6) and (5.7) to estimate MFPCA basis functions and the corresponding scores by

$$\hat{\psi}_k^{(m)}(t^{(m)}) = \sum_{r=1}^{K_m} [\hat{\mathbf{c}}_k]_r^{(m)} \hat{\phi}_r^{(m)}(t^{(m)}) \quad \text{and} \quad \hat{\gamma}_{ik} = \sum_{m=1}^{p} \sum_{r=1}^{K_m} [\hat{\mathbf{c}}_k]_r^{(m)} \hat{\xi}_{ir}^{(m)},$$

respectively, for $t_m \in \mathcal{T}_m$ and $k = 1, \ldots, K_+$.

As in the univariate case, the first K estimated MFPCA scores $\hat{\boldsymbol{\gamma}}_i = [\hat{\gamma}_{i1}, \ldots, \hat{\gamma}_{iK}]$ can be used as the lower-dimensional representation of a multivariate functional marker $\mathbf{X}_i$, where the choice of K_m and $K \leq K_+$ can be guided by the proportion of explained variance, Akaike information criterion (AIC) or cross-validation (Happ and Greven, 2018). Herein, each $\hat{\boldsymbol{\psi}}_k = [\hat{\psi}_{i1}^{(1)}, \ldots, \hat{\psi}_{ik}^{(p)}]^T$ can be viewed as a covarying pattern of p curves on respective time domains $\mathcal{T}* = \{\mathcal{T}_1, \ldots \mathcal{T}_p\}$, and $\hat{\gamma}_{ik}$ describes how strongly $\mathbf{X}_i$ follows this pattern; therefore, $\hat{\boldsymbol{\gamma}}_i$ can be used as a collection of scalar markers for disease in subsequent ROC analysis.

5.2.4 FPCA-based ROC analysis without gold standard

In this section, we describe an ROC approach for evaluating the diagnostic accuracy of univariate or multivariate functional markers based on their estimated FPCA scores without a gold standard. For a chosen number of principal components K, let $\hat{\boldsymbol{\zeta}}_i = [\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{iK}]^T$ represent either an estimated univariate FPCA score vector $\hat{\boldsymbol{\xi}}_i$ or an estimated MFPCA score vector $\hat{\boldsymbol{\gamma}}_i$ extracted from the ith subject's univariate or multivariate functional marker, respectively. Since each $\hat{\zeta}_{ik}$ represents a strength of distinct changing pattern of a functional marker, which is often predictive of a disease, a vector $\hat{\boldsymbol{\zeta}}_i$ can be viewed as a set of useable scalar markers whose diagnostic accuracy can be assessed using ROC curves and AUC.

Let D_i denote a binary indicator of disease for the ith subject with $D_i = 1$ if the subject is diseased and $D_i = 0$ otherwise. Because the gold standard test is not available, the disease status D_i is latent. Hui and Walter (1980) showed that a latent model for two binary non-gold standard tests on a single population is not identifiable, but using two or more populations with different prevalences can circumvent the problem of non-identifiability. In our case, we assume that the disease prevalence depends on covariates as an alternative to using multiple populations (Jones et al.,

2009, Yu et al., 2011). Specifically, denoting the covariate vector of the ith subject by $\mathbf{w}_i$, the disease prevalence is assumed to follow a logistic model

$$\Pr(D_i = 1 \mid \mathbf{w}_i) = \pi_i = \frac{\exp(\mathbf{w}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{w}_i^T \boldsymbol{\beta})}, \tag{5.9}$$

where $\boldsymbol{\beta}$ is the vector of coefficients.

Assume that FPCA scores $\hat{\boldsymbol{\zeta}}_i$ is conditionally independent of covariates $\mathbf{w}_i$ given D_i and follow the multivariate binormal latent model given as:

$$D_i \mid \mathbf{w}_i \sim \text{Bernoulli}(\pi_i), \quad \hat{\boldsymbol{\zeta}}_i \mid D_i = d \sim N_K(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d) \quad (d = 0, 1) \tag{5.10}$$

where $\pi_i = P(D_i = 1 \mid \mathbf{w}_i)$ is the disease prevalence provided in (5.9), and $N_K(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$ represents a K-variate normal distribution given true disease status $D = d$ with mean vector $\boldsymbol{\mu}_d = [\mu_{d1}, \ldots, \mu_{dK}]^T$ and covariance matrix $\boldsymbol{\Sigma}_d = \{\sigma_{d,kk'}\}$ $(k, k' = 1, \ldots, K)$.

We will follow the convention that each FPCA score for diseased subjects tends to be greater than those for non-diseased subjects, that is, $\mu_{1k} > \mu_{0k}$. The ROC curve of kth score ζ_{ik} can be expressed by plotting pairs of 1-specificity (x-axis) and sensitivity (y-axis) for given cutoff values $c \in (\infty, \infty)$, namely,

$$\text{ROC}_k(c) = \left[1 - \Phi\left(\frac{c - \mu_{0k}}{\sqrt{\sigma_{0,kk}}}\right), 1 - \Phi\left(\frac{c - \mu_{1k}}{\sqrt{\sigma_{1,kk}}}\right)\right], \tag{5.11}$$

where $\Phi(\cdot)$ denotes a cumulative distribution function (cdf) of the standard normal distribution. The AUC is frequently interpreted as the probability that a score of a randomly chosen subject from the diseased group ($\hat{\zeta}_k^+$) is higher than that of a chosen subject from the nondiseased group ($\hat{\zeta}_k^-$) (Krzanowski and Hand, 2009). Thus, the AUC of kth score $\hat{\zeta}_{ik}$ can be calculated under the binormal model as

$$\text{AUC}_k = \Pr(\hat{\zeta}_{ik}^+ > \hat{\zeta}_{ik}^-) = \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}}\right). \tag{5.12}$$

Note that in real-world applications, we usually do not have *a priori* information on whether FPCA scores $\hat{\zeta}_{ik}$ of diseased subjects are on average higher or lower than those for non-diseased subjects. Thus, if $\mu_{1k} < \mu_{0k}$, we can simply re-define the two diagnostic accuracy measures by taking one minus the coordinates of the $\mathrm{ROC}_k(c)$ in (5.11) and $1 - \mathrm{AUC}_k$ in (5.12).

To assess overall accuracy of a functional marker, it is desirable to combine information carried by multiple scores using their linear combination, where the composite test $\hat{\zeta}_i^* = \mathbf{a}^T \hat{\boldsymbol{\zeta}}_i$. For the weight vector, we choose $\mathbf{a} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_0)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)$, which, under the binormal model (5.10), is known to provide the best linear combination of the scores in the sense that AUC is maximized among all possible linear combinations (Su and Liu, 1993). The corresponding AUC based on such construction ("combined AUC" or "cAUC") is

$$\mathrm{cAUC} = \Phi\Big\{\sqrt{\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}\Big\}. \tag{5.13}$$

5.2.5 Estimation and inference of the ROC model

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denote the collection of all parameters to be estimated. For estimation, we consider the complete data $\mathcal{G} = \{(\mathbf{w}_i, D_i, \hat{\boldsymbol{\zeta}}_i),\ i = 1, \ldots, n\}$, which include covariates, latent disease status and estimated FPCA scores for n subjects. Then the complete-data likelihood function is given by

$$L_c(\boldsymbol{\theta} \mid \mathcal{G}) = \prod_{i=1}^{n} \{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\}^{D_i} \{(1-\pi_i) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}^{1-D_i}, \tag{5.14}$$

where $g(\cdot|\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$ $(d = 0, 1)$ denotes the K-variate normal density with mean $\boldsymbol{\mu}_d$ and covariance $\boldsymbol{\Sigma}_d$ given the true disease status $D_i = d$. Here, we employ expectation-maximization (EM) algorithm (Dempster et al., 1977) to find the maximum likelihood (ML) estimates of $\boldsymbol{\theta}$ by maximizing (5.14). More details regarding the EM algorithm implementation (E-step & M-step) are given in Appendix D.1.

Once the ML estimates of the parameters, $\hat{\boldsymbol{\theta}}$, are obtained, they can replace the corresponding parameters $\boldsymbol{\theta}$ in definitions (5.11), (5.12) and (5.13) to estimate $\mathrm{ROC}_k(c)$, AUC_k and cAUC, respectively. The standard errors of the estimates $\widehat{\mathrm{AUC}}_k$ and $\widehat{\mathrm{cAUC}}$ can be estimated based on the observed information matrix and delta method, and their closed-form formulas are presented in Appendix D.2.

One can use normal approximation to construct confidence intervals (CIs) of AUC_k and cAUC. Since AUC measures are bounded between 0 and 1, adopting a logistic transformation may accelerate the convergence of the corresponding AUC estimate to asymptotic normality, especially when it is close to the boundary. Specifically, let $\widehat{\mathrm{AUC}}$ denote either $\widehat{\mathrm{AUC}}_k$ or $\widehat{\mathrm{cAUC}}$ and $\hat{s}$ denote its estimated standard error. Define $l(x) = \ln\{(x/(1-x)\}$, $l'(x) = dl(x)/dx$, and denote $l^{-1}(\cdot)$ as the inverse function of $l(\cdot)$. Using the delta method, the $100(1-\alpha)\%$ CI for the AUC measures can be constructed as

$$\Big[l^{-1}\big(\tilde{l} - z_{1-\alpha/2}\cdot\tilde{l}'\hat{s}\big), l^{-1}\big(\tilde{l} + z_{1-\alpha/2}\cdot\tilde{l}'\hat{s}\big)\Big],$$

where $\tilde{l} \equiv l(\widehat{\mathrm{AUC}})$, $\tilde{l}' \equiv l'(\widehat{\mathrm{AUC}})$ and $z_{1-\alpha/2}$ denotes the $100(1-\alpha/2)^{\text{th}}$ percentile of $N(0,1)$.

5.2.6 FPCA-based approach to predict disease status of future observations

In this section, we describe an FPCA-based approach to predict disease outcome D_{new} for a new subject (not in the original training dataset) with covariate $\mathbf{w}_{\text{new}}$ and univariate functional marker measurements $\{X_{\text{new}}(t_j),\ t_j \in \mathcal{T},\ j = 1,\ldots,N\}$. Firstly, refer to formula (5.8) to compute the new subject's FPCA scores $\hat{\boldsymbol{\xi}}_{\text{new}} = [\hat{\xi}_{\text{new},1},\ldots,\hat{\xi}_{\text{new},K}]^T$ via numerical integration:

$$\hat{\xi}_{\text{new},k} = \int_{\mathcal{T}} \{X_{\text{new}}(t) - \hat{\mu}(t)\}\hat{\phi}_k(t)dt, \quad k = 1,\ldots,K, \tag{5.15}$$

where $\hat{\mu}(t)$ and $\hat{\phi}_k(t)$ are obtained using the original training dataset following the procedure described in Section 5.2.3. Secondly, to maximize predictability, compute the new subject's composite score $\hat{\xi}^*_{\text{new}} = \hat{\mathbf{a}}^T\hat{\boldsymbol{\xi}}_{\text{new}}$ using optimal weights $\hat{\mathbf{a}} = (\hat{\boldsymbol{\Sigma}}_1 + \hat{\boldsymbol{\Sigma}}_0)^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)$ estimated from the training dataset. Thirdly, estimate the new subject's predictive probability of disease given $\mathbf{w}_{\text{new}}$ and $\hat{\xi}^*_{\text{new}}$ using the Bayes theorem:

$$\begin{aligned}&\widehat{\Pr}(D_{\text{new}} = 1 \mid \mathbf{w}_{\text{new}}, \hat{\xi}^*_{\text{new}}; \hat{\boldsymbol{\theta}}) \\ &= \frac{\hat{\pi}_{\text{new}} g(\hat{\xi}^*_{\text{new}} \mid \hat{\mathbf{a}}^T\hat{\boldsymbol{\mu}}_1, \hat{\mathbf{a}}^T\hat{\boldsymbol{\Sigma}}_1\hat{\mathbf{a}})}{(1-\hat{\pi}_{\text{new}}) g(\hat{\xi}^*_{\text{new}} \mid \hat{\mathbf{a}}^T\hat{\boldsymbol{\mu}}_0, \hat{\mathbf{a}}^T\hat{\boldsymbol{\Sigma}}_0\hat{\mathbf{a}}) + \hat{\pi}_{\text{new}} g(\hat{\xi}^*_{\text{new}} \mid \hat{\mathbf{a}}^T\hat{\boldsymbol{\mu}}_1, \hat{\mathbf{a}}^T\hat{\boldsymbol{\Sigma}}_1\hat{\mathbf{a}})},\end{aligned} \tag{5.16}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$, $\hat{\mathbf{a}}$ and $\hat{\pi}_{\text{new}} = \exp(\mathbf{w}^T_{\text{new}}\hat{\boldsymbol{\beta}})/\{1 + \exp(\mathbf{w}^T_{\text{new}}\hat{\boldsymbol{\beta}})\}$ are obtained from the training dataset, and $g(\cdot|\hat{\mathbf{a}}^T\hat{\boldsymbol{\mu}}_d, \hat{\mathbf{a}}^T\hat{\boldsymbol{\Sigma}}_d\hat{\mathbf{a}})$ $(d = 0, 1)$ denotes a normal density function with mean $\hat{\mathbf{a}}^T\hat{\boldsymbol{\mu}}_d$ and variance $\hat{\mathbf{a}}^T\hat{\boldsymbol{\Sigma}}_d\hat{\mathbf{a}}$. Finally, applying a cutoff (e.g., $v = 0.5$) to the estimated predictive probability yields a method that predicts the latent disease status of this new subject; that is, the subject is predicted to have disease if $\widehat{\Pr}(D_{\text{new}} = 1 \mid \mathbf{w}_{\text{new}}, \hat{\boldsymbol{\xi}}_{\text{new}}; \hat{\boldsymbol{\theta}}) > v$.

The proposed prediction method can be easily extended to a new subject with covariate $\mathbf{w}_{\text{new}}$ and multivariate functional marker measurements $\{X^{(m)}_{\text{new}}(t^{(m)}_j),\ t^{(m)}_j \in \mathcal{T}_m,\ j = 1, \ldots, N_m,\ m = 1, \ldots, p\}$. Let $\hat{\mu}_m(t^{(m)}) = \sum_{i=1}^n X_i(t^{(m)})/n$ and $\hat{\psi}^{(m)}_k(t^{(m)})$ denote the mth element of the estimated mean function and mth element of the kth MFPCA basis function, respectively, obtained using the original dataset. The new subject's MFPCA scores $\hat{\boldsymbol{\gamma}}_{\text{new}} = [\hat{\gamma}_{\text{new},1}, \ldots, \hat{\gamma}_{\text{new},K}]^T$ can be computed via numerical integration as

$$\gamma_{\text{new},k} = \sum_{m=1}^{p} \int_{\mathcal{T}_m} \{X^{(m)}_{\text{new}}(t^{(m)}) - \hat{\mu}_m(t^{(m)})\}\hat{\psi}^{(m)}_k(t^{(m)})dt_m, \quad k = 1, \ldots, K.$$

Then, we can combine these MFPCA scores to produce the new subject's composite score $\hat{\gamma}^*_{\text{new}} = \hat{\mathbf{a}}^T\hat{\boldsymbol{\gamma}}_{\text{new}}$, which can replace $\hat{\xi}^*_{\text{new}}$ in formula (5.16) to calculate the

corresponding predictive probability of disease $\widehat{\Pr}(D_{\text{new}} = 1 \mid \mathbf{w}_{\text{new}}, \hat{\gamma}^*_{\text{new}}; \hat{\boldsymbol{\theta}})$.

5.3 A FPLS Approach to Incorporate Imperfect Reference Test

One main limitation of the FPCA approach is that it only takes into account information about the functional marker X, and therefore maybe suboptimal for predicting the latent disease status D. In particular, the first K FPCA basis functions $\phi_1, \ldots, \phi_K$ contain information only related to the covariance of X, and thus the resulting order of the principal components may not indicate the order of their predictive power; that is, all or some of the most important terms explaining the interaction between D and X might come from later principal components (Delaigle and Hall, 2012). In this section, we propose incorporating diagnostic result from an imperfect reference standard via FPLS to efficiently extract changing patterns of a functional marker that are more relevant for predicting the underlying disease status.

Let Y denote an imperfect reference test score and Y_i $(i = 1, \ldots, n)$ denote its independent random realization for each subject. Assume for notational simplicity that $E(Y_i) = 0$ for all i. Assuming that Y_i is a reasonably accurate marker for the underlying disease status D_i, our approach is to treat Y_i as a surrogate indicator for the underlying disease and directly link it with the corresponding univariate functional marker X_i using the functional linear model given by

$$Y_i = \int_{\mathcal{T}} B(t) X_i(t) dt + \epsilon, \tag{5.17}$$

where B is a function-valued parameter, and ϵ is a mean-zero random error term.

The basic idea of FPLS is to maximize the predictive performance of the model (5.17) by simultaneously decomposing the functional predictor X_i and the scalar

response Y_i in terms of FPLS scores $\nu_{i1}, \nu_{i2}, \ldots$ that have mean zero and uncorrelated with eachother.

$$X_i(t) = \sum_{k=1}^{\infty} \nu_{ik}\rho_k(t) \quad \text{and} \quad Y_i = \sum_{k=1}^{\infty} \nu_{ik}\beta_k + \epsilon,$$

where $\rho_1, \rho_2, \ldots$ are FPLS basis functions (Preda and Saporta, 2005, Febrero-Bande et al., 2017). An iterative algorithm proposed by Delaigle and Hall (2012) that sequentially estimates the FPLS scores and basis functions using observed data $\{(Y_i, X_i(t_j)),$ $t_j \in \mathcal{T},\ j = 1, \ldots, N,\ i = 1, \ldots n\}$ is described in Appendix D.3.

A vector of the first K estimated FPLS scores $[\hat{\nu}_{i1}, \ldots, \hat{\nu}_{iK}]^T$ can then be used as the lower-dimensional representation of X_i. As with the FPCA scores, each FPLS score $\hat{\nu}_{ik}$ represents how strongly X_i follows the pattern traced by $\hat{\rho}_k$. The main advantage of the FPLS approach is that the sequence of FPLS scores $\hat{\nu}_{ik}, \ldots, \hat{\nu}_{iK}$ is naturally sorted in increasing order of importance of explaining the total variance of Y_i (Febrero-Bande et al., 2017), allowing practitioners to efficiently extract features of functional markers that are highly predictive of the latent disease status.

FPLS-based ROC analysis of functional markers can be proceeded as in the FPCA-based framework by replacing $\hat{\boldsymbol{\zeta}}_i$ with the FPLS score vector $\hat{\boldsymbol{\nu}}_i = [\hat{\nu}_{i1}, \ldots, \hat{\nu}_{iK}]^T$ and following the procedures described in Sections 5.2.5 and 5.2.5. For a new subject with covariate $\mathbf{w}_{\text{new}}$ and functional marker measurements $\{X_{\text{new}}(t_j),\ t_j \in \mathcal{T},\ j = 1, \ldots, N\}$, but without the imperfect reference test result, the FPLS scores $\hat{\boldsymbol{\nu}}_{\text{new}} = [\hat{\nu}_{\text{new},1}, \ldots, \hat{\nu}_{\text{new},K}]^T$, the composite test $\hat{\nu}^*_{\text{new}} = \hat{\mathbf{a}}^T\hat{\boldsymbol{\nu}}_{\text{new}}$ and the corresponding predictive probability of disease $\widehat{\Pr}(D_{\text{new}} = 1 \mid \mathbf{w}_{\text{new}}, \hat{\nu}^*_{\text{new}}; \hat{\boldsymbol{\theta}})$ can be computed. See Appendix D.3 for details.

5.4 Simulation Study

In this section, we conduct a simulation study with three settings to evaluate the proposed method. In Setting I, the lower-dimensional representations of (univariate or multivariate) functional markers, i.e., the estimated FPCA scores $\hat{\boldsymbol{\zeta}}_i = [\hat{\zeta}_{i1}, \hat{\zeta}_{i2}, \hat{\zeta}_{i3}]^T$, are directly generated from the binormal model given in (5.10). Different values for the conditional mean ($\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$) and covariance ($\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_0$) parameters are selected in order to investigate the performance of proposed method given different discriminative abilities of the FPCA scores. Specifically, we consider three cases: (1) the first two scores have good ($0.8 < \text{AUC} < 0.9$) discriminative abilities and the last score has moderate ($0.7 < \text{AUC} < 0.8$) discriminative ability; (2) the first score has good discriminative ability, and the last two scores have moderate discriminative abilities; and (3) all three scores have moderate discriminative abilities. Note that in every case, $\boldsymbol{\zeta}_i$, marginally, has mean zero and a diagonal covariance matrix with decreasing variances, ensuring that our data generation scheme is on a par with the classical formulation of FPCA (Ramsay and Silverman, 2005). Specific values of the parameters for the three cases are presented in Appendix D.4.

The disease prevalence is assumed to depend on two binary covariates w_{1i} and w_{2i} according to the logistic model given in (5.9); that is, $\pi_i = \exp(\beta_1 w_{1i} + \beta_2 w_{2i})/\{1 + \exp(\beta_1 w_{1i} + \beta_2 w_{2i})\}$, from which the latent disease status of each subject D_i is generated. We set $\beta_1 = 2$ and $\beta_2 = -2$ so that the two covariates have equal but opposite effect on the disease prevalence.

For each case, we simulate $M = 1000$ samples, each of which consists of $n = 200$ or 400 subjects, and equal numbers of subjects, $n/4$, are assigned to the four subpopulations defined by values of the covariates. The $\widehat{\text{AUC}}_k$ ($k = 1, 2, 3$) and $\widehat{\text{cAUC}}$ are obtained for each sample, and several statistics are calculated to assess the performance of the proposed method by summarizing the 1000 simulated sets. The bias of each AUC estimate is calculated as the mean of 1000 differences between the

estimates and true values (MeanBias), and the empirical standard deviation of the 1000 AUC estimates (EmpSD) is compared to the mean of the 1000 standard error estimates (MeanSE) in order to investigate the validity of our proposed procedure based on the observed information matrix and delta method. The actual coverage rate (Cov95) is calculated as proportion of the 95% CIs (constructed based on logistic transformation) containing the true AUC value. We eliminated the rare cases (less than 5%) when convergence or numerical problems occurred in the EM or Newton-Rahpson algorithm.

