Published in final edited form as:

Biometrics. 2019 December ; 75(4): 1367–1379. doi:10.1111/biom.13063.

Assessing alignment between functional markers and ordinal outcomes based on broad sense agreement

Jeong Hoon Jang, Limin Peng, Amita K. Manatunga

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia

Abstract

Functional markers and their quantitative features (eg, maximum value, time to maximum, area under the curve [AUC], etc) are increasingly being used in clinical studies to diagnose diseases. It is thus of interest to assess the diagnostic utility of functional markers by assessing alignment between their quantitative features and an ordinal gold standard test that reflects the severity of disease. The concept of broad sense agreement (BSA) has recently been introduced for studying the relationship between continuous and ordinal measurements, and provides a promising tool to address such a question. Our strategy is to adopt a general class of summary functionals (SFs), each of which flexibly captures a different quantitative feature of a functional marker, and study its alignment according to an ordinal outcome via BSA. We further illustrate the proposed framework using three special classes of SFs (AUC-type, magnitude-specific, and time-specific) that are widely used in clinical settings. The proposed BSA estimator is proven to be consistent and asymptotically normal given a consistent estimator for the SF. We further provide an inferential framework for comparing a pair of candidate SFs in terms of their importance on the ordinal outcome. Our simulation results demonstrate satisfactory finite-sample performance of the proposed framework. We demonstrate the application of our methods using a renal study.

Keywords

alignment; broad sense agreement; curve data; functional marker; nonparametric estimation; ordinal outcome

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 κ coefficient or the weighted κ coefficient (Cohen, 1960; 1968; Fleiss, 1971; Kraemer, 1980); with paired continuous measurements, intraclass correlation coefficient (Bartko, 1966) and concordance correlation coefficient (CCC; Lin, 1989) are popular measures of agreement. Some raters, however, may use different measurement

Correspondence: Amita K. Manatunga, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, Georgia 30322. amanatu@emory.edu.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

processes with distinctive point systems, resulting in paired measurements with different scales (eg, 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 BSA (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 prespecified 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 (renal obstruction) is a serious clinical problem that can lead to irreversible loss of renal function if not properly treated. In the diagnosis of renal 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., 2008). 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 1). Second renogram curves (called postfurosemide) are also obtained with an additional 20 minutes after an intravenous injection of furosemide, a potent diuretic (see the right panel in Figure 1). In the absence of a gold standard for the presence of renal obstruction, consensus ordinal ratings on each subject's obstruction status (1: nonobstructed; 2: equivocal; and 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 renal 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 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 (eg, 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 renal 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 renal 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* (SFs), 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 SFs 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 article is organized as follows. In Section 2, we first review the existing BSA framework, followed by an introduction of a general class of SFs 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 (CIs) are also presented. In Section 3, we illustrate the proposed framework based on BSA using several concrete classes of SFs. In Section 4, we describe the inferential framework for comparing SFs regarding their alignment with the ordinal outcomes. In Section 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 6. Finally, we conclude with some remarks in Section 7.

2 | METHODS

2.1 | Review of BSA

The concept of 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 BSA (or disagreement) between X and Y is defined as the existence of an increasing (or decreasing) step function $\mathcal{V}$ from $\mathcal{D}_X$ and $\mathcal{D}_Y$ such that $Y = \mathcal{V}(X)$ with probability of 1. That is, if $X_{(*k)}$ denotes the randomly selected X given $Y = k (k = 1, 2, \dots, K)$, a perfect BSA (disagreement) case implies $X_{(*1)} < X_{(*2)} < \dots < X_{(*K)}$ ($X_{(*1)} > X_{(*2)} > \dots > X_{(*K)}$) 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 perfect BSA among a group of continuous measurements. Specifically, denote the ranks of $\{X_{(*1)}, X_{(*2)}, \dots, X_{(*K)}\}$ by $\{R_1, R_2, \dots, R_K\}$. Then the proposed BSA measure is defined as

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

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 BSA (disagreement). A value close to 0 indicates independence between X and Y .

A nonparametric 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 $\{R_1, \dots, R_K\}$ which consists of $K!$ permutations of the elements of $\vec{k} = [1, \dots, K]'$. Suppose the observed data consist of n pairs of (X_i, Y_i) , $i = 1, \dots, n$. 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 ($1 \leq s_k \leq n_k$) continuous measurement among those that fall into the k th ordinal category. Denote $\vec{R}(\cdot)$ as a mapping from $\mathcal{D}_X^K$ to Θ_K : $\vec{R}(\vec{x}) = [r_1, \dots, r_K]'$, where $\vec{x} = [x_1, \dots, x_K]' \in \mathcal{D}_X^K$ and $r_k = \sum_{m=1}^K I(x_k \geq x_m)$ representing the rank of x_k among $\{x_1, \dots, x_K\}$. Then it can be shown that the estimator takes the form of (Peng et al., 2011)

$$\hat{\rho}_{\text{bsa}}(X, Y) = 1 - \frac{\left(\prod_{k=1}^K n_k\right)^{-1} \sum_{s_1=1}^{n_1} \dots \sum_{s_K=1}^{n_K} \|\vec{k} - \vec{R}(X_{(*1), s_1}, \dots, X_{(*K), s_K})\|^2}{(K^3 - K)/6}, \quad (2)$$