The summary of bias, MeanSE and EmpSD and Cov95 are shown in Table 5.1. For all three cases, we see that the AUC estimates have negligible bias. The estimated standard errors agree with the empirical standard deviations, suggesting that the proposed approach based on the observed information matrix and delta method provides fairly accurate standard error estimates. The 95% coverage is close to the nominal level for all cases, implying that the proposed confidence interval formula based on the logistic transformation works well even when the AUC value is close to 1. Better diagnostic accuracy of the FPCA scores tends to increase the speed of convergence of the AUC estimates to the true value and asymptotic normality, although its impact is minimal.

In Setting 2, we aim to assess the learning and prediction performance of our proposed method. Consider univariate functional markers X_i that take the form of

$$X_i(t) = \xi_{i1} \cdot \sqrt{2}\sin(2\pi t) + \xi_{i2} \cdot \sqrt{2}\cos(2\pi t) + \xi_{i3} \cdot \sqrt{2}\sin(4\pi t) + \xi_{i4} \cdot \sqrt{2}\cos(4\pi t),$$

for $t \in \mathcal{T} = [0, 1]$, and are observed at $N = 20$ or 60 discrete time points $\{(t_1, t_2, \ldots, t_N) \in \mathcal{T} : 0 = t_1 < t_2 < \cdots < t_{N-1} < t_N = 1\}$. For the sake of achieving our objective of this simulation study, we generate the true univariate FPCA scores $\boldsymbol{\xi}_i = [\xi_{i1}, \xi_{i2}, \xi_{i3}, \xi_{i4}]^T$ from the binormal model (5.10). The conditional mean and

Table 5.1: Simulation results for Setting 1. The averages of 1000 biases (MeanBias) and standard errors (MeanSE), the standard deviation of the 1000 estimated AUC estimates (EmpSD) and the proportion 95% CIs containing the true AUC estimate in 1000 simulations (Cov95) are presented.

Case	Sample Size	True AUC	MeanBias	EmpSD	MeanSE	Cov95
1	200	$AUC_1 = 0.807$	0.001	0.036	0.037	0.961
		$AUC_2 = 0.807$	0.000	0.037	0.037	0.955
		$AUC_3 = 0.725$	-0.001	0.042	0.042	0.943
		cAUC = 0.984	0.000	0.009	0.008	0.955
	400	$AUC_1 = 0.807$	0.002	0.026	0.026	0.941
		$AUC_2 = 0.807$	0.000	0.026	0.026	0.948
		$AUC_3 = 0.725$	-0.002	0.028	0.029	0.957
		cAUC = 0.984	0.000	0.006	0.005	0.957
2	200	$AUC_1 = 0.807$	0.000	0.046	0.042	0.925
		$AUC_2 = 0.732$	0.001	0.050	0.047	0.945
		$AUC_3 = 0.725$	-0.003	0.051	0.048	0.924
		cAUC = 0.954	0.001	0.024	0.020	0.922
	400	$AUC_1 = 0.807$	0.001	0.031	0.030	0.937
		$AUC_2 = 0.732$	0.000	0.034	0.034	0.953
		$AUC_3 = 0.725$	-0.002	0.034	0.034	0.947
		cAUC = 0.954	0.001	0.015	0.014	0.944
3	200	$AUC_1 = 0.748$	0.003	0.056	0.050	0.930
		$AUC_2 = 0.732$	0.003	0.056	0.051	0.926
		$AUC_3 = 0.736$	-0.001	0.058	0.051	0.916
		cAUC = 0.925	0.008	0.034	0.030	0.910
	400	$AUC_1 = 0.748$	0.001	0.040	0.036	0.922
		$AUC_2 = 0.732$	0.000	0.039	0.037	0.937
		$AUC_3 = 0.736$	-0.001	0.039	0.037	0.940
		cAUC = 0.925	0.002	0.026	0.023	0.923

covariance parameters of the FPCA scores are selected in a way that the first three scores explain approximately 90% of the total variance in generated data. Three cases are considered to assess prediction performance under varying diagnostic accuracy of the first three scores: (1) the first two scores have good discriminative abilities and the third score has moderate discriminative ability; (2) the first score has good discriminative ability, and the next two scores have moderate discriminative abilities; and (3) the first three scores have moderate discriminative abilities. Specific values of the parameters for the three cases (including the parameter values for the fourth score) are presented in Appendix D.4.

There are two binary covariates w_{1i} and w_{2i}, and the latent disease variables D_i are generated according to the logistic model (5.9), with $\beta_1 = 2$ and $\beta_2 = -2$. We generate $M = 1000$ samples, in each of which the training sample size is $n_1 =$ 160 or 320, and the testing sample size is $n_2 = 100$. In each training and testing dataset, equal numbers of subjects ($n_1/4$ and $n_2/4$) are assigned to the four covariate-defined subpopulations. We first perform FPCA to the generated functional markers X_i ($i = 1, \ldots n_1$) in the training dataset, and obtain the estimated mean function ($\hat{\mu}$) and the first three estimated FPCA basis functions ($\hat{\phi}_1$, $\hat{\phi}_2$, $\hat{\phi}_3$), which explain approximately 90% of variability of the data. Since the estimated basis functions are only unique up to their respective signs (Kokoszka and Reimherr, 2017), we use $s_k = \text{sign}\{\int_{\mathcal{T}} \phi_k(t)\hat{\phi}_k(t)dt\}$ to set $\hat{\phi}_k \coloneqq s_k\hat{\phi}_k$ so that the signs are consistent throughout the datasets. Then for each subject in the testing dataset, the first three FPCA scores and the corresponding predictive probability of disease based on the best composite test are obtained using formulas (5.15) and (5.16), respectively. 0.5 is used as a cutoff for the predictive probability to determine diseased/nondiseased diagnosis. As in the first setting, the rare cases where convergence or numerical problems occurred were eliminated.

The average AUC and cAUC estimates obtained using the first three estimated

Table 5.2: Simulation results for Setting 2. The averages (over 1000 simulations) of the AUC and cAUC estimates of the first three estimated FPCA scores in the training dataset of size $n_1 = 160$ and 320, and the percentages correctly classified (PCC) in the testing data of size $n_2 = 100$ are presented.

AUC_y	Training Sample Size (n_1)	Number of Time Points	Average Estimated AUC in Training Data $(\widehat{AUC_1}, \widehat{AUC_2}, \widehat{AUC_3})$, $\widehat{cAUC}$	Average PCC in Testing Data
1	160	20	(0.807, 0.827, 0.669), 0.990	94.2%
		60	(0.806, 0.812, 0.681), 0.987	93.6%
	320	20	(0.806, 0.831, 0.665), 0.989	94.5%
		60	(0.806, 0.811, 0.681), 0.985	93.9%
2	160	20	(0.806, 0.762, 0.675), 0.977	91.8%
		60	(0.805, 0.738, 0.686), 0.972	91.1%
	320	20	(0.806, 0.766, 0.670), 0.970	92.2%
		60	(0.806, 0.744, 0.688), 0.964	91.4%
3	160	20	(0.746, 0.736, 0.734), 0.939	86.9%
		60	(0.748, 0.729, 0.734), 0.937	86.6%
	320	20	(0.748, 0.743, 0.733), 0.935	88.3%
		60	(0.748, 0.736, 0.732), 0.932	88.0%

FPCA scores in the training dataset and the corresponding average percentage of correct classification (PCC) in the testing dataset over 1000 replications are presented in Table 5.2. The results show that when one or two scores have good discriminative abilities, the proposed method can achieve very good prediction accuracy; the average PCCs in the testing data are above 90% for the first two cases. Furthermore, our proposed method is found capable of producing satisfactory prediction accuracy (PCCs $>$ 85%) even when all three scores have only moderate diagnostic accuracy. The prediction performance generally improves as the training sample size (n_1) increases. The average PCC values slightly decrease as N increases implying that the cAUC tends to be overestimated with relatively small N. Overall, the trend is fairly consistent across different n_1 and N combinations, lending support to the robustness of the proposed prediction method to the structure and size of the training dataset.

Setting 3 is identical in design to Setting 2, but the goal here is to illustrate the advantages of the FPLS approach over the FPCA approach provided that the imperfect reference test results are available in the training dataset. For this purpose, we again consider the third case of Setting 2 where the first three FPCA scores, which explain about 90% of variability in the original functional marker data, have only moderate discriminative abilities, but the fourth FPCA score, which is not chosen in the subsequent ROC analysis, has good diagnostic accuracy. This corresponds to a situation where information about the latent disease status moves further away in the sequence of principal components. Imperfect reference tests results of training subjects are generated under $N(\mu_{y,1}, \sigma^2_{y,1})$ and $N(\mu_{y,0}, \sigma^2_{y,0})$ distributions for the diseased and nondiseased, respectively, and are used in the iterative algorithm (Delaigle and Hall, 2012) presented in Appendix D.3 to estimate the first three FPLS scores and the corresponding AUC estimates. Then, based on the procedure outlined in Appendix C, we extract and combine the FPLS scores of the functional markers in the testing dataset to predict their latent disease status. We also investigate the sensitivity of

the proposed approach to different values of the diagnostic accuracy of the imperfect reference test by varying $\mu_{y,1}$ and $\mu_{y,0}$. Values of the parameters used for generating the imperfect reference tests are presented in Appendix D.4.

Table 5.3: Simulation results for Setting 3. The averages (over 1000 simulations) of the AUC and cAUC estimates of the first three estimated FPLS scores in the training dataset of size $n_1 = 160$ and 320, and the percentages correctly classified (PCC) in the testing data of size $n_2 = 100$ are presented.

AUC_y	Training Sample Size (n_1)	Number of Time Points	Average Estimated AUC in Training Data ($\widehat{AUC_1}, \widehat{AUC_2}, \widehat{AUC_3}$), $\widehat{cAUC}$	Average PCC in Testing Data
0.901	160	20	(0.883, 0.820, 0.638), 0.996	94.9%
		60	(0.883, 0.820, 0.638), 0.996	94.9%
	320	20	(0.881, 0.828, 0.641), 0.997	95.5%
		60	(0.881, 0.828, 0.641), 0.997	95.5%
0.802	160	20	(0.882, 0.806, 0.633), 0.995	94.6%
		60	(0.882, 0.806, 0.633), 0.995	94.6%
	320	20	(0.881, 0.821, 0.637), 0.996	95.4%
		60	(0.881, 0.821, 0.637), 0.996	95.4%
0.700	160	20	(0.876, 0.774, 0.625), 0.989	93.6%
		60	(0.876, 0.774, 0.625), 0.989	93.6%
	320	20	(0.881, 0.800, 0.630), 0.993	94.9%
		60	(0.881, 0.800, 0.630), 0.993	94.9%
0.600	160	20	(0.807, 0.705, 0.612), 0.969	90.8%
		60	(0.807, 0.705, 0.612), 0.969	90.8%
	320	20	(0.851, 0.735, 0.619), 0.978	93.0%
		60	(0.851, 0.735, 0.619), 0.978	93.0%

The results in Table 5.3 show that by incorporating an imperfect reference test via FPLS, we can efficiently extract a composite test with excellent diagnostic accuracy (cAUC > 0.95) and accordingly achieve outstanding prediction performance (PCC $> 90\%$). This demonstrates a clear advantage over the FPCA approach, which only takes into account information about the variability in functional markers and exhibits

suboptimal prediction performance below 90% (see the last four rows of Table 5.2). Also, the first one or two AUC estimates are higher, while the third AUC estimates are lower than those of the FPCA approach. This finding suggests that most information about the latent disease status is concentrated in the first few FPLS scores, allowing a parsimonious interpretation of functional markers. Regarding the sensitivity analysis, we see that decreasing AUC of the imperfect reference test (AUC_y) tends to decrease the cAUC and prediction accuracy as expected. But the aforementioned benefits of using FPLS basis remain valid even when AUC_y is as low as 0.6, suggesting that the proposed FPLS approach is applicable over a wide range of scientifically reasonable values of the diagnostic accuracy of the imperfect reference test.

5.5 Application to Renal Study

In this section, we apply the proposed method to the renal study data described in Section 5.1. A total of 145 patients (75 men [52%], 70 women [48%]; mean age, 58 years; SD, 16 years; range, 18-87 years), that is, 290 kidneys, with suspected renal obstruction were enrolled in the study. Only 280 kidneys (138 left kidneys and 142 right kidneys) had complete baseline and post-furosemide renogram curve data, and were randomly divided into a training set and a testing set with sizes 230 and 50, respectively. Two covariates were considered in our models: gender (binary) and age (continuous). Furthermore, diagnoses from three nuclear medicine experts were available on the same kidneys. Each expert rated a kidney on a scale from -1.0 to +1.0. Scores approaching -1.0 indicate greater confidence in the absence of obstruction, and scores closer to 1.0 denote greater confidence in the diagnosis of obstruction. The consensus interpretation (expert consensus rating) of the three experts was determined by majority vote unless there was substantial disagreement and was considered as the imperfect reference test. The goal of our data analysis is twofold: 1) to evaluate the

diagnostic accuracy of renogram curves for renal obstruction using training data; and 2) to utilize this information to predict obstruction status of kidneys in the testing dataset.

Taylor et al. (2008c) noted that hospitals can save time and medical costs required to perform a furosemide administration if the baseline renogram alone can exclude renal obstruction in practice. Therefore, we are first interested in assessing the diagnostic accuracy of the baseline renogram curve and predicting obstruction status of a kidney based upon it. The mean function $\mu(t)$, eigenvalue λ_k and FPCA basis function $\phi_k(t)$ of the baseline renogram curves in the training dataset were estimated by the procedure described in Section 5.2.3, and the resulting estimates were used in equation (5.8) to obtain the corresponding univariate FPCA scores $\hat{\xi}_{ik}$ for each subject. The first two scores $\hat{\boldsymbol{\xi}}_i = [\hat{\xi}_{i1}, \hat{\xi}_{i2}]^T$, which explain 95% of variability in the data, were then used in the subsequent ROC analysis. Specifically, we ran the EM algorithm (see Appendix D.1) using the first two FPCA scores and two covariates (age and gender), obtained the MLEs, and computed the corresponding AUCs and their 95% CIs.

Figure 5.2: Three plots related to the ROC and predictive analyses of the first two FPCA scores extracted from the baseline renogram curves: (a) the first two estimated FPCA basis functions; (b) the fitted mean curves by obstruction status; and (c) the predictive probabilities of renal obstruction cross-tabulated against the corresponding expert consensus ratings.

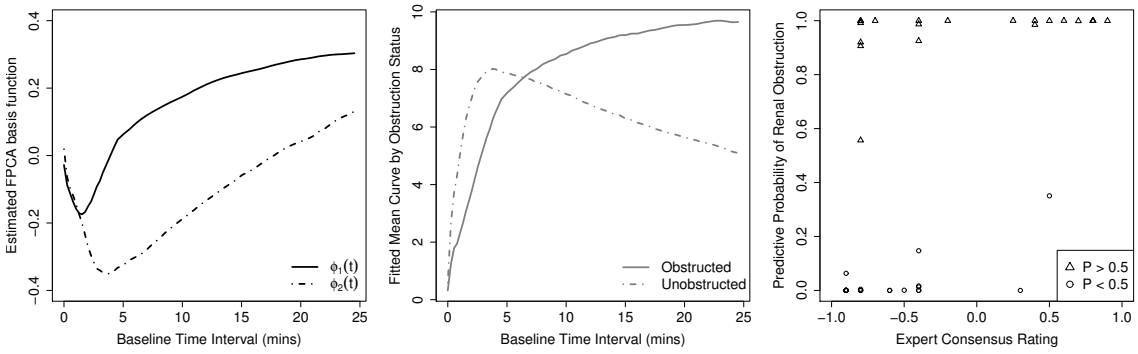


Table 5.4: Estimated conditional means $(\hat{\mu}_{1k}, \hat{\mu}_{0k})$, AUCs and combined AUC (and their 95% CIs) of the first two: (1) FPCA scores extracted from baseline renogram curves; (2) FPLS scores extracted from baseline renogram curves using expert consensus ratings; and (3) MFPCA scores extracted from both baseline and post-furosemide renogram curves.

Renogram	Method	Score	$(\hat{\mu}_{1k}, \hat{\mu}_{0k})$	AUC	95% CI of AUC
Baseline	FPCA	$\hat{\xi}_1$	$(7.61, -5.39)$	$\text{AUC}_1 = 0.925$	$[0.813, 0.972]$
		$\hat{\xi}_2$	$(2.73, -1.93)$	$\text{AUC}_2 = 0.832$	$[0.772, 0.878]$
		$0.37\hat{\xi}_1 + 0.63\hat{\xi}_2$	–	$\text{cAUC} = 0.997$	$[0.991, 0.999]$
Baseline	FPLS	$\hat{\nu}_1$	$(8.07, -5.75)$	$\text{AUC}_1 = 0.960$	$[0.881, 0.987]$
		$\hat{\nu}_2$	$(2.34, -1.67)$	$\text{AUC}_2 = 0.779$	$[0.690, 0.849]$
		$0.40\hat{\nu}_1 + 0.49\hat{\nu}_2$	–	$\text{cAUC} = 0.997$	$[0.991, 0.999]$
Both	MFPCA	$\hat{\gamma}_1$	$(34.09, -25.44)$	$\text{AUC}_1 = 0.924$	$[0.815, 0.971]$
		$\hat{\gamma}_2$	$(2.87, -2.13)$	$\text{AUC}_1 = 0.672$	$[0.544, 0.779]$
		$0.05\hat{\gamma}_1 + 0.11\hat{\gamma}_2$	–	$\text{cAUC} = 0.965$	$[0.926, 0.984]$

Each of the first two FPCA basis functions $(\hat{\phi}_1, \hat{\phi}_2)$ shown in the first figure of Figure 5.2 represents a particular changing pattern, and the corresponding FPCA score $\hat{\xi}_{ik}$ $(k = 1, 2)$ describes how strongly the baseline renogram curve from subject i follow this pattern. The AUC of respective FPCA scores, that is, the diagnostic accuracy of respective changing patterns of the baseline renogram curve, is presented in Table 5.4. We see that the first FPCA score has excellent AUC ($\text{AUC}_1 = 0.925$, 95% CI $= [0.813, 0.972]$) with larger mean for the obstructed ($\hat{\mu}_{11} = 7.61$ vs. $\hat{\mu}_{01} = -5.39$), suggesting that a relatively low MAG3 count during the first five minutes of the scan, followed by its high and increasing trend in the later period, is highly predictive of renal obstruction. The second FPCA score, which has good AUC ($\text{AUC}_2 = 0.832$, 95% CI $= [0.772, 0.878]$) with larger mean for the obstructed ($\hat{\mu}_{11} = 2.73$ vs. $\hat{\mu}_{01} = -1.93$), provides a similar story. The only difference is in the prolonged period (first 13–15 minutes) of relatively low MAG3 counts being predictive of renal obstruction.

To better understand how these two changing patterns characterize the base-

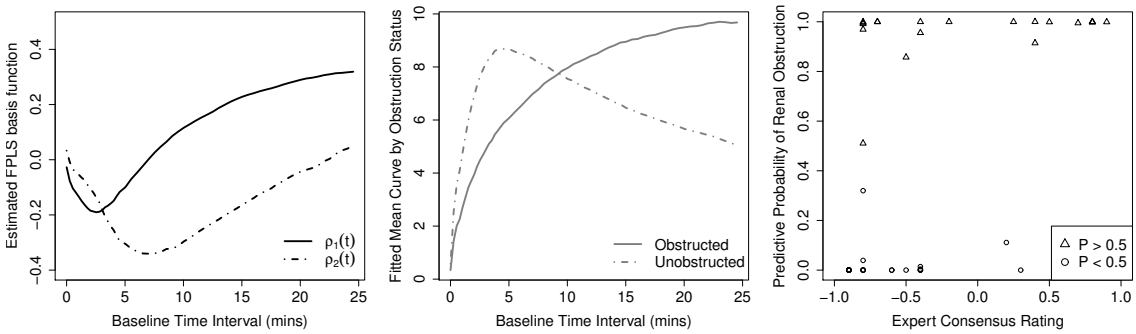
line renogram curve by obstruction status, we plotted the "fitted mean curve" for each obstructed and unobstructed kidney, that is, $\hat{X}_{obs}(t) = \hat{\mu}_{11}\hat{\phi}_1(t) + \hat{\mu}_{12}\hat{\phi}_2(t)$ and $\hat{X}_{un}(t) = \hat{\mu}_{01}\hat{\phi}_1(t) + \hat{\mu}_{02}\hat{\phi}_2(t)$, respectively. The second figure of Figure 5.2 presents the fitted mean curves given each obstruction status. We see that the curve for the obstructed kidney gradually increases over the entire scan period. On the other hand, the curve for the nonobstructed kidney is characterized by a quick uptake of MAG3 followed by its slow drainage over time. This coincides with known knowledge about renography interpretations (compare with the curve patterns in the left panel of Figure 5.1) and is a clear evidence that our proposed framework can extract changing patterns of baseline renogram curves related to the obstruction mechanism.

The optimal composite test performs the best compared to the individual scores (cAUC = 0.997; 95% CI = $[0.991, 0.999]$), and can be used to predict the obstruction status of the kidneys in the testing dataset. We first assigned the FPCA scores $\hat{\boldsymbol{\xi}}_{\text{new}} = [\hat{\xi}_{\text{new},1}, \hat{\xi}_{\text{new},2}]^T$ for each testing kidney using formula (5.15) and derived the composite score $\hat{\mathbf{a}}\hat{\boldsymbol{\xi}}_{\text{new}}$ using the optimal weight $\hat{\mathbf{a}} = [0.36, 0.63]^T$ estimated from the training dataset. Then, the corresponding predictive probability of obstruction was computed by formula (5.16).

It is not straightforward to evaluate the performance of the proposed prediction method on the testing dataset due to the fact that the true obstruction status (gold standard) is unknown. As an alternative, we cross-tabulated the predictive probabilities against the expert consensus ratings, and the resulting plot is shown in the third figure of Figure 5.2. We can see that the predicted obstruction status generally agrees well with the expert ratings; that is, kidneys with higher expert ratings tend to have higher predictive probabilities. However, there are handful of kidneys who have negative expert ratings but are predicted to be obstructed (predictive probability greater than 0.5), because their baseline renogram curves are close in shape to that of obstructed kidneys. This is mostly due to the fact that experts have also taken

into account their post-furosemide renogram curves to diagnose renal obstruction as their baseline renogram alone could not exclude the disease.

Figure 5.3: Three plots related to the ROC and predictive analyses of the first two FPLS scores extracted from the baseline renogram curves: (a) the first two estimated FPLS basis functions; (b) the fitted mean curves by obstruction status; and (c) the predictive probabilities of renal obstruction in the testing dataset cross-tabulated against the corresponding expert consensus ratings.

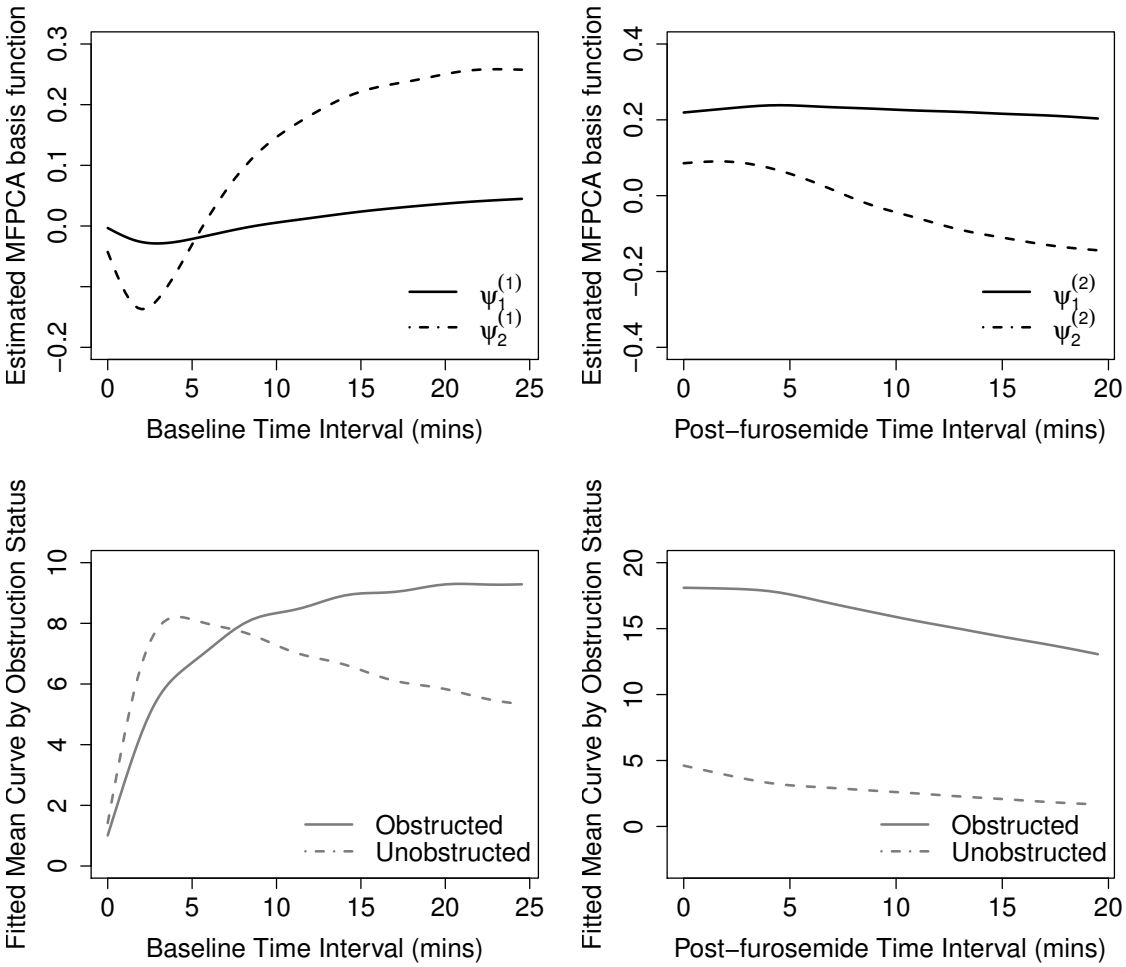


To better extract features of the baseline renogram curve that are relevant to predicting renal obstruction, a FPLS approach using expert consensus ratings as an imperfect reference test was applied to the training dataset. Specifically, we considered a functional linear model (5.17) with expert consensus as the response variable and baseline renogram curve as the explanatory function and ran the iterative algorithm presented in Appendix D.3 to extract the FPLS basis functions and the corresponding FPLS scores.

The first two FPLS basis functions, $\hat{\rho}_1(t)$ and $\hat{\rho}_2(t)$, are presented in the first figure of Figure 5.3, and the resulting conditional mean and AUC estimates of the first two FPLS scores are reported in Table 5.4. The changing patterns of the first two FPLS scores are similar to those of the first two FPCA functions discussed above. Accordingly, the AUC estimates of the first two FPLS scores ($\text{AUC}_2 = 0.960$, 95% CI $= [0.872, 0.988]$; $\text{AUC}_1 = 0.780$, 95% CI $= [0.641, 0.876]$) are close to those of the first two FPCA scores, though the information about the obstruction status is more

concentrated in the first FPLS component. This illustrates that the FPLS approach can capture relevant information with fewer terms, thus allowing a parsimonious interpretation of functional markers. The fitted mean curves are consistent with the typical patterns of obstructed and unobstructed kidneys (see the second figure of Figure 5.3). The predictive probability of obstruction based on the composite FPLS test (cAUC = 0.997, 95% CI = $[0.993, 0.999]$) was obtained for each testing kidney and is plotted against the expert consensus ratings in the third figure of Figure 5.3.

Figure 5.4: Plots related to the ROC analysis of the first two MFPCA scores jointly extracted from the baseline and post-furosemide renogram curves. First row: the first two estimated MFPCA basis functions; Second row: the fitted mean curves by obstruction status.



In practice, a baseline renogram curve often alone cannot exclude renal obstruction, requiring a joint analysis of both baseline and post-furosemide renogram curves for accurate diagnosis. We thus treated the renogram data as multivariate functional markers $\mathbf{X}(\mathbf{t})$, where the baseline and post-furosemide renogram curves constitute the first element $X^{(1)}$ and second element $X^{(2)}$, respectively. Then, MFPCA was used to extract the covarying patterns of both renogram curves potentially relevant for diagnosing renal obstruction. Specifically, we obtained the estimated MFPCA basis functions $\hat{\boldsymbol{\psi}}_k = [\hat{\psi}_k^{(1)}, \hat{\psi}_k^{(2)}]^T$ and scores $\hat{\boldsymbol{\gamma}}_i = [\hat{\gamma}_{i1}, \hat{\gamma}_{i2}]^T$ for each training kidney based on the univariate FPCA expansion of each of the renogram curves. We then selected the first two components that explain 98% of variability in the data. The first two MFPCA basis functions are shown in the first row of Figure 5.4, and the AUC estimates of the corresponding MFPCA scores are listed in Table 5.4. We can see that the first MFPCA score has excellent diagnostic accuracy ($\text{AUC}_1 = 0.924$, 95% CI = $[0.815, 0.971]$) with larger mean for the obstructed ($\hat{\mu}_{11} = 34.09$ vs. $\hat{\mu}_{01} = -25.44$). This implies that the following changing pattern of each of the renogram curves is highly predictive of renal obstruction: 1) a relatively low MAG3 count during the first ten minutes of the baseline renogram, followed by its high and increasing trend in the later period; and 2) an elevated MAG3 count over the entire period of the post-furosemide renogram.

We further plotted the fitted mean curves of the baseline and post-furosemide renogram curves given each obstruction status, $\hat{\mathbf{X}}_{\text{obs}}(\mathbf{t}) = \hat{\mu}_{11}\hat{\boldsymbol{\psi}}_1(\mathbf{t}) + \hat{\mu}_{12}\hat{\boldsymbol{\psi}}_2(\mathbf{t})$ and $\hat{\mathbf{X}}_{\text{un}}(\mathbf{t}) = \hat{\mu}_{21}\hat{\boldsymbol{\psi}}_1(\mathbf{t}) + \hat{\mu}_{22}\hat{\boldsymbol{\psi}}_2(\mathbf{t})$, to better understand how their changing patterns jointly characterize obstructed and unobstructed kidneys (see the second row of Figure 5.4). The fitted mean curves of the baseline renogram closely resemble those of the univariate approaches (FPCA and FPLS). For the post-furosemide renogram, we see that the fitted mean curve of obstructed kidneys maintains a consistently high level; this perfectly agrees with medical knowledge about renal obstruction such that

obstructed kidneys suffer from improper drainage for a prolonged period of time (compare with the curve patterns in the right panel of Figure 5.1).

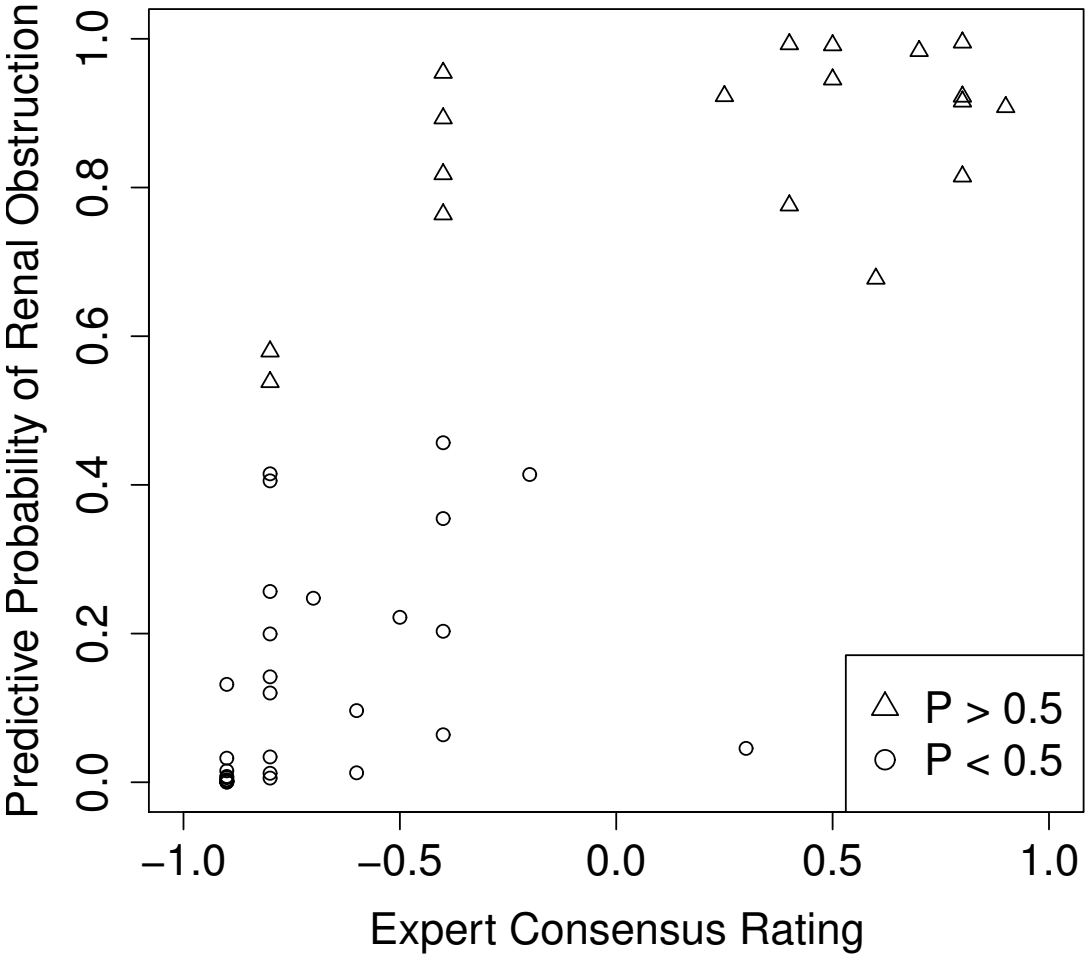


Figure 5.5: The predictive probabilities of renal obstruction in the testing dataset based on the first two MFPCA scores. They are cross-tabulated against the corresponding expert consensus ratings.

To predict obstruction status of testing kidneys by using information from both baseline and post-furosemide renogram curves, we considered an optimal composite test based on combination of the first two MFPCA scores with weights $\hat{\mathbf{a}} = [0.05, 0.11]^T$. The use of this composite test is justified by the fact that it has excel-

lent discriminative ability (cAUC = 0.965, 95% CI = $[0.926, 0.984]$) in the training dataset. The predictive probability of disease for each testing kidney was computed based on the newly derived composite score $\hat{\mathbf{a}}\hat{\boldsymbol{\gamma}}_{\text{new}}$ and two covariate values. In the absence of a gold standard, the predictive probabilities were cross-tabulated against the corresponding expert consensus for evaluation, and the resulting plot is shown in the third row of Figure 5.5. We can see that the predictive probabilities agree better with the expert consensus compared to the univariate approaches. Specifically, the predictive probabilites of kidneys that were in the upper-left corner (low expert rating but high predictive probability) in the previous analyses are now closer to 0, suggesting that the MFPCA approach overcomes the uncertainty in diagnosis posed by inconclusive baseline renogram curves and provides a firm basis for closely replicating experts' opinions in practice by incorporating both renograms for prediction.

5.6 Discussion

In this Chapter, we focused on the dynamic, interpretative changing patterns of functional markers that are potentially useful for understanding the disease progression and develop a statistical framework for rigorously evaluating their diagnostic and prognostic utility without a gold standard. Specifically, we employed a FPCA approach to capture several changing patterns of functional markers in a systematic and parsimonious manner. For multivariate functional marker data, we proposed to utilize a MFPCA approach to characterize their joint changing patterns. Once the changing patterns are extracted, ROC analysis can be performed based on a multivariate latent binormal model in which the unknown true disease status is treated as a latent variable. We further extended the framework to allow prediction of a new subject's disease status based on an optimal composite test. If results from an imperfect reference test are available, we proposed utilizing a FLPS approach to ex-

ploit this information and achieve superior prediction performance compared to the FPCA approach. The ROC performance metrics can be estimated via EM algorithm, and their standard errors are can be computed based on the observed information matrix.

Chapter 6

Future Research

In Chapter 2, we have introduced a set of indices (ODI, OCP and RAUOCPC) that quantifies agreement among multiple raters by expressing the distance (RM-SPD) among their clinical measurements. In future work, we plan to develop a semi-parametric method to describe the distribution of the distance over time and the influence of covariates on this distribution. Specifically, let Y_{i1t} and Y_{i2t} denote measurements of the ith subject ($i = 1, \ldots, n$) at time t ($t = 1, \ldots, T_i$) from two raters. For a given time t, the distance between the paired measurements can be characterized by the absolute error $D_{it} = |Y_{i1t} - Y_{i2t}|$, which reflects the error of one rater with respect to the other. Using this characterization, Lin (2000) proposed to quantify agreement between the two methods using TDI, which is defined as the solution to $\tau = P(D_{it} < \text{TDI}_\tau)$ given $0 < \tau < 1$. The smaller the TDI_τ value, the better the agreement between the two methods. An appealing feature of TDI is its interpretation tied to the original measurement unit.

The extension of TDI to a longitudinal study has not been done and will be the focus of future research. We will propose to longitudinally model the TDI with respect to baseline (e.g., gender, weight, site, etc.) and time-dependent covariates (e.g, patients CD4 count over time). Define the conditional TDI_τ given $\mathbf{x}_{it}$ at time t as $\text{TDI}_\tau(t \mid \mathbf{x}_{it}) = \inf\{d : P(D_{it} \leq d \mid \mathbf{x}_{it}) \geq \tau\}$. We will consider a semi-parametric marginal regression model

$$\text{TDI}_\tau(t \mid \mathbf{x}_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta}_\tau, \tag{6.1}$$

where $\boldsymbol{\beta}_\tau$ is a vector of unknown regression coefficients (function of τ) that allows inhomogeneous covariate effects on TDI across different τ values. The model (6.1) marginally specifies the relationship between TDI_τ and covariates at time t. A positive coefficient indicates that an increase in a corresponding covariate impairs agreement.

We will estimate $\boldsymbol{\beta}_\tau$ by solving a system of estimating equations of the kind

$$n^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T_i}\mathbf{x}_{it}(\tau - I(D_{it} - \mathbf{x}_{it}^T\boldsymbol{\beta}_\tau \leq 0)) = \mathbf{0}, \tag{6.2}$$

which assumes the working independent correlation matrix but yields consistent estimates (He et al., 2003, Yin and Cai, 2005). We can extend (6.2) to explicitly account for correlation to enhance efficiency (Jung, 1996, Tang and Leng, 2011, Leng and Zang, 2014). We will use bootstrap procedures to make inference; for example, we can generate a bootstrap by randomly selecting n subjects with replacement.

In Chapter 3, we have developed a statistical framework based on BSA to study alignment between various quantitative features of a functional marker and an ordinal gold standard test. It is possible that variations of BSA exist among different subpopulations of subjects, and our framework can be extended in several ways to adjust for covariates that characterize these subpopulations. Suppose we are interested in examining whether the BSA values are the same over two covariate levels (strata), say males and females. Then given a chosen summary functional, the null hypothesis $H_0 : \rho_{\text{bsa,M}} = \rho_{\text{bsa,F}}$ (BSA measures for males and females, respectively) can be tested based on the procedure described in Section 3.4. That is, one can use a Wald-type test statistic (3.8) based on the two BSA estimates computed for the two gender groups. Recently, Rahman et al. (2017) proposed a non-parametric regression framework that enables a further investigation into population heterogeneity in BSA by allowing nonlinear covariate effects. The nonparametric regression approach for BSA can be extended to evaluate the potential variations in the alignment between a functional marker and an ordinal scale according to a continuous covariate.

In practical situations, predicting ordinal response using quantitative features of a functional marker may be of great interest. We expect that fitting a generalized linear model with ordinal measurements as response and summary functionals with high

BSA values as predictors provides a basic framework for prediction, provided variable selection and multicollinearity are addressed appropriately. Future work needed in this direction includes extending our framework to select candidate quantitative features in a purely data-driven manner as well as further investigating the possibility of combining multiple summary functionals to reduce dimension and maximize prediction performance.

In Chapter 4, we have proposed a statistical framework for evaluating the diagnostic accuracy of quantitative features using AUC and describing its heterogeneity among different subpopulations. The proposed framework assumes an existence of already widely-used quantitative features or those that are newly chosen based on *a priori* scientific information. But this is not always the case in some studies, especially for those that involve recently introduced devices. Future work thus includes extending the proposed framework to select candidate quantitative features in a purely data-driven manner using modern analytic methods, e.g., tree-based methods. In the long term, we plan to derive an empirical summary functional form that produces maximum AUC given any functional marker data.

Our framework can be extended to evaluate features from a 2D image marker, which is a generalization of a 1D functional marker considered in our work. For instance, radiomics is an emerging field which seeks to take full advantage of all the information contained in multiple medical imaging modalities (Florez et al., 2018). Several quantitative features are derived to identify important regions of interest and discriminate normal healthy pattern from abnormal pattern: run length, intensity, distance zone entropy, coarseness, gray-level variance and many more. The main challenges are to identify a sensible smoothing method that allows accurate and efficient estimation of these features from observed pixel-wise data and to formulate a generalized summary functional concept that projects images not only to a real number space but also to a space of arrays and vectors.

Appendix A

A.1 Derivation of quadratic form (2.5)

In this section, we derive the quadratic form the extended measure of distance D_k in terms of distinct pairwise differences $\mathbf{X}$. Firstly, consider

$$\sum_{1\leq p\leq q\leq k}(Y_p-Y_q)^2=\sum_{q=2,\ldots,k}(Y_1-Y_q)^2+\sum_{q=3,\ldots,k}(Y_3-Y_q)^2+\ldots+\sum_{q=k}(Y_{k-1}-Y_q)^2. \quad \text{(A.1)}$$

Also, define a $(k-1)\times(k-1)$ matrix $\mathbf{M_s}$ with $(m,n)^{\text{th}}$ element given as: $\{M_s\}_{mn}=0$ if $\min(m,n)<s-1$, and $\{M_s\}_{mn}=k-\max(m,n)$ otherwise. Then, we can express the first and second terms on the right-hand side of equation (A.1) as $\sum_{1\leq p\leq q\leq k}(Y_p-Y_q)^2=\mathbf{X^T M_1 X}$ and $\sum_{q=3,\ldots,k}(Y_2-Y_q)^2=\mathbf{X^T M_2 X}$, respectively. Repeat such computation till the last term on the right-hand side of equation (A.1). Then add the results to re-express the left-hand side of equation (A.1) as $\sum_{1\leq p<q\leq k}(Y_p-Y_q)^2=\mathbf{X^T}(\sum_{s=1,\ldots,k-1}\mathbf{M_s})\mathbf{X}$, where $(m,n)^{\text{th}}$ element of the matrix $\sum_{s=1,\ldots,k-1}\mathbf{M_s}$ can be derived as: $\{\sum_{s=1,\ldots,k-1}M_s\}_{m,n}=m(k-n)$ if $1\leq m\leq n\leq k-1$, and $\{\sum_{s=1,\ldots,k-1}M_s\}_{m,n}=n(k-m)$ if $1\leq n<m\leq k-1$. By applying some basic linear algebra, we find that $\sum_{s=1,\ldots,k-1}\mathbf{M_s}=\text{adj}(\mathbf{AA^T})$, where $\text{adj}(\mathbf{AA^T})$ denotes the adjugate matrix of $\mathbf{AA^T}$.

Consequently, we can derive quadratic form (2.5) as

$$\sum_{1\le p<q\le k}(Y_p-Y_q)^2=\mathbf{X^T}\text{adj}(\mathbf{AA^T})\mathbf{X} \implies \mathbf{X^T}(\mathbf{AA^T})^{-\mathbf{1}}\mathbf{X}=\frac{\sum\limits_{1\le p<q\le k}(Y_p-Y_q)^2}{k}$$
$$\implies \mathbf{X^T}\Big\{\frac{2}{k-1}(\mathbf{AA^T})^{-\mathbf{1}}\Big\}\mathbf{X}=D_k^2,$$

by noting that $\det(\mathbf{AA^T})=k$ and $(\mathbf{AA^T})^{-\mathbf{1}}=\text{adj}(\mathbf{AA^T})/\det(\mathbf{AA^T})$.

A.2 Steps for two-sample hypothesis testing

Step 1: Denote $\mathbf{X}_{\text{obs}}^{(\mathbf{1})}$ as observations from a first group of raters and $\mathbf{X}_{\text{obs}}^{(\mathbf{2})}$ as observations from a second group of raters. Take B bootstrap samples from the observed data matrix $\mathbf{X}_{\text{obs}}^{(\mathbf{1})}$ and $\mathbf{X}_{\text{obs}}^{(\mathbf{2})}$ at the subject level with replacement, respectively.

Step 2: Compute $g(\hat{\theta}_k^{(1)(b)})$ and $g(\hat{\theta}_k^{(2)(b)})$ for each bootstrap sample $\mathbf{X}_{\text{obs}}^{(\mathbf{1})(\mathbf{b})}$ and $\mathbf{X}_{\text{obs}}^{(\mathbf{2})(\mathbf{b})}$, respectively, $b=1,2,...,B$.

Step 3: Compute B differences $g(\hat{\theta}_k^{(D)(b)})=g(\hat{\theta}_k^{(1)(b)})-g(\hat{\theta}_k^{(2)(b)})$, $b=1,2,...,B$

Step 4: Compute the standard deviation of B differences between the two ODI estimates, which gives a bootstrap estimate of the standard error:

$$\widehat{\text{SE}}_B\{g(\hat{\theta}_k^{(D)(b)})\}=\left[\frac{1}{B}\sum_{b=1}^{B}\left\{g(\hat{\theta}_k^{(D)(b)})-\overline{g(\hat{\theta}_k^{(D)(b)})}_B\right\}^2\right]^{1/2}, \tag{A.2}$$

where $\overline{g(\hat{\theta}_k^{(D)(b)})}_B=\frac{1}{B}\sum\limits_{b=1}^{B}g(\hat{\theta}_k^{(D)(b)})$.

Step 5: Given type I error rate of α and bootstrap standard error estimate (A.2), reject the null hypothesis if

$$\left|\frac{g(\hat{\theta}_k^{(D)})}{\widehat{\text{SE}}_B\{g(\hat{\theta}_k^{(D)})\}}\right|\ge z_{1-\alpha/2},$$

and conclude that degrees of agreement are not equal between two groups of raters.

Appendix B

B.1 Proof of Theorem 3.2.1

We adopt the notation provided by Peng et al. (2011). Let $Z_{Ni} = (\phi_{N_i}(W_i), Y_i)$ and $\Omega_{n,K} = \{(s_1, \ldots, s_K) : 1 \leq s_k \leq n, s_1, \ldots, s_K \text{ are distinct}\}$. For $(m_1, \ldots, m_K) \in \Omega_{n,K}$, define

$$\psi(Z_{Nm_1}, \ldots, Z_{Nm_K}) = I\{(Y_{m_1}, \ldots, Y_{m_K}) \in \Theta_K\} \times \sum_{p=1}^{K} \Big[Y_{m_p} - \sum_{q=1}^{K} I\{\phi_{N_{m_p}}(W_{m_p}) \geq \phi_{N_{m_q}}(W_{m_q})\}\Big]^2.$$

Denote $p_k = Pr(Y = k)$ for $k = 1, \ldots, K$, $C_K = (K^3 - K)/6$ and $\gamma_K = 1/(C_K \cdot K!)$. Define $h(Z_{Nm_1}, \ldots, Z_{Nm_K}) = 1 - \psi(Z_{Nm_1}, \ldots, Z_{Nm_K})\gamma_K(\prod_{k=1}^{K} p_k)^{-1}$, $h_1(z_{N1}) = E\{h(z_{N1}, Z_{N2}, \ldots, Z_{NK})\}$ and $\widetilde{h}_1(z_{N1}) = h_1(z_{N1}) - \rho_{\text{bsa}}(\phi_N(W), Y)$.

Furthermore, let

$$D_{Npq} = \phi_N(W_{(*p)}) - \phi_N(W_{(*q)}) \quad \text{and} \quad D_{pq} = \phi(X_{(*p)}) - \phi(X_{(*q)}),$$

for $p, q = 1, \ldots, K$ and $p \neq q$. Then it has been shown that the true BSA measures with respect to $\rho_{\text{bsa}}(\phi_N(W), Y)$ and $\rho_{\text{bsa}}(\phi(X), Y)$ can be written as (Dai et al., 2015):

$$\rho_{\text{bsa}}(\phi_N(W), Y) = \frac{2\sum_{q=1}^{K-1}\sum_{p=q+1}^{K}(p-q)Pr(D_{Npq} > 0)}{C_K} - 1 \tag{B.1}$$

and

$$\rho_{\text{bsa}}(\phi(X), Y) = \frac{2\sum_{q=1}^{K-1}\sum_{p=q+1}^{K}(p-q)Pr(D_{pq}>0)}{C_K} - 1, \tag{B.2}$$

respectively.

The regularity conditions include:

(A1) $p_k > 0$ for $k = 1, \ldots, K$;

(A2) $\text{Var}\{h_1(Z_{N1})\} > 0$;

(A3) 0 is the continuity point for the distribution function of D_{pq};

(A4) $E|\widetilde{h}_1(Z_{N1})|^3 < \infty$;

We first prove the consistency of the proposed estimator by considering two separate parts:

$$\begin{aligned}&\hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi(X), Y)\\ &= \hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi_N(W), Y) + \rho_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi(X), Y) = T_1 + T_2\end{aligned} \tag{B.3}$$

Firstly, by assuming conditions A1 and A2, and regarding $\phi_N(W)$ as a continuous random variable for fixed N_i values, it follows from the proof of Theorem 1 in Peng et al. (2011) that $T_1 \to 0$ as $n \to \infty$. Secondly, $\phi_N(W_{(*p)}) \xrightarrow{d} \phi(X_{(*p)})$ (provided that $P_N \to 0$ as $n \to \infty$), implies $[\phi_N(W_{(*p)}), \phi_N(W_{(*q)})]^T \xrightarrow{d} [\phi(X_{(*p)}), \phi(X_{(*q)})]^T$ (since $\phi_N(W_{(*p)})$ and $\phi_N(W_{(*q)})$ are independent), which in turn implies that $D_{Npq} \xrightarrow{d} D_{pq}$ by the continuous mapping theorem. Then, under condition A3, we obtain by (B.1) and (B.2) that

$$T_2 = \frac{2}{C_K}\sum_{q=1}^{K-1}\sum_{p=q+1}^{K}(p-q)\Big\{Pr(D_{Npq}>0) - Pr(D_{pq}>0)\Big\} \to 0, \text{ as } n \to \infty.$$

Therefore, $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ is a consistent estimator for $\rho_{\text{bsa}}(\phi(X), Y)$.

Now we prove the asymptotic normality of the proposed estimator. Consider the

decomposition:

$$\sqrt{n}\big|\hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi(X), Y)\big| = \sqrt{n}T_1 + \sqrt{n}T_2,$$

where T_1 and T_2 are defined as in (B.3). Under conditions A1 and A2, by applying the projection method of the U-statistic (Shao, 2003) and some standard asymptotic arguments, it can be shown as for the proof of Theorem 2 in Peng et al. (2011) that

$$T_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_{Ni} + o(n^{-1/2}),$$

where

$$\xi_{Ni} = K\widetilde{h}_1(Z_{Ni}) - E\{\psi(Z_{N1}, \ldots, Z_{NK})\} \times \frac{\sum_{k=1}^{K} \gamma_K \cdot \big(\prod_{1\le j\le K, j\neq k} p_j\big) \cdot \{I(Y_i = k) - p_k\}}{\big(\prod_{k=1}^{K} p_k\big)^2},$$

given fixed N_i values. Furthermore, under condition A4, it is straightforward to infer that $E|\xi_{Ni}|^3 < \infty$. Then, noting that $\xi_{N1}, \ldots, \xi_{Nn}$ are independent, we can invoke the Liapounov's Central Limit Theorem to obtain (Greene, 2011):

$$\sqrt{n}T_1 \xrightarrow{d} N(0, \sigma^2_{\text{bsa}}),$$

where $\sigma^2_{\text{bsa}} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E(\xi^2_{Ni})$.

To complete the proof, it now suffices to show that $\sqrt{n}T_2 \to 0$ as $n \to \infty$ and invoke the Slutsky's theorem. Suppose $\sup_i |\phi_{N_i}(W_i) - \phi(X_i)| \le P^*_N$ with probability 1 for some $P^*_N \le \mathcal{O}_p(P_N)$. Furthermore, let $\mathcal{D}$ be the support of D_{pq} and f be the

density function of D_{pq} such that $|f(x)| \leq M, \forall x \in \mathcal{D}$ for some $0 \leq M < \infty$. Then,

$$\begin{aligned} Pr(D_{Npq} > 0) &= Pr\{\phi_N(W_{(*p)}) > \phi_N(W_{(*q)})\} \\ &\leq Pr\{P_N^* - \phi(X_{(*q)}) > -P_N^* - \phi(X_{(*p)})\} \\ &= Pr(D_{pq} > -2P_N^*). \end{aligned}$$

This implies that

$$\begin{aligned} |Pr(D_{Npq} > 0) - Pr(D_{pq} > 0)| &\leq |Pr(D_{pq} > -2P_N^*) - Pr(D_{pq} > 0)| \\ &= \Big| \int_{-2P_N^*}^{\infty} f(u)du - \int_0^{\infty} f(u)du \Big| \\ &\leq M \cdot 2P_N^* \leq \mathcal{O}_p(P_N), \end{aligned}$$

for some $0 \leq M < \infty$. Therefore, provided that $\sqrt{n}P_N \to 0$ as $n \to \infty$, we can establish by (B.1) and (B.2) that

$$\begin{aligned} \sqrt{n}T_2 &= \frac{2\sqrt{n}}{C_K} \sum_{q=1}^{K-1} \sum_{p=q+1}^{K} (p-q)\Big\{P(D_{Npq} > 0) - P(D_{pq} > 0)\Big\} \\ &\leq M^* \cdot \sqrt{n} \cdot \mathcal{O}_p(P_N) \longrightarrow 0, \text{ as } n \to \infty, \end{aligned}$$

where M^* is some constant such that $0 \leq M^* < \infty$.

B.2 Specification of Kernel Function

Kernel function K_ν of order (ν, ω) is defined as a Lipschitz continuous function characterized by vanishing $(\omega - 2)$ moments (Gasser and Müller, 1984, Müller, 1984,

1985):

$$\int K_\nu(u)u^q du = \begin{cases} 0, & 0 \le q \le \omega - 1, q \ne \nu, \\ (-1)^\nu \nu!, & q = \nu, \\ C, & q = \omega, \end{cases}$$

where $C = \int K_\nu(u)u^\omega du \neq 0$. If $\omega = \nu + 2$, K_ν is called a standard kernel function; if $\omega > \nu + 2$, K_ν is called a higher-order kernel (Beran et al., 2013).

B.3 Consistency of the estimators for the three special-case summary functionals

The following theorem establishes the consistency of the estimators for the three special-case summary functionals discussed in Section 3.3.

Theorem B.3.1. *Suppose that $\phi(X_i)$ $(i = 1, \ldots, n)$ is one of AUC-type, magnitude- or time-specific functionals, and $\phi_{N_i}(W_i)$ is its estimator defined in Section 3.3. Then, under the conditions B1 and B2, if the regularity conditions C1-C2, C3-C4 and C5-C6 hold for AUC-type functionals, magnitude-specific functionals and time-specific functionals, respectively, we have $|\phi_{N_i}(W_i) - \phi(X_i)| = \mathcal{O}_p(\beta_{N_i}^\theta)$, where $0 < \theta \le 1$ and β_{N_i} (defined below) is a function of bandwidth b_{N_i} that approaches zero as $N_i \to \infty$ for a suitable choice of b_{N_i}.*

This theorem implies that each estimator for three special formulations of summary functionals is consistent; that is, $\phi_{N_i}(W_i)$ satisfies the condition in Theorem 3.2.1 by setting $\sup_i \beta_{N_i}^\theta = P_N$.

The first step of the proof of this theorem is to establish the consistency of the Gasser-Müller kernel estimators. Assume the following preliminary conditions.

(B1) $E|\epsilon_1|^r < \infty$ for some $r > 2$;

(B2) $\liminf_{N_i \to \infty} N_i b_{N_i}^k > 0$, $\liminf_{N_i \to \infty} N_i^{1-2/r} b_{N_i} (\log N_i)^{-1} > 0$.

Let $\beta_{N_i} = b_{N_i}^{\omega-\nu} + \{(\log N_i)/(N_i b_{N_i}^{2\nu+1})\}^{1/2}$. Then, it follows from Lemma 2.2 of Müller (1985) that

$$\sup_{t\in[0,1]} |\hat{W}_i^{[\nu]}(t) - X_i^{(\nu)}(t)| = \mathcal{O}(\beta_{N_i}) \quad \text{a.s..} \tag{B.4}$$

If the bandwidth is chosen as $b_{N_i} = \mathcal{O}\{(\log N_i/N_i)^{1/(2\omega+1)}\}$, we can establish that $\beta_{N_i} = (\log N_i/N_i)^{(\omega-\nu)/(2\omega+1)}$, which goes to zero as $N_i \to \infty$.

Based on this result, we prove that each of the estimators for the three special cases of summary functionals is consistent in the following subsections. For the sake of simplicity, we will drop the index i throughout the proof.

B.3.1 AUC-type functionals

Consider the collection $\mathcal{P}$ of partitions of the time interval $[0,1]$, that is, $\mathcal{P}$ consists of a finite sequence of output design points of the form $0 = t_1 < t_2 < \cdots < t_N = 1$. Then the total variation of any realization $\hat{w}^{[\nu]}$ of $\hat{W}^{[\nu]}$ is given by

$$TV(\hat{w}^{(\nu)}) = \sup\Big\{ \sum_{j=1}^{N-1} |\hat{w}^{[\nu]}(t_{j+1}) - \hat{w}^{[\nu]}(t_j)| : (t_1, t_2, \ldots, t_N) \in \mathcal{P}\Big\}.$$

Provided that $TV(\hat{w}^{[\nu]})$ is finite, it has been shown that (Hua and Wang, 1981)

$$\Big| \int_0^1 \hat{w}^{[\nu]}(t)dt - \sum_{j=1}^{N-1} \hat{w}^{[\nu]}(t_j)(t_{j+1} - t_j)\Big| < TV(\hat{w}^{[\nu]}) \max_{0\le j\le N-1} |t_{j+1} - t_j|. \tag{B.5}$$

Now let $\mathcal{G}$ denote the space consisting of the Gasser-Müller kernel-based estimators $\hat{W}^{(\nu)}$ and assume the following regularity conditions:

(C1) $\sup_{\hat{w}^{[\nu]}\in\mathcal{G}} TV(\hat{w}^{[\nu]}) = \mathcal{O}_p(1)$;

(C2) $\max_{0\le j\le N-1} |t_{j+1} - t_j| = \mathcal{O}(1/N)$.

Then, by conditions C1 and C2, and using (B.4) and (B.5), it follows that

$$\begin{aligned}
\big|\phi^{[\nu]}_{N,\mathrm{AUC}}(W) - \phi^{[\nu]}_{\mathrm{AUC}}(X)\big| &= \Big|\sum_{j=1}^{N-1} \hat{W}^{[\nu]}(t_j)(t_{j+1}-t_j) - \int_0^1 X^{(\nu)}(t)dt\Big| \\
&< \Big|\sum_{j=1}^{N-1} \hat{W}^{[\nu]}(t_j)(t_{j+1}-t_j) - \int_0^1 \hat{W}^{[\nu]}(t)dt\Big| \\
&+ \Big|\int_0^1 \hat{W}^{[\nu]}(t)dt - \int_0^1 X^{(\nu)}(t)dt\Big| \\
&< TV(\hat{w}^{[\nu]}) \max_{0\le j\le N-1} |t_{j+1}-t_j| + \mathcal{O}_p(\beta_N) \\
&= \mathcal{O}_p\Big(\frac{1}{N}\Big) + \mathcal{O}_p(\beta_N).
\end{aligned}$$

B.3.2 Magnitude-specific functionals

For a general magnitude-specific summary functional, it is obvious to see that $|\phi^{[\nu]}_{N,\mathrm{MAG}(t^*)}(W) - \phi^{[\nu]}_{\mathrm{MAG}(t^*)}(X)| = \mathcal{O}_p(\beta_N)$ for a given time point $t^* \in [0,1]$ by (B.4).

We now turn to establishing the consistency of $\phi^{[\nu]}_{N,\mathrm{MAX}}(W)$. Let $t_{\max}$ and $\hat{t}_{\max}$ denote the times at which the maximum values are achieved by $X^{(\nu)}$ and $\hat{W}^{[\nu]}$, respectively. Then, under the following conditions

(C3) $\exists$ a and b $(0 < a < t_{\mathrm{MAX}} < b < 1)$ such that $X^{(\nu)}$ is monotonously increasing on $[a, t_{\mathrm{MAX}}]$ and monotonously decreasing on $[t_{\mathrm{MAX}}, b]$,

(C4) $\exists$ $c > 0$ and $\theta \le 1$ such that $|X^{(\nu)}(t) - X^{(\nu)}(t_{\max})| > c|t - t_{\max}|^{\theta}$,

it follows from Lemma 2.2 of Müller (1985) that $|\phi^{[\nu]}_{N,\mathrm{MAX}}(W) - \phi^{[\nu]}_{\mathrm{MAX}}(X)| = |\hat{W}^{[\nu]}(\hat{t}_{\max}) - X^{(\nu)}(t_{\max})| = \mathcal{O}_p(\beta_N)$ given that $X^{(\nu)}$ has a unique maximum. On the other hand, assuming that $X^{(\nu)}$ is monotonously decreasing (increasing) on $[a, t_{\mathrm{MAX}}]$ $([t_{\mathrm{MAX}}, b])$ and $|X^{(\nu)}(t) - X^{(\nu)}(t_{\min})| > c|t - t_{\min}|^{\theta}$ in place of condition C3 and C4, respectively, a similar proof establishes $|\phi^{[\nu]}_{N,\mathrm{MIN}}(W) - \phi^{[\nu]}_{\mathrm{MIN}}(X)| = |\hat{W}^{[\nu]}(\hat{t}_{\min}) - X^{(\nu)}(t_{\min})| = \mathcal{O}_p(\beta_N)$ given that $X^{(\nu)}$ has a unique minimum.

B.3.3 Time-specific functionals

Firstly, under conditions C3 and C4, it follows from Lemma 2.3 of that Müller (1985) $|\phi^{[\nu]}_{N,\text{tMAX}}(W) - \phi^{[\nu]}_{\text{tMAX}}(X)| = \mathcal{O}_p(\beta^\theta_N)$ and $|\phi^{[\nu]}_{N,\text{tMIN}}(W) - \phi^{[\nu]}_{\text{tMIN}}(X)| = \mathcal{O}_p(\beta^\theta_N)$.

We now turn to establishing the consistency of a general time-specific summary functional $\phi^{(\nu)}_{N,\text{TIME}(\eta)}(W)$. Let t_η and $\hat{t}_\eta$ denote the times at which $X^{(\nu)}$ and $\hat{W}^{[\nu]}$ attain a threshold value η, respectively. Similar to the conditions outlined in Lemmas 2.3 and 2.4 of Müller (1985), we assume:

(C5) $\exists$ a and b $(0 < a < t_\eta < b < 1)$ such that $X^{(\nu)}$ is strictly monotonous on $[a, b]$;

(C6) $\exists$ $c > 0$ and $\theta \leq 1$ such that $|X^{(\nu)}(t) - X^{(\nu)}(t_\eta)| > c|t - t_\eta|^\theta$.

Given that η is a unique threshold value that $X^{(\nu)}$ attains, we can always find $\delta > 0$ such that $|X^{(\nu)}(t_\eta) - X^{(\nu)}(u)| > \delta$ a.s. for $u \notin [a, b]$. For sufficiently large N, we have $c\beta_N < \delta$ and thus

$$\begin{aligned}
|X^{(\nu)}(t_\eta) - X^{(\nu)}(\hat{t}_\eta)| &= |X^{(\nu)}(t_\eta) - \hat{W}^{[\nu]}(\hat{t}_\eta) + \hat{W}^{[\nu]}(\hat{t}_\eta) - X^{(\nu)}(\hat{t}_\eta)| \\
&\leq |X^{(\nu)}(t_\eta) - \hat{W}^{[\nu]}(\hat{t}_\eta)| + |\hat{W}^{[\nu]}(\hat{t}_\eta) - X^{(\nu)}(\hat{t}_\eta)| \\
&< |\eta - \eta| + \delta \\
&= \delta \quad \text{a.s.},
\end{aligned}$$

which implies that $\hat{t}_\eta \in [a, b]$ a.s. With out loss of generality, assume $\hat{W}^{[\nu]}(\hat{t}_\eta) - \hat{W}^{[\nu]}(t_\eta) \geq 0$. Then, it follows that

$$\begin{aligned}
|\hat{t}_\eta - t_\eta|^\theta &< c^{-1}|X^{(\nu)}(t_\eta) - X^{(\nu)}(\hat{t}_\eta)| \\
&\leq c^{-1}|\hat{W}^{[\nu]}(\hat{t}_\eta) - \hat{W}^{[\nu]}(t_\eta) + X^{(\nu)}(t_\eta) - X^{(\nu)}(\hat{t}_\eta)| \\
&\leq c^{-1}|\hat{W}^{[\nu]}(\hat{t}_\eta) - X^{(\nu)}(\hat{t}_\eta)| + |X^{(\nu)}(t_\eta) - \hat{W}^{[\nu]}(t_\eta)| \\
&= \mathcal{O}_p(\beta_N).
\end{aligned}$$

Therefore, $|\phi^{[\nu]}_{N,\text{TIME}(\eta)}(W) - \phi^{[\nu]}_{\text{TIME}(\eta)}(X)| = \mathcal{O}_p(\beta^\theta_N)$.

B.4 Additional Simulations

B.4.1 Evaluation of the proposed hypothesis testing procedure

We examined the empirical rejection rates of the hypothesis testing procedure presented in Section 3.4. Firstly, consider testing the null H_0: $\rho_{\text{bsa}}(\phi_{\text{AUC},1}(X), Y) = \rho_{\text{bsa}}(\phi_{\text{AUC},2}(X), Y)$, where $\phi_{\text{AUC},1}(X)$ and $\phi_{\text{AUC},2}(X)$ are based on the sub-time intervals $\mathcal{T}_1 = [0, 0.5]$ and $\mathcal{T}_2 = [0.5, 1]$, respectively. Data are generated under the first two scenarios presented in Section 3.5. The empirical rejection rates in scenario 1 represent the empirical sizes of the test as the two true BSA measures based on the respective sub-time intervals are equal, that is, $\rho_{\text{bsa}}(\phi_{\text{AUC},1}(X), Y) = \rho_{\text{bsa}}(\phi_{\text{AUC},2}(X), Y) = 0.690$. On the other hand, the empirical rejection rates in scenario 2 correspond to the empirical power of the test as the BSA increase over time, that is, $\rho_{\text{bsa}}(\phi_{\text{AUC},1}(X), Y) = 0.212$ and $\rho_{\text{bsa}}(\phi_{\text{AUC},2}(X), Y) = 0.566$.

We further consider a hypothesis test that involves comparing BSA measures that arise from two different types of summary functionals. Specifically, the null hypothesis H_0: $\rho_{\text{bsa}}(\phi_{\text{AUC}}(X), Y) = \rho_{\text{bsa}}(\phi_{\text{MAG}(\frac{1}{4})}(X), Y)$ is tested using data generated under scenario 2; that is, the empirical power of the test that compares between $\rho_{\text{bsa}}(\phi_{\text{AUC}}(X), Y) = 0.425$ and $= \rho_{\text{bsa}}(\phi_{\text{MAG}(\frac{1}{4})}(X), Y) = 0.208$ is assessed.

Data are generated under the five study designs (a) – (e) used in Section 3.5 to simulate varying density of time points. We adopt the same measurement error model and smoothing technique that are used in Section 3.5. The empirical rejection rates are calculated as the proportion of 1000 simulated data sets of size $n = 40$ and 60 for which the null hypothesis is rejected with significance $\alpha = 0.05$.

Empirical rejection rates for respective null hypotheses (H_0) reported in Table B.1 demonstrate satisfactory performance of the proposed hypothesis testing procedure. In scenario 1, the empirical rejection rates rapidly approach the nominal level of 0.05

Table B.1: Simulation results on empirical rejection rates of the proposed hypothesis testing procedure.N denotes the five study designs: (a) unbalanced design with N_i following a Poisson distribution with mean 20; (b) unbalanced design with N_i following a Poisson distribution with mean 40; (c) balanced design with $N_i = 20$; (d) balanced design with $N_i = 40$; and (e) balanced design with $N_i = 60$.

Scenario	True Values	N	Rejection Rate ($n = 40$)	Rejection Rate ($n = 60$)
1	$\rho_{\text{bsa}}(\phi_{\text{AUC},1}(X), Y)$	(a)	0.035	0.039
	$= \rho_{\text{bsa}}(\phi_{\text{AUC},2}(X), Y)$	(b)	0.043	0.050
	$= 0.690$	(c)	0.042	0.046
		(d)	0.031	0.049
		(e)	0.034	0.040
2	$\rho_{\text{bsa}}(\phi_{\text{AUC},1}(X), Y) = 0.212$	(a)	0.868	0.977
	vs.	(b)	0.856	0.964
	$\rho_{\text{bsa}}(\phi_{\text{AUC},2}(X), Y) = 0.566$	(c)	0.857	0.984
		(d)	0.863	0.975
		(e)	0.872	0.985
2	$\rho_{\text{bsa}}(\phi_{\text{AUC}}(X), Y) = 0.425$	(a)	0.843	0.980
	vs.	(b)	0.853	0.978
	$\rho_{\text{bsa}}(\phi_{\text{MAG}(\frac{1}{4})}(X), Y) = 0.208$	(c)	0.834	0.976
		(d)	0.857	0.985
		(e)	0.845	0.986

as sample size increases. In scenario 2, the empirical power of the test under both cases appears satisfactory even with relatively small sample size and sparse data.

B.4.2 Finite-sample performance at the first derivative level of the summary functionals

We conducted further simulation studies to assess the finite-sample performance of the proposed approaches to evaluate alignment between the summary functionals based on the first derivatives of functional markers and the corresponding ordinal outcomes. Following the lines of previous simulation study in Section 3.5, performances of BSA estimators based on three special cases of summary functionals (AUC-type, magnitude-specific, and time-specific) were assessed. We first set $K = 3$ and generate ordinal outcomes Y from the multinomial distribution with equal probabilities, that is, $Pr(Y = k) = 1/3$, $k = 1, 2, 3$.

Given each $Y = k$, the true functional markers X are generated over a time interval $\mathcal{T} = [0, 1]$ under five different scenarios depending on the type of a summary functional to be analyzed. For the AUC-type summary functionals, we generate $X(t)$ as a Gaussian process with mean functions $\mu(t) = kt$ (scenario 1) and $\mu(t) = (k/2)t^2$ (scenario 2). Scenarios 1 and 2 represent a constant and improving degrees of alignment in terms of the first derivative AUC over the time interval, respectively. Performances based on the magnitude-specific summary functionals are evaluated using a Gaussian process with mean function $\mu(t) = -(k/\pi)\cos(\pi t)$, whose unique maximum value 1 is attained at time 1/2 (scenario 3). In all configurations involving the generation of Gaussian processes (scenarios 1-3), we adopted a common covariance function $\mathrm{Cov}(X(s), X(t)) = \exp\{-(s-t)^2\}$, $s, t \in \mathcal{T}$. We consider two different scenarios for evaluating the finite-sample performance based on the time-specific summary functionals. In scenario 4, if $Y = 1$, $X(t) = -(2\pi)^{-1}\cos(2\pi t)$ with probability 1; if $Y = 2$, $X(t) = -(0.25\pi)^{-1}\cos(0.25\pi t)$ with probability 1; and if

$Y = 3$, $X(t) = -(0.5\pi)^{-1}\cos(0.5\pi t)$ with probability 1. In scenario 5, if $Y = 1$, $X(t) = -(2\pi)^{-1}\cos(2\pi t)$ with probability 1; if $Y = 2$, $X(t) = -(0.66\pi)^{-1}\cos(0.66\pi t)$ with probability 1; and if $Y = 3$, $X(t) = -\pi^{-1}\cos(\pi t)$ with probability 1.

Each functional markers are generated with measurement error under the five study designs (a) – (e) provided in Section 3.5. We obtain the Gasser-Müller kernel estimators of the first derivatives evaluated on 300 output design points using a polynomial kernel of degree 3 (Müller, 1984) and an automatically adapted global "plug-in" bandwidth that is asymptotically optimal with respect to the mean integrated square error (MISE) (Gasser et al., 1991). Results presented in Table B.2 are based on 1000 simulated datasets of size $n = 40$ and 60.

From Table B.2, we see that the empirical biases maintain a pretty high level when data are sparse and are generally greater than those obtained in Table 3.1 from Section 3.5. This is partly due to a slower rate of convergence for estimated summary functionals involving the first derivative of functional markers. Therefore, we recommend using functional markers collected at least 40 time points to obtain reliable BSA estimates, especially when magnitude-specific or time-specific summary functional is considered. Despite slower convergence, we see that empirical biases eventually approach 0 as the sample size and number of time points increase. Likewise, the estimated standard errors and coverage probabilities approach the empirical standard deviations and nominal level of 95%, respectively, as the sample size and number of time points increase, suggesting that the proposed estimation and inference procedures work fairly well at the level of the first derivative of functional markers.

Table B.2: Simulation results on proposed BSA measures: mean of 1000 biases (EmpBias), standard deviation of 1000 BSA estimates (EmpSD), mean of 1000 standard error estimates (EstSE) and proportion of 95% CIs containing the true BSA value (Cov95). N denotes the five study designs: (a) unbalanced design with N_i following a Poisson distribution with mean 20; (b) unbalanced design with N_i following a Poisson distribution with mean 40; (c) balanced design with $N_i = 20$; (d) balanced design with $N_i = 40$; and (e) balanced design with $N_i = 60$.

			$n = 40$				$n = 60$			
Scenario	True BSA Values	N	EmpBias	EmpSD	EstSE	Cov95	EmpBias	EmpSD	EstSE	Cov95
1	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{AUC}}(X), Y)$	(a)	0.006	0.113	0.113	0.940	0.003	0.089	0.092	0.952
	$= 0.631$	(b)	0.003	0.112	0.113	0.937	0.002	0.093	0.091	0.929
		(c)	0.006	0.113	0.113	0.932	0.003	0.093	0.092	0.939
		(d)	0.003	0.113	0.115	0.930	0.002	0.088	0.091	0.950
		(e)	0.006	0.114	0.112	0.918	0.001	0.092	0.092	0.935
2	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{AUC}}(X), Y)$	(a)	0.006	0.157	0.160	0.947	0.005	0.125	0.128	0.959
	$= 0.359$	(b)	0.004	0.155	0.159	0.947	0.001	0.128	0.128	0.949
		(c)	0.008	0.158	0.160	0.941	0.004	0.125	0.128	0.947
		(d)	0.002	0.158	0.161	0.949	0.001	0.121	0.128	0.956
		(e)	0.003	0.160	0.159	0.941	0.003	0.125	0.128	0.946
3	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{MAG}(1/2)}(X), Y)$	(a)	-0.060	0.136	0.146	0.944	-0.059	0.111	0.116	0.931
	$= 0.533$	(b)	-0.015	0.136	0.137	0.941	-0.012	0.105	0.110	0.949
		(c)	-0.056	0.142	0.144	0.932	-0.060	0.116	0.116	0.919
		(d)	-0.008	0.133	0.136	0.942	-0.008	0.109	0.110	0.934
		(e)	-0.002	0.127	0.135	0.949	-0.001	0.107	0.108	0.943
3	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{MAX}}(X), Y)$	(a)	-0.048	0.145	0.152	0.953	-0.048	0.115	0.123	0.938
	$= 0.464$	(b)	-0.026	0.145	0.150	0.948	-0.025	0.116	0.119	0.938
		(c)	-0.042	0.143	0.152	0.953	-0.044	0.115	0.122	0.948
		(d)	-0.018	0.143	0.149	0.957	-0.014	0.118	0.119	0.939
		(e)	-0.001	0.141	0.146	0.957	0.003	0.115	0.116	0.943
4	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{TIME}(1/2)}(X), Y)$	(a)	0.017	0.029	0.027	0.971	0.015	0.022	0.020	0.975
	$= 0.500$	(b)	-0.011	0.042	0.028	0.886	-0.010	0.031	0.024	0.970
		(c)	0.016	0.028	0.024	0.960	0.014	0.020	0.020	0.988
		(d)	-0.013	0.043	0.028	0.877	-0.012	0.033	0.024	0.963
		(e)	-0.010	0.022	0.014	0.911	-0.009	0.020	0.018	0.959
5	$\rho_{\text{bsa}}(\phi^{[1]}_{\text{tMAX}}(X), Y)$	(a)	-0.029	0.077	0.085	0.969	-0.027	0.068	0.064	0.949
	$= 0.500$	(b)	-0.015	0.060	0.050	0.978	-0.016	0.048	0.042	0.981
		(c)	-0.027	0.084	0.074	0.972	-0.025	0.066	0.062	0.952
		(d)	-0.016	0.061	0.048	0.981	-0.017	0.050	0.041	0.983
		(e)	-0.009	0.053	0.036	0.966	-0.008	0.042	0.031	0.987

Appendix C

C.1 Proof of Theorem 4.3.1

For simplicity of notation, let $\theta = \theta(\phi)$, $\theta_N = \theta(\phi_N)$, $\widehat{\theta} = \widehat{\theta}(\phi)$ and $\widehat{\theta}_N = \widehat{\theta}(\phi_N)$. Denote $h_{ij} = I\{\phi(X_i^D) > \phi(X_j^{\overline{D}})\}$, $h_{Nij} = I\{\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})\}$, $h_{N11} = E(h_{Nij} \mid W_i^D = w^D)$, $h_{N12} = E(h_{Nij} \mid W_j^{\overline{D}} = w^{\overline{D}})$, $\widetilde{h}_{N11} = h_{N11} - \theta_N$, $\widetilde{h}_{N12} = h_{N12} - \theta_N$, $h_{N11i} = E(h_{Nij} \mid W_i^D)$ and $h_{N12j} = E(h_{Nij} \mid W_j^{\overline{D}})$. The regularity conditions include:

(A1) $\phi_{N_i}(W_i^D)$ $(i = 1, \ldots, n_D)$ and $\phi_{N_j}(W_j^{\overline{D}})$ $(j = 1, \ldots, n_{\overline{D}})$ are iid within groups and mutually independent between groups, given sufficiently large N_i and N_j values;

(A2) $N_i, N_j \to \infty$ as $n \to \infty$, that is, $N_i = N_{i,n}$ and $N_j = N_{j,n}$ are sequences that tend to infinity;

(A3) $E\{|h_{Nij}|(\log^+ |h_{Nij}|)\} < \infty$;

(A4) $n_D/n \to \lambda$ as $n \to \infty$, where $\lambda \in (0, 1)$;

(A5) $0 < \mathrm{Var}(h_{Nij}) < \infty$.

• Proof of consistency

Consider the decomposition:

$$\widehat{\theta}_N - \theta = (\widehat{\theta}_N - \theta_N) + (\theta_N - \theta) = T_1 + T_2 \tag{C.1}$$

For sufficiently large fixed N_i and N_j values, $\widehat{\theta}_N$ is a two-sample U-statistic under condition A1 and is an unbiased estimator of θ_N. Then by conditions (A1) and (A2) and the strong law of large numbers for generalized U-statistics (Sen, 1977), we have $T_1 \xrightarrow{a.s.} 0$ as $n \to \infty$.

Secondly, $\phi_{N_i}(W_i^D) \xrightarrow{d} \phi(X_i^D)$ and $\phi_{N_j}(W_j^{\overline{D}}) \xrightarrow{d} \phi(X_j^{\overline{D}})$ for all i and j (provided that $P_N \to 0$ as $n \to \infty$) imply $[\phi_{N_i}(W_i^D), \phi_{N_j}(W_j^{\overline{D}})]^T \xrightarrow{d} [\phi(X_i^D), \phi(X_j^{\overline{D}})]^T$ as $n \to \infty$ under conditions (A1) and (A2). Then, by the continuous mapping theorem, $\phi_{N_i}(W_i^D) - \phi_{N_j}(W_j^{\overline{D}}) \xrightarrow{d} \phi(X_i^D) - \phi(X_j^{\overline{D}})$ as $n \to \infty$, which in turn implies that:

$$T_2 = \theta_N - \theta = \Pr(\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})) - \Pr(\phi(X_i^D) > \phi(X_j^{\overline{D}})) \longrightarrow 0, \quad \text{as} \quad n \to \infty,$$

given that 0 is a continuity point of the distribution of $\phi(X_i^D) - \phi(X_j^{\overline{D}})$. Therefore, (C.1) converges in probability to zero as $n \to \infty$, establishing that $\widehat{\theta}_N$ is consistent for θ.

• Proof of asymptotic normality

Consider the decomposition:

$$\sqrt{n}\big(\widehat{\theta}_N - \theta\big) = \sqrt{n}T_1 + \sqrt{n}T_2,$$

where T_1 and T_2 are defined as in (C.1). By the method of projection of U-statistics (Serfling, 1980) and some standard asymptotic arguments, it can be shown, for sufficiently large fixed N_i and N_j values, that

$$\begin{aligned}\sqrt{n}T_1 &= \frac{\sqrt{n}}{n_D}\sum_{i=1}^{n_D}(h_{N11i} - \theta_N) + \frac{\sqrt{n}}{n_{\overline{D}}}\sum_{j=1}^{n_{\overline{D}}}(h_{N12j} - \theta_N) + o_p(1) \\ &= \Big(\frac{n}{n_D}\Big)\frac{1}{\sqrt{n}}\sum_{i=1}^{n_D}(h_{N11i} - \theta_N) + \Big(\frac{n}{n_{\overline{D}}}\Big)\frac{1}{\sqrt{n}}\sum_{j=1}^{n_{\overline{D}}}(h_{N12j} - \theta_N).\end{aligned}$$

Then, under conditions (A1), (A4) and (A5), and noting that h_{N11i} and h_{N12j} are

independent across i and j, respectively, and mutually independent, we have

$$\sqrt{n}T_1 \xrightarrow{d} N(0, \sigma^2_{\theta(\phi)}) \quad \text{as} \quad n \to \infty,$$

where

$$\sigma^2_{\theta(\phi)} = \lim_{n\to\infty} \Big\{ \frac{1}{\lambda}\sigma^2_{N10} + \frac{1}{1-\lambda}\sigma^2_{N01} \Big\},$$

with

$$\sigma^2_{N10} = \mathrm{Cov}(h_{Nij}, h_{Nil}),\ j \neq l,$$
$$\sigma^2_{N01} = \mathrm{Cov}(h_{Nij}, h_{Nlj}),\ i \neq l.$$

To complete the proof, it now suffices to show that $\sqrt{n}T_2 \to 0$ as $n \to \infty$ and invoke the Slutsky's theorem. Suppose $\sup_i |\phi_{N_i}(W_i^D) - \phi(X_i^D)| \leq \widetilde{P}_N^D$ and $\sup_j |\phi_{N_j}(W_j^{\overline{D}}) - \phi(X_j^{\overline{D}})| \leq \widetilde{P}_N^{\overline{D}}$ with probability 1 for some $\widetilde{P}_N^D, \widetilde{P}_N^{\overline{D}} \leq \mathcal{O}_p(P_N)$. Furthermore, let $\mathcal{D}$ be the support of $\phi(X_i^D) - \phi(X_j^{\overline{D}})$, and f be the density function of $\phi(X_i^D) - \phi(X_j^{\overline{D}})$ such that $|f(x)| \leq M, \forall x \in \mathcal{D}$ for some $0 \leq M < \infty$. Then,

$$\begin{aligned}
\theta_N &= \Pr(\phi_{N_i}(W_i^D) - \phi_{N_j}(W_j^{\overline{D}}) > 0) \\
&= \Pr\{\phi_{N_i}(W_i^D) - \phi(X_i^D) - \phi(X_j^{\overline{D}}) > \phi_{N_j}(W_j^{\overline{D}}) - \phi(X_i^D) - \phi(X_j^{\overline{D}})\} \\
&\leq \Pr\{\widetilde{P}_N^D - \phi(X_j^{\overline{D}}) > -\widetilde{P}_N^{\overline{D}} - \phi(X_i^D)\} \\
&= \Pr\{\phi(X_i^D) - \phi(X_j^{\overline{D}}) > -\widetilde{P}_N^D - \widetilde{P}_N^{\overline{D}}\}
\end{aligned}$$

which implies

$$\begin{aligned}
|\theta_N - \theta| &\leq |\Pr\{\phi(X_i^D) - \phi(X_j^{\overline{D}}) > -\widetilde{P}_N^D - \widetilde{P}_N^{\overline{D}}\} - \Pr(\phi(X_i^D) - \phi(X_j^{\overline{D}}) > 0)| \\
&= \Big| \int_{-\widetilde{P}_N^D - \widetilde{P}_N^{\overline{D}}}^{\infty} f(u)du - \int_0^{\infty} f(u)du \Big| \\
&\leq M \cdot (\widetilde{P}_N^D + \widetilde{P}_N^{\overline{D}}) \leq \mathcal{O}_p(P_N).
\end{aligned}$$

Therefore, provided that $\sqrt{n}P_N \to 0$ as $n \to \infty$, we can establish that $\sqrt{n}T_2 \to 0$ as $n \to \infty$.

C.2 Proof of Theorem 4.4.1

Without loss of generality, consider the estimating equations and *estimated* estimating equations of the form:

$$S_n(\boldsymbol{\beta}) = \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} \Psi_{ij}(U_{ij} - \theta_{ij}) = \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} S_{ij}(\boldsymbol{\beta}),$$

and

$$S_{Nn}(\boldsymbol{\beta}) = \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} \Psi_{ij}(U_{Nij} - \theta_{ij}) = \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} S_{Nij}(\boldsymbol{\beta}),$$

where $\Psi_{ij} = (\partial\theta_{ij}/\partial\boldsymbol{\beta})\Omega_{ij}^{-1}$. Let $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}_N$ denote the solutions of $S_n(\boldsymbol{\beta}) = \mathbf{0}$ and $S_{Nn}(\boldsymbol{\beta}) = \mathbf{0}$, respectively. The regularity conditions include:

(B1) For a given ϕ, $[\{\phi(X_i^D), Z_i^D\},\ i = 1, \ldots, n_D]$ and $[\{\phi(X_j^{\overline{D}}), Z_j^{\overline{D}}\},\ j = 1, \ldots, n_{\overline{D}}]$ are iid within groups and mutually independent between groups. The same assumptions hold for $[\{\phi_{N_i}(W_i^D), Z_i^D\},\ i = 1, \ldots, n_D]$ and $[\{\phi_{N_j}(W_j^{\overline{D}}), Z_j^{\overline{D}}\},\ j = 1, \ldots, n_{\overline{D}}]$ given sufficiently large fixed N_i and N_j values;

(B2) $N_i, N_j \to \infty$ as $n \to \infty$, i.e., $N_i \equiv N_{i,n}$ and $N_j \equiv N_{j,n}$ are sequences that go to infinity;

(B3) $n_D/n \to \lambda$ as $n \to \infty$, where $\lambda \in (0, 1)$;

(B4) The parameter space of $\boldsymbol{\beta}$, denoted as Θ, is compact in $\mathbb{R}^p$, and there exist unique $\boldsymbol{\beta}_0 \in \Theta$ such that $E\{S_n(\boldsymbol{\beta}_0)\} = \mathbf{0}$;

(B5) The matrix $E\{\partial S_{ij}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^T\}$ is negative-definite;

(B6) $g(\cdot)$ is monotone increasing and three-times differentiable with bounded derivatives.

(B7) The covariate space is bounded.

(B8) There exists $\epsilon > 0$ such that $\Omega_{ij} > \epsilon$ for $\boldsymbol{\beta} \in \mathcal{N} \equiv N_\delta(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| < \delta\}$;

• Preliminary lemmas

We first state lemmas that are used in subsequent proofs.

Lemma C.2.1. $\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}} S_n(\boldsymbol{\beta})$, $\frac{1}{n_D n_{\overline{D}}} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} S_n(\boldsymbol{\beta})$, $\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{E}\{\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}} S_n(\boldsymbol{\beta})\}$, $\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}} S_{Nn}(\boldsymbol{\beta})$, $\frac{1}{n_D n_{\overline{D}}} \frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} S_{Nn}(\boldsymbol{\beta})$ *and* $\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{E}\{\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}} S_{Nn}(\boldsymbol{\beta})\}$ *are uniformly bounded for* $\boldsymbol{\beta} \in \mathcal{N}$.

The uniform boundedness of the six terms follow from (B6)–(B8) (Dodd and Pepe, 2003).

Lemma C.2.2. *For fixed* $\boldsymbol{\beta} \in \mathcal{N}$, $\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}) \xrightarrow{p} 0$ *as* $n \to \infty$.

Proof: Consider the decomposition:

$$
\begin{aligned}
&\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}) \\
&= \left[\frac{1}{n_D n_{\overline{D}}} S_{ij}(\boldsymbol{\beta}) - E\{S_{ij}(\boldsymbol{\beta})\}\right] + \left[E\{S_{Nij}(\boldsymbol{\beta})\} - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})\right] \\
&+ \left[E\{S_{ij}(\boldsymbol{\beta})\} - E\{S_{Nij}(\boldsymbol{\beta})\}\right] \\
&= A_1 + A_2 + A_3.
\end{aligned}
$$

Firstly, we consider A_1 and its decomposition as the following:

$$\begin{aligned} A_1 &= \frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - E\{S_{ij}(\boldsymbol{\beta})\} \\ &= \left[\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - \frac{1}{n_D} \sum_i E\{S_{ij}(\boldsymbol{\beta}) \mid \phi(X_i^D)\} \right] \\ &+ \left[\frac{1}{n_D} \sum_i E\{S_{ij}(\boldsymbol{\beta}) \mid \phi(X_i^D)\} - E\{S_{ij}(\boldsymbol{\beta})\} \right] \\ &= B_1 + B_2, \end{aligned}$$

where $E\{S_{ij}(\boldsymbol{\beta}) \mid \phi(X_i^D)\}$ is a random variable that is independent across i. For B_1, we have:

$$\begin{aligned} B_1 &= \frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - \frac{1}{n_D} \sum_i E\{S_{ij}(\boldsymbol{\beta}) \mid \phi(X_i^D)\} \\ &= \frac{1}{n_D} \sum_i \left[\frac{1}{n_{\overline{D}}} \sum_j S_{ij}(\boldsymbol{\beta}) - E\{S_{ij}(\boldsymbol{\beta}) \mid \phi(X_i^D)\} \right]. \end{aligned}$$

Since the terms inside the square bracket $[\cdot]$ are iid across j for fixed i, we can apply the weak law of large numbers (WLLN) to establish that $B_1 \xrightarrow{p} 0$ as $n \to \infty$. $B_2 \xrightarrow{p} 0$ as $n \to \infty$ can be also established by applying WLLN. Hence, $A_1 = B_1 + B_2 \xrightarrow{p} 0$ as $n \to \infty$.

For sufficiently large fixed N_i and N_j values and given (B1), we can similarly show that $A_2 \xrightarrow{p} 0$ as $n \to \infty$. In this case, we decompose A_2 using independent random variables $E\{S_{Nij}(\boldsymbol{\beta}) \mid \phi_{N_i}(W_i^D)\}$ and apply WLLN to each decomposition as above.

Now consider A_3. Under conditions (B2), (B6)–(B8) and assuming $\phi_{N_i}(W_i^D) \xrightarrow{p}$

$\phi(X_i^D)$, $\phi_{N_j}(W_j^{\overline{D}}) \xrightarrow{p} \phi(X_j^{\overline{D}})$ as $n \to \infty$ and $\Pr(\phi(X_i^D) = \phi(X_j^{\overline{D}})) = 0$, it follows that

$$\begin{aligned}
A_3 &= E\{S_{ij}(\boldsymbol{\beta})\} - E\{S_{Nij}(\boldsymbol{\beta})\} \\
&= E\left[\Psi_{ij}(U_{ij} - \theta_{ij}) - \Psi_{ij}(U_{Nij} - \theta_{ij})\right] \\
&\leq M \cdot E(U_{ij} - U_{Nij}), \quad \text{for some } 0 < M < \infty \\
&= M[\Pr\{\phi(X_i^D) > \phi(X_j^{\overline{D}})\} - \Pr\{\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})\}] \longrightarrow 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Hence, the lemma is proved.

Lemma C.2.3. *$E\{\partial S_{Nij}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^T\} - E\{\partial S_{Nij}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^T\} \to 0$ as $n \to \infty$ for $\boldsymbol{\beta} \in \mathcal{N}$.*

Proof: Under conditions (B1), (B2), (B6)–(B8) and assuming $\phi_{N_i}(W_i^D) \xrightarrow{p} \phi(X_i^D)$, $\phi_{N_j}(W_j^{\overline{D}}) \xrightarrow{p} \phi(X_j^{\overline{D}})$ as $n \to \infty$ and $\Pr(\phi(X_i^D) = \phi(X_j^{\overline{D}})) = 0$, it follows that

$$\begin{aligned}
&E\{\partial S_{Nij}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} - E\{\partial S_{ij}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T\} \\
&= E\left[\frac{\partial}{\partial\boldsymbol{\beta}^T}\left\{\Psi_{ij}(U_{Nij} - \theta_{ij})\right\} - \frac{\partial}{\partial\boldsymbol{\beta}^T}\left\{\Psi_{ij}(U_{ij} - \theta_{ij})\right\}\right] \\
&= \frac{\partial\Psi_{ij}}{\partial\boldsymbol{\beta}^T} \cdot E(U_{Nij} - U_{ij}) \\
&\leq M\{E(U_{ij}) - E(U_{Nij})\}, \quad \text{for some } 0 < M < \infty \\
&= M[\Pr\{\phi(X_i^D) > \phi(X_j^{\overline{D}})\} - \Pr\{\phi_{N_i}(W_i^D) > \phi_{N_j}(W_j^{\overline{D}})\}] \longrightarrow 0 \quad \text{as } n \to \infty.
\end{aligned}$$

Hence, the lemma is proved.

Lemma C.2.4. *For sufficiently large fixed N_i and N_j values, $\frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})$ is asymptotically equivalent to the expression*

$$\frac{1}{n_D n_{\overline{D}}} S_{Nn}^P(\boldsymbol{\beta}) = \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} \Psi_{ij}\{(\xi_{Ni} - \theta_{ij}) + (\xi_{Nj} - \theta_{ij})\},$$

where $\xi_{Ni} = E\{U_{Nij} \mid \phi_{N_i}(W_i^D), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}\}$ and $\xi_{Nj} = E\{U_{Nij} \mid \phi_{N_j}(W_j^{\overline{D}}), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}\}$. In fact, $S_{Nn}^P(\boldsymbol{\beta}) - S_{Nn}(\boldsymbol{\beta}) = o_p(n^{3/2})$.

Proof: Notice that $\frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})$ is a two-sample U-statistic under (B1). Then, it follows from the Projection Theorem of U-statistics (Serfling, 1980) that $\frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})$ and

$$\begin{aligned}
\frac{1}{n_D n_{\overline{D}}} S^P_{Nn}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} E\left\{ \frac{1}{n_D n_{\overline{D}}} \sum_{s=1}^{n} \sum_{j=1}^{n_{\overline{D}}} \Psi_{sj}(U_{Nsj} - \theta_{sj}) \mid \phi_{N_i}(W_i^D), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}} \right\} \\
&+ \sum_{j=1}^{n_{\overline{D}}} E\left\{ \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n} \sum_{k=1}^{n_{\overline{D}}} \Psi_{ik}(U_{Nik} - \theta_{ik}) \mid \phi_{N_j}(W_j^{\overline{D}}), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}} \right\} \\
&= \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n} \sum_{j=1}^{n_{\overline{D}}} \Psi_{ij} \big[E\left\{ (U_{Nij} - \theta_{ij}) \mid \phi_{N_i}(W_i^D), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}} \right\} \\
&+ E\left\{ (U_{Nij} - \theta_{ij}) \mid \phi_{N_j}(W_j^{\overline{D}}), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}} \right\} \big] \\
&= \frac{1}{n_D n_{\overline{D}}} \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\overline{D}}} \Psi_{ij} \{ (\xi_{Ni} - \theta_{ij}) + (\xi_{Nj} - \theta_{ij}) \},
\end{aligned}$$

are asymptotically equivalent. The convergence rate is established in Serfling (1980).

• Proof of consistency

By Theorem 1 of Dodd and Pepe (2003), the solution of $S_n(\boldsymbol{\beta}) = \mathbf{0}$ is consistent under (B1), (B3), (B4)–(B8) and Lemma C.2.1; that is, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$. Then, showing that $\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta})$ and $\frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})$ converge in probability to the same limit uniformly for $\boldsymbol{\beta} \in \mathcal{N}$ will ensure that the solution of $S_{Nn}(\boldsymbol{\beta}) = \mathbf{0}$ is consistent; that is, $\widehat{\boldsymbol{\beta}}_N \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$ (Li et al., 2016). Given (B3), we can first find a finite union of open intervals with known length that cover $\mathcal{N}$. Specifically, for $\zeta > 0$, define intervals $C_k = (\boldsymbol{\beta}_k, \boldsymbol{\beta}_{k+1})$ such that $|\boldsymbol{\beta}_{k+1} - \boldsymbol{\beta}_k| < \zeta$ for all k and $\mathcal{N} \subseteq \cup_{k-1}^{K} C_k$. The

triangle inequality then gives the following:

$$
\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{N}}\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta})-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta})\right| \\
&=\max_k \sup_{\boldsymbol{\beta}\in C_k}\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta})-\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)+\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta}_k)-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta})\right. \\
&\left.+\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta}_k)\right| \\
&\le \max_k \sup_{\boldsymbol{\beta}\in C_k}\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta})-\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)\right| \\
&+\max_k \sup_{\boldsymbol{\beta}\in C_k}\left|\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta}_k)-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta})\right| \\
&+\max_k \sup_{\boldsymbol{\beta}\in C_k}\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta}_k)\right| \\
&= A_{1n}+A_{2n}+A_{3n}
\end{aligned}
$$

The mean value theorem and uniform boundedness of $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}S_n(\boldsymbol{\beta})$ within $\mathcal{N}$ give the following result for A_{1n}:

$$
\begin{aligned}
A_{1n} &= \max_k \sup_{\boldsymbol{\beta}\in C_k}\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta})-\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)\right| \\
&= \max_k \sup_{\boldsymbol{\beta}\in C_k}(\boldsymbol{\beta}-\boldsymbol{\beta}_k)\cdot\left\{\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}S_n(\boldsymbol{\beta}^*)\right\}, \quad \text{for } \boldsymbol{\beta}^*\in(\boldsymbol{\beta}_k,\boldsymbol{\beta}) \\
&\le \zeta M_1, \quad \text{where } M_1<\infty
\end{aligned}
$$

Similarly, for fixed N_i and N_j values. the mean value theorem and uniform boundedness of $\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial\boldsymbol{\beta}}S_{Nn}(\boldsymbol{\beta})$ within $\mathcal{N}$ imply $A_{2n}<\zeta M_2$, where $M_2<\infty$. Finally, Lemma C.2.2 implies that for any $\epsilon>0$ and any $\gamma>0$ there exists a number n_0 such that for all $n\ge n_0$

$$
\Pr\left\{\left|\frac{1}{n_D n_{\overline{D}}}S_n(\boldsymbol{\beta}_k)-\frac{1}{n_D n_{\overline{D}}}S_{Nn}(\boldsymbol{\beta}_k)\right|>\frac{\epsilon}{2}\right\}<\frac{\gamma}{K},
$$

for a given k. This implies that

$$\begin{aligned}
&\Pr\left(A_{3n} > \frac{\epsilon}{2}\right) \\
&= \Pr\left\{\max_k \sup_{\boldsymbol{\beta}\in C_k} \left|\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}_k) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}_k)\right| > \frac{\epsilon}{2}\right\} \\
&= \Pr\left\{\max_k \left|\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}_k) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}_k)\right| > \frac{\epsilon}{2}\right\} \\
&< \sum_k \Pr\left\{\left|\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}_k) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}_k)\right| > \frac{\epsilon}{2}\right\} \\
&< \sum_k \frac{\gamma}{K} = \gamma.
\end{aligned}$$

Now choose an arbitrarily small ζ such that $\zeta(M_1 + M_2) < \epsilon/2$. It then follows that

$$\Pr\left\{\sup_{\boldsymbol{\beta}\in\mathcal{N}} \left|\frac{1}{n_D n_{\overline{D}}} S_n(\boldsymbol{\beta}) - \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta})\right| > \epsilon\right\} \leq \Pr\left(A_{1n} + A_{2n} + A_{3n} > \epsilon\right) < \gamma.$$

Therefore, $\widehat{\boldsymbol{\beta}}_N$ is a consistent estimator of $\boldsymbol{\beta}_0$.

• Proof of asymptotic normality

By Taylor's expansion, we obtain the following expression:

$$\begin{aligned}
0 &= \sqrt{\frac{n_D n_{\overline{D}}}{n}} \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\widehat{\boldsymbol{\beta}}_N) \\
&= \sqrt{\frac{n_D n_{\overline{D}}}{n}} \frac{1}{n_D n_{\overline{D}}} S_{Nn}(\boldsymbol{\beta}_0) + \frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}^T} S_{Nn}(\widetilde{\boldsymbol{\beta}}) \sqrt{\frac{n_D n_{\overline{D}}}{n}} (\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0),
\end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}$ is an intermediate value between $\widehat{\boldsymbol{\beta}}_N$ and $\boldsymbol{\beta}_0$. By applying Lemma C.2.2 of Dodd and Pepe (2003), it can be shown that $\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}^T} S_{Nn}(\boldsymbol{\beta})$ - $E\{\frac{\partial S_{Nij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\}$ converges in probability to 0 as $n \to \infty$ for $\boldsymbol{\beta} \in \mathcal{N}$, given sufficiently large fixed N_i and N_j values. This in turn implies, by Lemma C.2.3 and the Slutsky's theorem, that $\frac{1}{n_D n_{\overline{D}}} \frac{\partial}{\partial \boldsymbol{\beta}^T} S_{Nn}(\boldsymbol{\beta})$ converges in probability to $E\{\frac{\partial S_{ij}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\}$ for $\boldsymbol{\beta} \in \mathcal{N}$. Then since $\widehat{\boldsymbol{\beta}}_N$ is

a consistent estimator of $\boldsymbol{\beta}_0$ (see above), the following convergence result follows:

$$-\frac{1}{n_D n_{\overline{D}}}\frac{\partial}{\partial \boldsymbol{\beta}^T} S_{Nn}(\boldsymbol{\beta}) \xrightarrow{p} -E\left\{\frac{\partial S_{ij}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\right\} \equiv \mathbf{Q}.$$

Using this result and applying Lemma C.2.4, we can write:

$$\begin{aligned}
\sqrt{\frac{n_D n_{\overline{D}}}{n}}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) &= \mathbf{Q}^{-1}\sqrt{\frac{n_D n_{\overline{D}}}{n}}\frac{1}{n_D n_{\overline{D}}} S^P_{Nn}(\boldsymbol{\beta}_0) + o_p(1) \\
&= \mathbf{Q}^{-1}\sqrt{\frac{n_D n_{\overline{D}}}{n}}\frac{1}{n_D n_{\overline{D}}}\sum_{i=1}^{n_D}\sum_{j=1}^{n_{\overline{D}}}\Psi_{ij}\{(\xi_{Ni}-\theta_{0ij})+(\xi_{Nj}-\theta_{0ij})\} \\
&\quad + o_p(1) \\
&= \mathbf{Q}^{-1}\left[\frac{1}{\sqrt{n_D}}\sum_{i=1}^{n_D}\left\{\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n_{\overline{D}}}}\sum_{j=1}^{n_{\overline{D}}}\Psi_{ij}\left(\xi_{Ni}-\theta_{0ij}\right)\right\} + o_p(1)\right. \\
&\quad \left. + \frac{1}{\sqrt{n_{\overline{D}}}}\sum_{j=1}^{n_{\overline{D}}}\left\{\frac{1}{\sqrt{n}}\frac{1}{\sqrt{n_D}}\sum_{i=1}^{n_D}\Psi_{ij}\left(\xi_{Nj}-\theta_{0ij}\right)\right\}\right] \\
&= \mathbf{Q}^{-1}\left\{\frac{1}{\sqrt{n_D}}\sum_{i=1}^{n_D}V_{Ni} + \frac{1}{\sqrt{n_{\overline{D}}}}\sum_{j=1}^{n_{\overline{D}}}V_{Nj}\right\} + o_p(1)
\end{aligned}$$

where

$$\begin{aligned}
\theta_{0ij} &= g^{-1}(\mathbf{Z}\boldsymbol{\beta}_0) = E(\xi_{Ni}) = E(\xi_{Nj}) \\
V_{Ni} &= \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n_{\overline{D}}}}\sum_{j=1}^{n_{\overline{D}}}\Psi_{ij}\left(\xi_{Ni}-\theta_{0ij}\right) \\
V_{Nj} &= \frac{1}{\sqrt{n}}\frac{1}{\sqrt{n_D}}\sum_{i=1}^{n_D}\Psi_{ij}\left(\xi_{Nj}-\theta_{0ij}\right).
\end{aligned}$$

Then, by applying a central limit theorem for triangular arrays to respective sums of mean-zero independent random variables V_{Ni} and V_{Nj} (Greene, 2011), we can show that

$$\sqrt{\frac{n_D n_{\overline{D}}}{n}}(\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0},\, \mathbf{Q}^{-1}\boldsymbol{\Sigma}\mathbf{Q}^{-1}\right) \quad \text{as } n \to \infty,$$

where

$$\boldsymbol{\Sigma} = (1-\lambda)\lim_{n\to\infty}\left\{\frac{1}{n_D}\sum_{i=1}^{n_D}\frac{1}{n_{\overline{D}}^2}\sum_{j=1}^{n_{\overline{D}}}\sum_{l=1}^{n_{\overline{D}}}\Psi_{ij}\Psi_{il}^T\mathrm{Cov}(\xi_{Ni}^{(j)},\xi_{Ni}^{(l)})\right\}$$
$$+\lambda\lim_{n\to\infty}\left\{\frac{1}{n_{\overline{D}}}\sum_{j=1}^{n_{\overline{D}}}\frac{1}{n_D^2}\sum_{i=1}^{n_D}\sum_{k=1}^{n_D}\Psi_{ij}\Psi_{kj}^T\mathrm{Cov}(\xi_{Nj}^{(i)},\xi_{Nj}^{(k)})\right\},$$

with $\xi_{Ni}^{(j)} = E\{U_{Nij} \mid \phi_{N_i}(W_i^D), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}\}$, $\xi_{Ni}^{(l)} = E\{U_{Nij} \mid \phi_{N_i}(W_i^D), \mathbf{Z}_i^D, \mathbf{Z}_l^{\overline{D}}\}$, $\xi_{Nj}^{(i)} = E\{U_{Nij} \mid \phi_{N_j}(W_j^{\overline{D}}), \mathbf{Z}_i^D, \mathbf{Z}_j^{\overline{D}}\}$ and $\xi_{Nj}^{(k)} = E\{U_{Nij} \mid \phi_{N_j}(W_j^{\overline{D}}), \mathbf{Z}_k^D, \mathbf{Z}_j^{\overline{D}}\}$.

Appendix D

D.1 The EM Algorithm

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denote the collection of parameters and $\mathcal{G} = \{(\mathbf{w}_i, D_i, \hat{\boldsymbol{\zeta}}_i), i = 1, \ldots, n\}$ denote the complete data. The complete-data likelihood function is given by

$$L_c(\boldsymbol{\theta} \mid \mathcal{G}) = \prod_{i=1}^{n} \{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\}^{D_i} \{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}^{1-D_i},$$

from which the complete-data log-likelihood function can be derived as

$$l_c(\boldsymbol{\theta} \mid \mathcal{G}) = \sum_{i=1}^{n} D_i \ln\{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} + (1 - D_i) \ln\{(1 - \pi_i) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\},$$

where $g(\cdot \mid \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$ $(d = 0, 1)$ denotes the K-variate normal density of $\hat{\boldsymbol{\zeta}}_i$ with mean $\boldsymbol{\mu}_d$ and covariance $\boldsymbol{\Sigma}_d$ given the true disease status $D_i = d$.

E step

The expected value of the complete-data log-likelihood function $l(\boldsymbol{\theta} \mid \mathcal{G})$ with respect to the conditional distribution of latent data $\mathbf{D} = \{D_i, i = 1, \ldots, n\}$ given observed data $\mathbf{O} = \{(\mathbf{w}_i, \boldsymbol{\zeta}_i), i = 1, \ldots, n\}$ under the current estimate of the parameters

$\boldsymbol{\theta}^{(t)}$ can be calculated as

$$\begin{aligned} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) &= \mathrm{E}_{\mathbf{D}|\mathbf{O},\boldsymbol{\theta}^{(t)}}\{l_c(\boldsymbol{\theta} \mid \mathcal{G})\} \\ &= \sum_{i=1}^{n} \mathrm{E}_{\mathbf{D}|\mathbf{O},\boldsymbol{\theta}^{(t)}}(D_i)\ln\{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} \\ &+ \mathrm{E}_{\mathbf{D}|\mathbf{O},\boldsymbol{\theta}^{(t)}}(1-D_i)\ln\{(1-\pi_i) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\} \\ &= \sum_{i=1}^{n} P_i^{(t)}\ln\{\pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} + (1-P_i^{(t)})\ln\{(1-\pi_i) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)\}, \end{aligned}$$

where

$$\begin{aligned} P_i^{(t)} &= \mathrm{E}_{\mathbf{D}|\mathbf{O},\boldsymbol{\theta}^{(t)}}(D_i) = \Pr(D_i = 1 \mid \mathbf{w}_i, \hat{\boldsymbol{\zeta}}_i; \boldsymbol{\theta}^{(t)}) \\ &= \frac{\pi_i^{(t)} g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}{(1-\pi_i^{(t)}) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + \pi_i^{(t)} g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)}, \end{aligned}$$

with

$$\pi_i^{(t)} = \frac{\exp(\mathbf{w}_i^T \boldsymbol{\beta}^{(t)})}{1 + \exp(\mathbf{w}_i^T \boldsymbol{\beta}^{(t)})}.$$

M step

The updated estimate $\boldsymbol{\theta}^{(t+1)}$ can be obtained by maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$. Note that $\boldsymbol{\beta}$, $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ can be maximized independently since they all appear in separate linear terms.

To begin, consider $\boldsymbol{\beta}$:

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \arg\max_{\boldsymbol{\beta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} P_i^{(t)}\ln(\pi_i) + (1-P_i^{(t)})\ln(1-\pi_i) \\ &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} P_i^{(t)}\ln\Big\{\frac{\exp(\mathbf{w}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{w}_i^T\boldsymbol{\beta})}\Big\} + (1-P_i^{(t)})\ln\Big\{\frac{1}{1+\exp(\mathbf{w}_i^T\boldsymbol{\beta})}\Big\} \\ &= \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} P_i^{(t)}\mathbf{w}_i^T\boldsymbol{\beta} - \ln\{1+\exp(\mathbf{w}_i^T\boldsymbol{\beta})\}. \end{aligned}$$

This corresponds to a system of nonlinear equations whose solution cannot be derived algebraically. Thus, we propose to solve for $\boldsymbol{\beta}^{(t+1)}$ based on a numerical optimization technique, namely the Newton-Raphson method. Specifically, let $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_n]^T$ (i.e., covariate matrix) and $\mathbf{P}^{(t)} = [P_1^{(t)}, \ldots, P_n^{(t)}]^T$. It can be shown that the Newton-Raphson method can be implemented by first setting $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}_{<0>}$ as the initial value and then iteratively updating the beta coefficients (over s) using the formula

$$\boldsymbol{\beta}_{<s+1>} = \boldsymbol{\beta}_{<s>} + [\mathbf{W}^T \mathbf{H}_{<s>} \mathbf{W}]^{-1} \mathbf{W}^T (\mathbf{P}^{(t)} - \boldsymbol{\pi}_{<s>}),$$

where $\boldsymbol{\pi}_{<s>} = [\pi_{1<s>}, \ldots, \pi_{n<s>}]^T$ is the prevalence of each subject evaluated at the current estimate $\boldsymbol{\beta}_{<s>}$, and $\mathbf{H}_{<s>}$ denote a diagonal matrix with elements $\pi_{i<s>}(1 - \pi_{i<s>})$ on the diagonal and zeros everywhere else. The Newton-Rapshon algorithm continues until there is essentially no change between the elements of $\boldsymbol{\beta}$ from one iteration to the next, and the final estimate of $\boldsymbol{\beta}$ from this algorithm can be used as the updated estimate $\boldsymbol{\beta}^{(t+1)}$ in the EM algorithm.

For the next estimates of $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$:

$$\begin{aligned}(\boldsymbol{\mu}_1^{(t+1)}, \boldsymbol{\Sigma}_1^{(t+1)}) &= \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) = \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} \sum_{i=1}^{n} P_i^{(t)} \ln\{g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\} \\ &= \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} \sum_{i=1}^{n} P_i^{(t)} \big\{ -\frac{1}{2} |\boldsymbol{\Sigma}_1| - -\frac{1}{2} (\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_1) \big\},\end{aligned}$$

so that

$$\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^n P_i^{(t)} \hat{\boldsymbol{\zeta}}_i}{\sum_{i=1}^n P_i^{(t)}} \quad \text{and} \quad \boldsymbol{\Sigma}_1^{(t+1)} \frac{\sum_{i=1}^n P_i^{(t)} (\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_1^{(t+1)})(\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_1^{(t+1)})^T}{\sum_{i=1}^n P_i^{(t)}}.$$

Similarly, we can obtain:

$$\boldsymbol{\mu}_0^{(t+1)} = \frac{\sum_{i=1}^n (1 - P_i^{(t)}) \hat{\boldsymbol{\zeta}}_i}{\sum_{i=1}^n (1 - P_i^{(t)})}$$

and

$$\boldsymbol{\Sigma}_0^{(t+1)} = \frac{\sum_{i=1}^n (1 - P_i^{(t)})(\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_0^{(t+1)})(\hat{\boldsymbol{\zeta}}_i - \boldsymbol{\mu}_0^{(t+1)})^T}{\sum_{i=1}^n (1 - P_i^{(t)})}.$$

The E step and M step are repeated until the algorithm converges.

Choice of initial values

The speed of convergence of the EM algorithm and its ability to locate the global maximum depends on the choice of initial values (Karlis and Xekalaki, 2003). In our model, only $\boldsymbol{\beta}$ that determines the subject-specific mixture proportion needs to be given an informative initial value, as (possibly non-informative) initial values of other parameters are automatically updated based upon this value and observed data during the first iteration of the M-step. Accordingly, it is recommended to set good initial values for $\boldsymbol{\beta}$ based on *a priori* knowledge of the covariate effects on the disease. For other parameters, we can simply assign non-informative initial values (mean parameters) or begin with estimates obtained by the FPCA (covariance parameters); that is, $\boldsymbol{\mu}_1^{(1)} = \boldsymbol{\mu}_0^{(1)} = \mathbf{0}$ and $\boldsymbol{\Sigma}_1^{(1)} = \boldsymbol{\Sigma}_1^{(0)} = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$.

D.2 Standard Error Estimation

Firstly, we estimate the covariance matrix of the ML estimates, $\hat{\boldsymbol{\theta}}$, obtained in Appendix A by the inverse of the observed information matrix evaluated at $\hat{\boldsymbol{\theta}}$, that is, $\mathbf{I}_n^{-1}(\hat{\boldsymbol{\theta}})$. Since we have independent data, $\mathbf{I}_n(\hat{\boldsymbol{\theta}})$ can be approximated in terms of the gradient vector of the observed log-likelihood function (McLachlan and Basford, 1988). Specifically, let $f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta})$ denote the likelihood function based on the single observation vector $(\mathbf{w}_i, \boldsymbol{\zeta}_i)$; that is,

$$f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta}) = \pi_i g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1 - \pi_i) g(\hat{\boldsymbol{\zeta}}_i \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

Then the approximation to $\mathbf{I}_n(\hat{\boldsymbol{\theta}})$ is given by

$$\mathbf{I}_n(\hat{\boldsymbol{\theta}}) \approx \sum_{i=1}^{n} \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^T \tag{D.1}$$

where $\hat{\mathbf{h}}_i = \partial \ln\{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta})\}/\partial\boldsymbol{\theta}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ is the gradient vector of the log-likelihood based on the single observation evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

Define

$$\hat{\pi}_i = \frac{\exp(\mathbf{w}_i\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{w}_i\hat{\boldsymbol{\beta}})}, \quad \hat{\tau}_{1i} = \frac{\hat{\pi}_i g(\hat{\boldsymbol{\zeta}}_i \mid \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1)}{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \hat{\boldsymbol{\theta}})} \quad \text{and} \quad \hat{\tau}_{0i} = \frac{(1-\hat{\pi}_i) g(\hat{\boldsymbol{\zeta}}_i \mid \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)}{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \hat{\boldsymbol{\theta}})}.$$

and consider the partition of the gradient vector $\hat{\mathbf{h}}_i$:

$$\hat{\mathbf{h}}_i = [\hat{\mathbf{h}}_{i,\boldsymbol{\beta}}, \hat{\mathbf{h}}_{i,\boldsymbol{\mu}_1}, \hat{\mathbf{h}}_{i,\boldsymbol{\Sigma}_1}, \hat{\mathbf{h}}_{i,\boldsymbol{\mu}_0}, \hat{\mathbf{h}}_{i,\boldsymbol{\Sigma}_0}]^T,$$

The gradient vectors corresponding to the beta and mean parameters are ($d = 0, 1$)

$$\hat{\mathbf{h}}_{i,\boldsymbol{\beta}} = \left.\frac{\partial \ln\{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta})\}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \{\hat{\tau}_{1i}(1-\hat{\pi}_i) - \hat{\tau}_{0i}\hat{\pi}_i\}\mathbf{w}_i,$$

$$\hat{\mathbf{h}}_{i,\boldsymbol{\mu}_d} = \left.\frac{\partial \ln\{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta})\}}{\partial \boldsymbol{\mu}_d}\right|_{\boldsymbol{\mu}_d=\hat{\boldsymbol{\mu}}_d} = \hat{\tau}_{di}\hat{\boldsymbol{\Sigma}}_d^{-1}(\mathbf{w_i} - \hat{\boldsymbol{\mu}}_d),$$

respectively. If the mth element of $\hat{\mathbf{h}}_{i,\boldsymbol{\Sigma}_d}$ corresponds to differentiation with respect to $\sigma_{d,kk'}$ ($k \le k'$), then it takes the form of

$$\begin{aligned}(\hat{\mathbf{h}}_{i,\boldsymbol{\Sigma}_d})_m = \left.\frac{\partial \ln\{f(\hat{\boldsymbol{\zeta}}_i, \mathbf{w}_i \mid \boldsymbol{\theta})\}}{\partial \sigma_{d,kk'}}\right|_{\sigma_{d,kk'}=\hat{\sigma}_{d,kk'}} &= \frac{1}{2}\hat{\tau}_{di}(2-\delta_{kk'})[-(\hat{\boldsymbol{\Sigma}}_d^{-1})_{kk'} \\ &\quad + \{(\mathbf{w}_i - \hat{\boldsymbol{\mu}}_d)^T\hat{\boldsymbol{\sigma}}_{dk}^{-1}\}\{(\mathbf{w}_i - \hat{\boldsymbol{\mu}}_d)^T\hat{\boldsymbol{\sigma}}_{dk'}^{-1}\},\end{aligned}$$

where $\delta_{kk'}$ is Kronecker delta that equals 1 when $k = k'$ and 0 otherwise, $(\hat{\boldsymbol{\Sigma}}_d^{-1})_{kk'}$ is the (k, k')th element of $\hat{\boldsymbol{\Sigma}}_d^{-1}$, and $\hat{\boldsymbol{\sigma}}_{dk}^{-1}$ is the kth column of $\hat{\boldsymbol{\Sigma}}_d^{-1}$.

Now we move on to estimate the standard errors of AUC_k and cAUC. Let $\Phi'(x) = d\Phi(x)/dx$, I denote an identity matrix and "$\circ$" denote a hadamard (element-wise) product of matrices. The standard error of $\widehat{\text{AUC}}_k$ can be obtained using the observed information matrix $\mathbf{I}_n(\hat{\boldsymbol{\theta}})$ and the delta method as

$$\widehat{\text{SE}}(\widehat{\text{AUC}}_k)$$
$$= \left(\left[\frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right]^T \mathbf{I}_n^{-1}(\hat{\boldsymbol{\theta}}) \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^{\frac{1}{2}},$$

where elements of the vector $\frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\}$ are

$$\frac{\partial}{\partial \mu_{1k}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} = \Phi'\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \frac{1}{\sqrt{\sigma_{1,kk} + \sigma_{0,kk}}},$$

$$\frac{\partial}{\partial \sigma_{1,kk}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} = -\frac{1}{2} \Phi'\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) (\mu_{1k} - \mu_{0k})(\sigma_{1,kk} + \sigma_{0,kk})^{-\frac{3}{2}}$$

$$\frac{\partial}{\partial \mu_{1k}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} = -\Phi'\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \frac{1}{\sqrt{\sigma_{1,kk} + \sigma_{0,kk}}},$$

$$\frac{\partial}{\partial \sigma_{0,kk}} \left\{ \Phi\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) \right\} = -\frac{1}{2} \Phi'\left(\frac{\mu_{1k} - \mu_{0k}}{\sqrt{\sigma_{0,kk} + \sigma_{1,kk}}} \right) (\mu_{1k} - \mu_{0k})(\sigma_{1,kk} + \sigma_{0,kk})^{-\frac{3}{2}},$$

and 0 elsewhere.

The standard error of $\widehat{\text{cAUC}}$ can also be obtained using the observed information matrix and the delta method as

$$\widehat{\text{SE}}(\widehat{\text{cAUC}})$$
$$= \left\{ \left(\frac{\partial}{\partial \boldsymbol{\theta}} \left[\Phi\left\{ \sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} \right\} \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^T \mathbf{I}_n^{-1}(\hat{\boldsymbol{\theta}}) \cdot \frac{\partial}{\partial \boldsymbol{\theta}} \left[\Phi\left\{ \sqrt{\mathbf{a}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} \right\} \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right\}^{\frac{1}{2}},$$

where elements of the vector $\frac{\partial}{\partial \boldsymbol{\theta}}[\Phi\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\}]$ are:

$$\frac{\partial}{\partial \boldsymbol{\mu}_1}\Big[\Phi\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\Big] = \Phi'\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\cdot\frac{1}{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}}\cdot\boldsymbol{a},$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_1}\Big[\Phi\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\Big] = \Phi'\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\cdot\frac{1}{2\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}} \cdot\{-2\mathbf{a}\mathbf{a}^T+(\mathbf{a}\mathbf{a}^T\circ I)\}$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_0}\Big[\Phi\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\Big] = -\Phi'\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\cdot\frac{1}{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}}\cdot\boldsymbol{a},$$

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_0}\Big[\Phi\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\Big] = \Phi'\big\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\big\}\cdot\frac{1}{2\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}} \cdot\{-2\mathbf{a}\mathbf{a}^T+(\mathbf{a}\mathbf{a}^T\circ I)\}$$

and 0 elsewhere. Note that only the upper triangular elements of the $\frac{\partial}{\partial \boldsymbol{\Sigma}_d}[\Phi\{\sqrt{\boldsymbol{a}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}\}]$ $(d=0,1)$ should be used in the computation.

D.3 Estimation and Prediction for FPLS

Algorithm for estimating FPLS:

The following algorithm for estimating functional partial least squares (FPLS) has been proposed by Delaigle and Hall (2012). Consider n independent data pairs $\mathcal{X} = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$. For $i=1,\ldots,n$, first introduce the centered data $X_i^{[1]}(t) = X_i(t)-\hat{\mu}(t)$ and $Y_i^{[1]} = Y_i-\bar{Y}$, where $\hat{\mu}(t)=\sum_{i=1}^n X_i(t)/n$ and $\bar{Y} = \sum_{i=1}^n Y_i/n$. Then for $k=1,\ldots,K$, the algorithm iterates through the following three steps:

(1) Estimate the kth FPCA score as

$$\hat{\nu}_{ik} = \int_{\mathcal{T}} X_i^{[k]}(t)\delta_k(t)dt,$$

where the function $\delta_k \in L^2(\mathcal{T})$ is chosen maximize $\{\widehat{\text{Cov}}(Y_i^{[k]}, \hat{\nu}_{ik})\}^2$; that is,

$$\hat{\delta}_k(t) = \frac{\widehat{\text{Cov}}(Y_i, X_i^{[k]})}{||\widehat{\text{Cov}}(Y_i, X_i^{[k]})||} = \frac{\sum_{i=1}^{n} Y_i^{[k]} X_i^{[k]}(t)}{\Big[\int_{\mathcal{T}} \Big\{\sum_{i=1}^{n} Y_i^{[k]} X_i^{[k]}(t)\Big\}^2 dt\Big]^{1/2}}.$$

(2) Estimate β_k and the kth FPLS basis function $\hat{\rho}_k(t)$ as:

$$\hat{\beta}_k = \frac{\widehat{\text{Cov}}(Y_i^{k]}, \hat{\nu}_{ik})}{\widehat{\text{Var}}(\hat{\nu}_{ik})} = \frac{\sum_{i=1}^{n} Y_i^{[k]}\hat{\nu}_{ik}}{\sum_{i=1}^{n} \hat{\nu}_{ik}^2} \quad \text{and} \quad \hat{\rho}_k(t) = \frac{\widehat{\text{Cov}}(X_i^{k]}, \hat{\nu}_{ik})}{\widehat{\text{Var}}(\hat{\nu}_{ik})} = \frac{\sum_{i=1}^{n} X_i^{[k]}(t)\hat{\nu}_{ik}}{\sum_{i=1}^{n} \hat{\nu}_{ik}^2},$$

respectively.

(3) Calculate:

$$X_i^{[k+1]}(t) = X_i^{[k]}(t) - \hat{\rho}_k(t)\hat{\nu}_{ik} \quad \text{and} \quad Y_i^{[k+1]} = Y_i^{[k]} - \hat{\beta}_k\hat{\nu}_{ik}.$$

Note that numerical integration methods based on the observed time points $\{t_1, \ldots, t_N\}$ can be used to evaluate the integrals.

Prediction for a new subject

Suppose we want to predict the disease status of a new subject (not in the orginal dataset) with covariate $\mathbf{w}_{\text{new}}$ and functional biomarker measurements $\{X_{\text{new}}(t_j),\ t_j \in \mathcal{T},\ j = 1, \ldots, N\}$, but whose imperfect reference test result is not available. Firstly, introduce the demeaned functional measurements $X_{\text{new}}^{[1]}(t) = X_{\text{new}}(t) - \hat{\mu}(t)$, where the sample mean function $\hat{\mu}(t)$ is computed using the original training dataset. Then

for $k = 1, \ldots, K$, repeat the following two steps:

(1) Estimate the kth FPLS score as

$$\hat{\nu}_{\text{new},k} = \int_{\mathcal{T}} X_{\text{new}}^{[k]}(t)\hat{\delta}_k(t)dt,$$

where $\hat{\delta}_k$ is obtained from the above algorithm using the training dataset. Use the numerical integration methods based on the observed time points $\{t_1, \ldots, t_N\}$ can be used to evaluate the integral.

(2) Set

$$X_{\text{new}}^{[k+1]}(t) = X_{\text{new}}^{[k]}(t) - \hat{\rho}_k(t)\hat{\nu}_{\text{new},k}$$

where $\hat{\rho}_k$ is obtained from the above algorithm using the training dataset.

Then, we can combine these FPLS scores $\hat{\boldsymbol{\nu}}_{\text{new}} = [\hat{\nu}_{\text{new},1}, \ldots, \hat{\nu}_{\text{new},K}]^T$ to produce the new subject's composite test $\hat{\nu}^*_{\text{new}} = \hat{\mathbf{a}}^T\hat{\boldsymbol{\nu}}_{\text{new}}$, which can replace $\hat{\xi}^*_{\text{new}}$ in formula (5.16) to calculate the corresponding predictive probability of disease $\widehat{\Pr}(D_{\text{new}} = 1 \mid \mathbf{w}_{\text{new}}, \hat{\nu}^*_{\text{new}}; \hat{\boldsymbol{\theta}})$.

D.4 Parameter Setup for Simulation Settings

* Parameter setup for Setting 1.

Case 1:

$\boldsymbol{\mu}_1 = [1.3, 0.7, 0.4]^T$, $\boldsymbol{\mu}_0 = [-1.3, -0.7, -0.4]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.50 & -0.91 & -0.52 \\ -0.91 & 1.30 & -0.28 \\ -0.52 & -0.28 & 0.90 \end{bmatrix}.$$

Case 2:

$\boldsymbol{\mu}_1 = [1.3, 0.5, 0.4]^T$, $\boldsymbol{\mu}_0 = [-1.3, -0.5, -0.4]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.50 & -0.65 & -0.52 \\ -0.65 & 1.30 & -0.20 \\ -0.52 & -0.20 & 0.90 \end{bmatrix}.$$

Case 3:

$\boldsymbol{\mu}_1 = [1, 0.5, 0.4]^T$, $\boldsymbol{\mu}_0 = [-1, -0.5, -0.4]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.5 & -0.5 & -0.4 \\ -0.5 & 1.3 & -0.2 \\ -0.4 & -0.2 & 0.8 \end{bmatrix}.$$

* Parameter setup for Setting 2.

Case 1:

$\boldsymbol{\mu}_1 = [1.3, 0.7, 0.4, 0.5]^T$, $\boldsymbol{\mu}_0 = [-1.3, -0.7, -0.4, -0.5]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.50 & -0.91 & -0.52 & -0.65 \\ -0.91 & 1.30 & -0.28 & -0.35 \\ -0.52 & -0.28 & 0.90 & -0.20 \\ -0.65 & -0.35 & -0.20 & 0.70 \end{bmatrix}.$$

Case 2:

$\boldsymbol{\mu}_1 = [1.3, 0.5, 0.4, 0.5]^T$, $\boldsymbol{\mu}_0 = [-1.3, -0.5, -0.4, -0.5]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.50 & -0.65 & -0.52 & -0.65 \\ -0.65 & 1.30 & -0.20 & -0.25 \\ -0.52 & -0.20 & 0.90 & -0.20 \\ -0.65 & -0.25 & -0.20 & 0.70 \end{bmatrix}.$$

Case 3:

$\boldsymbol{\mu}_1 = [1, 0.5, 0.4, 0.4]^T$, $\boldsymbol{\mu}_0 = [-1, -0.5, -0.4, -0.4]^T$,

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_0 = \begin{bmatrix} 4.50 & -0.50 & -0.40 & -0.40 \\ -0.50 & 1.30 & -0.20 & -0.20 \\ -0.40 & -0.20 & 0.80 & -0.16 \\ -0.40 & -0.20 & -0.16 & 0.40 \end{bmatrix}.$$

* Parameter setup for Setting 3.

$\mu_{y,1} = 0.91$, $\mu_{y,0} = -0.91$, $\sigma^2_{y,1} = \sigma^2_{y,0} = 1 \implies \text{AUC}_y = 0.901$

$\mu_{y,1} = 0.6$, $\mu_{y,0} = -0.6$, $\sigma^2_{y,1} = \sigma^2_{y,0} = 1 \implies \text{AUC}_y = 0.802$

$\mu_{y,1} = 0.37$, $\mu_{y,0} = -0.37$, $\sigma^2_{y,1} = \sigma^2_{y,0} = 1 \implies \text{AUC}_y = 0.700$

$\mu_{y,1} = 0.18$, $\mu_{y,0} = -0.18$, $\sigma^2_{y,1} = \sigma^2_{y,0} = 1 \implies \text{AUC}_y = 0.600$

Bibliography

Albert, P. S. (2009). Estimating diagnostic accuracy of multiple binary tests with an imperfect reference standard. *Statistics in Medicine*, 28:780–797.

Albert, P. S. and Dodd, L. E. (2004). A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60:427–435.

Albert, P. S., McShane, L. M., and Shih, J. H. (2001). Latent class modeling approaches for assessing diagnostic error without a gold standard: with applications to p53 immunohistochemical assays in bladder tumors. *Biometrics*, 57:610–619.

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics*, 3:421–432.

Altman, D. G. and Bland, J. M. (1994). Diagnostic tests 2: Predictive values. *BMJ*, 309:102.

Altman, D. G. and Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080.

Arvesen, J. N. (1969). Jackknifing a U-statistic. *Annals of Mathematical Statistics*, 40:2076–2100.

Atkinson, J. and Nevill, A. (1997). Comment on the use of concordance correlation to assess the agreement between two variables. *Biometrics*, 53:775–777.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415.

Banhart, H. X. (2016). Assessing agreement with relative area under the coverage probability curve. *Statistics in Medicine*, 35:3153–3165.

Banhart, H. X., Haber, M., and Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58:1020–1027.

Banhart, H. X., Manatunga, A. K., and Williamson, J. M. (2001). Modeling concordance correlation via gee to evaluate reproducibility. *Biometrics*, 57(3):931–940.

Banhart, H. X., Song, J., and Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine*, 24:1371–1384.

Banhart, H. X., Yow, E., Crowley, A. L., Daubert, M. A., Rabineau, D., Bigelow, R., Pencina, M., and Douglas, P. S. (2016). Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Statistical Methods in Medical Research*, 25:2939–2958.

Bao, J., Manatunga, A., Binongo, J. N. G., and Taylor, A. T. (2011). Key variables for interpreting 99mTc-mercaptoacetyltriglycine diuretic scans: development and validation of a predictive model. *American Journal of Roentgenology*, 197(2):325–333.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19:3–11.

Beran, J., Feng, Y., Ghosh, S., and Kulik, R. (2013). *Long-memory processes: probabilistic properties and statistical methods*. Springer, Berlin, Heidelberg.

Berrendero, J., Justel, A., and Svarc, M. (2011). Principal components for multivariate functional data. *Computational Statistics & Data Analysis*, 55:2619–2634.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1:307–310.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8:135–160.

Brown, E. B., Iyer, H. K., and Wang, C. M. (1997). Tolerance intervals for assessing individual bioequivalence. *Statistics in Medicine*, 16:803–820.

Cai, T. (2004). Semiparametric ROC regression analysis with placement values. *Biostatistics*, 5:45–60.

Castro, P. E., Lawton, W. H., and Sylvestre, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics*, 28:329–337.

Chiou, J. M., Yang, Y. F., and Chen, Y. T. (2014). Multivariate functional principal component analysis: A normalization approach. *Statistica Sinica*, 24:1571–1596.

Choi, Y., Johnson, W. O., Collins, M. T., and Gardner, I. A. (2006a). Bayesian inferences for receiver operating characteristic curves in the absence of a gold standard. *Journal of Agricultural, Biological and Environmental Statistics*, 11:210–229.

Choi, Y. K., Johnson, W. O., and Thurmond, M. C. (2006b). Diagnosis using predictive probabilities without cut-offs. *Statistics in Medicine*, 25:699–717.

Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statstical Planning and Inference*, 138:1102–1115.

Choudhary, P. K. (2010). A unified approach for nonparametric evaluation of agreement in method comparison studies. *International Journal of Biostatistics*, 6:Article 19.

Chow, S. and Liu, J. P. (2008). *Design and Analysis of Bioavailability and Bioequivalence Studies, Third Edition*. Chapman & Hall/CRC, Boca Raton, FL.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:213–220.

Collins, J. and Albert, P. S. (2016). Estimating diagnostic accuracy without a gold standard: A continued controversy. *Journal of Biopharmaceutical Statistics*, 26:1078–1082.

Collins, J. and Huynh, M. (2014). Estimation of diagnostic test accuracy without full verification: a review of latent class methods. *Statistics in Medicine*, 33:4141–4169.

Craig, C. R. and Stitzel, R. E. (2004). *Modern pharmacology with clinical applications*. Lippincott Williams & Wilkins, Philadelphia, PA.

Dai, T., Guo, Y., Peng, L., and Manatunga, A. (2015). Nonparametric estimation of broad sense agreement between ordinal and censored continuous outcomes. Technical report, Emory University, Department of Biostatistics and Bioinformatics.

Delaigle, A. and Hall, P. (2012). Methodology and theory for partial least squares applied to functional data. *The Annals of Statistics*, 40:322–352.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dendukuri, N., Hadgu, A., and Wang, L. (2009). Modeling conditional dependence between diagnostic tests: A multiple latent variable model. *Statistics in Medicine*, 28:441–461.

Dendukuri, N. and Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, 57:158–167.

Dendukuri, N., Schiller, I., Joseph, L., and Pai, M. (2012). Bayesian meta-analysis of the accuracy of a test for tuberculous pleuritis in the absence of a gold standard reference. *Biometrics*, 68:1285–1293.

Dodd, L. E. and Pepe, M. S. (2003). Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*, 98:409–417.

Doornik, J. A. and Hansen, H. (2008). An omnibus test for univariate and multivariate normality*. *Oxford Bulletin of Economics and Statistics*, 70:927–939.

Efron, B. (1981). Nonparametric estimates of standard eror: the jackknife, the bootstrap and other methods. *Biometrika*, 68:589–599.

Erdogan, Z., Abdülrezzak, U., Silov, G., Özdal, A., and Turhal, O. (2014). Evaluation of interobserver variability of parenchymal phase of tc-99m mercaptoacetyltriglycine and tc-99m dimercaptosuccinic acid renal scintigraphy. *Indian Journal of Nuclear Medicine*, 29:87–91.

Eskild-Jensen, A., Gordon, I., Piepsz, A., and Frøkiær, J. (2004). Interpretation of the renogram: problems and pitfalls in hydronephrosis in children. *BJU International*, 94(6):887–892.

Esteves, F. P., Taylor, A., Manatunga, A., Folks, R. D., Krishnan, M., and Garcia, E. V. (2006). ^{99m}Tc-MAG3 renography: Normal values for MAG3 clearance and

curve parameters, excretory parameters, and residual urine volume. *American Journal of Roentgenology*, 187:W610–W617.

Farraggi, D. (2003). Adjusting receiver operating characteristics curves and related indices for covariates. *The Statistician*, 52:179–192.

Febrero-Bande, M., Galeano, P., and González-Manteiga, W. (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review*, 85:61–83.

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Florez, E., Fatemi, A., Claudio, P. P., and Howard, C. M. (2018). Emergence of radiomics: Novel methodology identifying imaging biomarkers of disease in diagnosis, response, and progression. *SM Journal of Clinical and Medical Imaging*, 4:1019.

Gasser, T., Kneip, A., and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *Journal of American Statistical Association*, 86:643–652.

Gasser, T. and Müller, H. G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, volume 757 of *Lecture Notes in Mathematics*, Berlin, Heidelberg. Springer.

Gasser, T. and Müller, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11:171–185.

Gonin, R., Lipsitz, S. R., Fitzmaurice, G. M., and Molenberghs, G. (2000). Regression modelling of weighted κ by using generalized estimating equations. *Journal of the Royal Statistical Society, Series C*, 49:1–18.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized*

Linear Models: A roughness penalty approach. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, London, UK.

Greene, W. (2011). *Econometric Analysis, Seventh Edition*. Prentice Hall, Upper Saddle River, NJ.

Hall, P., Müller, H. G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society, Series B*, 70:703–723.

Hall, P. and Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, 34:1493–1517.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113:649–659.

He, X., Fu, B., and Fung, W. K. (2003). Median regression for longitudinal data. *Statistics in Medicine*, 22:3655–3669.

Hua, K. and Wang, Y. (1981). *Applications of Number Theory to Numerical Analysis*. Springer, New York, NY.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36:167–171.

Hui, S. L. and Zhou, X. H. (1998). Evaluation of diagnostic tests without a gold standard. *Statistical Methods in Medical Research*, 7:354–370.

Imhof, J. P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48:419–426.

Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T. A., and Cadarso-Suárez, C. (2012). Extending induced ROC methodology to the functional context. *Biostatistics*, 13:594–608.

Jacques, J. and Preda, C. (2014). Model-based clustering for multivariate functional data. *Computational Statistics & Data Analysis*, 71:92–106.

Jafarzadeh, S. R., Johnson, W. O., and A., G. I. (2016). Bayesian modeling and inference for diagnostic accuracy and probability of disease based on multiple diagnostic biomarkers with and without a perfect reference standard. *Statistics in Medicine*, 35:859–876.

Jaksić, E., Beatović, S., Paunković, N., Stefanović, A., and Han, R. (2005). Variability in interpretation of static renal scintigraphy findings. *Vojnosanitetski pregled*, 62:189–193.

Janes, H. and Pepe, M. S. (2008). Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: An old concept in a new setting. *American Journal of Epidemiology*, 168:89–97.

Janes, H. and Pepe, M. S. (2009). Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*, 96:371–382.

Jang, J. H., Peng, L., and Manatunga, A. K. (2019). Assessing alignment between functional markers and ordinal outcomes based on broad sense agreement. *Biometrics*. in press.

Jones, G., Johnson, W. O., and Vink, W. D. (2009). Evaluating a continuous biomarker for infection by using observed disease status with covariate effects on disease. *Journal of the Royal Statistical Society, Series C*, 58:705–717.

Joseph, L., Gyorkos, T. W., and Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272.

Jung, S. (1996). Quasi-likelihood approach for median regression models. *Journal of the American Statistical Association*, 91:251–257.

Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590.

King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, 20:2131–2147.

Klar, N., Lipsitz, S. R., and G., I. J. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42:45–58.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to functional data analysis*. CRC Press, Boca Raton, FL.

Kotz, S., Johnson, N. L., and Boyd, D. W. (1967). Series representations of distributions of quadratic forms in normal variables II. non-central case. *Annals of Mathematical Statistics*, 38:838–848.

Kraemer, H. C. (1980). Extension of the kappa coefficient. *Biometrics*, 36:207–216.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. CRC Press, New York, NY.

Lee, J., Koh, D., and Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Computers in Biology and Medicine*, 19:61–70.

Leng, C. and Zang, W. (2014). Smoothing combined estimating equations in quantile regression for longitudinal data. *Statistics and Computing*, 24:123–136.

Li, K. and Luo, S. (2017). Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease. *Statistics in Medicine*, 36:3560–3572.

Li, R. and Chow, M. (2005). Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis*, 93:81–101.

Li, S., Sun, Y., Huang, C. Y., Follmann, D. A., and Krause, R. (2016). Recurrent event data analysis with intermittently observed time-varying covariates. *Statistics in Medicine*, 35:3049–3065.

Lin, L., Pan, Y., Hedayat, A. S., Banhart, H. X., and Haber, M. (2016). A simulation study of nonparametric total deviation index as a measure of agreement based on quantile regression. *Journal of Biopharmaceutical Statistics*, 26:937–950.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45:255–268.

Lin, L. I. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48:599–604.

Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, 19:255–270.

Lin, L. I., Hedayat, A. S., Sinha, B., and Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *Journal of the American Statistical Association*, 97:257–270.

Lin, L. I., Hedayat, A. S., and Wu, W. (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics*, 17(4):629–652.

Liu, B. and Müller, H. G. (2009). Estimating derivatives for samples of sparsely observed functions, with application to online auction dynamics. *Journal of the American Statistical Association*, 104:704–717.

Liu, D. and Zhou, X. H. (2013). Covariate adjustment in estimating the area under ROC curve with partially missing gold standard. *Biometrics*, 69:91–100.

Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis*, pages 853–856.

McLachlan, G. J. and Basford, K. E. (1988). *Mixture Models Inference and Applications to Clustering*. Marcel Dekker, Inc., New York, NY.

Mettler, F. A. and Guiberteau, M. J. (2012). *Essentials of nuclear medicine imaging*. Elsevier/Saunders, Philadelphia, PA.

Müller, H. G. (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, 12:766–774.

Müller, H. G. (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian Journal of Statistics*, 12:221–232.

O'Reilly, P., Aurell, M., Britton, K., Kletter, K., Rosenthal, L., and Testa, T. (1996). Consensus on diuresis renography for investigating the dilated upper urinary tract. *Journal of nuclear medicine*, 37:1872–1876.

Pearson, E. S. (1959). Note on an approximation to the distribution of non-central χ^2. *Biometrika*, 46:364.

Peng, L., Li, R., Guo, Y., and Manatunga, A. (2011). A framework for assessing broad sense agreement between ordinal and continuous measurements. *Journal of the American Statistical Association*, 106:1592–1601.

Peng, L. and Zhou, X. H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference*, 118:129–143.

Pepe, M. S. (1997). A regression modelling framework for ROC curves in medical diagnostic testing. *Biometrika*, 84:595–608.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*, 54:124–135.

Pepe, M. S. (2000). Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95:307–311.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classfication and prediction*. Oxford University Press, New York, NY.

Perez-Jaume, S. and Carrasco, J. L. (2015). A non-parametric approach to estimate the total deviation index for non-normal data. *Statistics in Medicine*, 34:3318–3335.

Preda, C. and Saporta, G. (2005). Pls regression on a stochastic process. *Computational Statistics & Data Analysis*, 48:149–158.

Provost, S. B. and Mathai, A. M. (1992). *Quadratic Forms in Random Variables: Theory and Applications/ A.M. Mathai, Serge B. Provost*. Statistics : textbooks and monographs. Marcel Dek ker.

Qu, Y., Tan, M., and Kutner, M. H. (1996). Random effects models in latent class analysis for evaluating accuracy of diagnostic tests. *Biometrics*, 52:797–810.

Rahman, A., Peng, L., Manatunga, A., and Guo, Y. (2017). Nonparametric regression method for broad sense agreement. *Journal of Nonparametric Statistics*, 29:280–300.

Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer, New York, NY.

Rathnayake, L. N. and Choudhary, P. K. (2016). Tolerance bands for functional data. *Biometrics*, 72:503–512.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:233–243.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J., and Cadarso-Suárez, C. (2011). ROC curve and covariates: extending the induced methodology to the non-parametic framework. *Statistics and Computing*, 21:483–485.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25:127–141.

Sen, P. K. (1977). Almost sure convergence of generalized U-statistics. *The Annals of Probability*, 5:287–290.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, New York, NY.

Shao, J. (1992). Jackknifing in generalized linear models. *Annals of the Institute of Statistical Mathematics*, 44:673–686.

Shao, J. (2003). *Mathematical Statistics, Second Edition*. Springer, New York, NY.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424):1350–1355.

Tang, C. Y. and Leng, C. (2011). Empirical likelihood and quantile regression in longitudinal data analysis. *Biometrika*, 89:1001–1006.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Taylor, A., Blaufox, M. D., De Palma, D., Dubovsky, E. V., Erbas, B., Eskild-Jensen, A., Frøkiær, J., Issa, M. M., Piepsz, A., and Prigent, A. (2012). Guidance document for structured reporting of diuresis renography. *Seminars in Nuclear Medicine*, 42:41–48.

Taylor, A., Charman, S. C., Lefere, P., McFarland, E. G., Paulson, E. K., Yee, J., Aslam, R., Barlow, J. M., Gupta, A., and Kim, D. H. (2008a). CT Colonography: investigation of the optimum reader paradigm by using Computer-aided Detection Software 1. *Radiology*, 246:463–471.

Taylor, A. and Garcia, E. V. (2014). Computer-assisted diagnosis in renal nuclear medicine: rationale, methodology, and interpretative criteria for diuretic renography. *Seminars in Nuclear Medicine*, 44:146–158.

Taylor, A., Garcia, E. V., Binongo, J. N. G., Manatunga, A., Halkar, R., Folks, R. D., and Dubovsky, E. (2008b). Diagnostic performance of an expert system for interpretation of 99mTc MAG3 scans in suspected renal obstruction. *Journal of Nuclear Medicine*, 49:216–224.

Taylor, A. T. (2014). Radionuclides in nephrourology, Part 2: pitfalls and diagnostic applications. *Journal of nuclear medicine*, 55(5):786–798.

Taylor, A. T., Manatunga, A., and Garcia, E. V. (2008c). Decision support systems in diuresis renography. *Seminars in Nuclear Medicine*, 38:67–81.

Thisted, R. A. (1988). *Elements of Statistical Computing: Numerical Computation*. Chapman and Hall, New York, NY.

Torrance-Rynard, V. L. and Walter, S. D. (1997). Effects of dependent errors in the assessment of diagnostic test performance. *Statistics in Medicine*, 16:2157–2175.

Vacek, P. M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41:959–968.

Wang, C., Turnbull, B. W., Gröhn, Y. T., and Nielsen, S. S. (2006). Estimating receiver operating characteristic curves with covariates when there is no perfect reference test for diagnosis of johne's disease. *Journal of Dairy Science*, 89:3038–3046.

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., and Alzheimer's Disease Neuroimaging Initiative (2012). The Alzheimer's disease neuroimaging initiative: A review of papers published since its inception. *Alzheimer's & Dementia*, 9:e111–e194.

Williamson, J. M., Manatunga, A. K., and Lipsitz, S. R. (2000). Modelling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1:191–202.

Xu, H. and Craig, B. A. (2009). A probit latent class model with general correlation structures for evaluating accuracy of diagnostic tests. *Biometrics*, 65:1145–1155.

Yan, F., Xiao, L., and Huang, X. (2017). Dynamic prediction of disease progression

for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *Annals of Applied Statistics*, 11:1649–1670.

Yang, I. and Becker, M. P. (1997). Latent variable modeling of diagnostic accuracy. *Biometrics*, 53(3):948–958.

Yao, F. and Lee, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 68:3–25.

Yao, F., Müller, H. G., Clifford, A. J., Dueker, S. R., Follett, J., Lin, Y., Buchholz, B. A., and Vogel, J. S. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59:676–685.

Yao, F., Müller, H. G., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590.

Yin, G. and Cai, J. (2005). Quantile regression models with multivariate failure time data. *Biometrics*, 61:151–161.

Yu, B., Zhou, C., and Bandinelli, S. (2011). Combining multiple continuous tests for the diagnosis of kidney impairment in the absence of a gold standard. *Statistics in Medicine*, 30:1712–1721.

Zhang, B., Chen, Z., and Albert, P. S. (2012). Estimating diagnostic accuracy of raters without a gold standard by exploiting a group of experts. *Biometrics*, 68:1294–1302.

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

Zhou, Y. and Sedransk, N. (2013). A new functional data-based biomarker for monitoring cardiovascular behavior. *Statistics in Medicine*, 32(1):153–164.

Zou, K. H., Hall, W. J., and Shapiro, D. E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–2156.