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

2.2 | General formulation of the SF

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 $\mathcal{F}_\omega = \{f: \mathcal{T} \rightarrow \mathbb{R}; f \text{ is square integrable and } \omega\text{-times continuously differentiable at any } t \in \mathcal{T}\}$, and assume $X \in \mathcal{F}_\omega$. A general formulation of the SF of X is simply defined as a map from $\mathcal{F}_\omega$ to $\mathbb{R}$, that is, $\phi: \mathcal{F}_\omega \rightarrow \mathbb{R}$. Our approach is to consider this general formulation of the SF that encompasses a wide class of quantitative features of a functional marker; we defer discussion of its specific examples until Section 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 SF. We see that the index is now essentially based on the comparison between the ranks of $\{\phi(X_{(*1)}), \phi(X_{(*2)}), \dots, \phi(X_{(*K)})\}$ and their anticipated ranks under the perfect BSA scenario $(1, 2, \dots, K)$; see (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 SF $\phi(X)$ and the ordinal outcome Y .

2.4 | Nonparametric estimation

Suppose that n functional markers $X_1, X_2, \dots, X_n$ are directly observable. Then, given a reasonable choice of a quantitative feature to be analyzed, this in turn implies that n SFs $\phi(X_1), \phi(X_2), \dots, \phi(X_n)$ 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 (2) with X replaced by $\phi(X)$.

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

$$W_i(t_{ij}) = X_i(t_{ij}) + \epsilon_i(t_{ij}), \quad (3)$$

where the random measurement error $\epsilon_i(t)$ follows an independent and identical distribution with $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma_{\epsilon}^2 < \infty$ for each t , mutually independent of the true function X_i . We assume that $N_i \rightarrow \infty$ as $n \rightarrow \infty$ 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 SF $\phi(X_i)$ is unknown but can be estimated based on the observed data as $\phi_{N_i}(W_i)$. For instance, a SF 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 SFs of interest and their respective ordinal outcomes $\{\phi_{N_1}(W_1), Y_1\}, \{\phi_{N_2}(W_2), Y_2\}, \dots, \{\phi_{N_n}(W_n), Y_n\}$.

We propose to estimate $\rho_{\text{bsa}}(\phi(X), Y)$ by $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$, which, under formulation (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 nonparametric estimator takes the form of

$$\hat{\rho}_{\text{bsa}}(\phi_N(W), Y) = 1 - \frac{\left(\prod_{k=1}^K n_k \right)^{-1} \sum_{s_1=1}^{n_1} \dots \sum_{s_K=1}^{n_K} \|\vec{\mathbf{k}} - \vec{\mathbf{R}}\{\phi_N(W_{(*1), s_1}), \dots, \phi_N(W_{(*K), s_K})\}\|^2}{(K^3 - K)/6}, \quad (4)$$

where $\vec{\mathbf{k}} = [1, \dots, K]'$ and $\vec{\mathbf{R}}(\cdot)$ is mapping from $\mathcal{D}_{\phi}^K$ to Θ_K : $\vec{\mathbf{R}}(\vec{\mathbf{w}}) = [r_1, \dots, r_K]'$, with $\vec{\mathbf{w}} = [\phi_N(w_{(*1)}), \dots, \phi_N(w_{(*K)})]' \in \mathcal{D}_{\phi}^K$ 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 $\{\phi_N(w_{(*1)}), \dots, \phi_N(w_{(*K)})\}$. Note that all the subscripts of N have been dropped in the right-hand side of the Equation (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 $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ given by (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 SFs and their anticipated ranks under the scenario of perfect BSA. The second source of error comes from replacing the true SFs $\phi(X)$ 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 (4).

2.5 | Asymptotic properties

In Theorem 1, we establish the consistency and asymptotic normality of the proposed estimator $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ given by (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 Supporting Information S1.

Theorem 1.—Suppose $\sup_{1 \leq i \leq n} |\phi_{N_i}(W_i) - \phi(X_i)| \leq O_p(P_N)$, where P_N is a nonnegative sequence. (a) If $P_N \rightarrow 0$ as $n \rightarrow \infty$ and the regularity conditions A1 to A3 provided in Supporting Information S1 hold, $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ is a consistent estimator for $\rho_{\text{bsa}}(\phi(X), Y)$. (b) If $\sqrt{n}P_N \rightarrow 0$ as $n \rightarrow \infty$ and the regularity conditions A1, A2, and A4 provided in Supporting Information S1 hold, when $|\rho_{\text{bsa}}(\phi(X), Y)| < 1$, $\sqrt{n}\{\hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi(X), Y)\}$ has an asymptotic normal distribution with mean zero and variance σ_{bsa}^2 , where σ_{bsa}^2 is defined in Supporting Information S1.

The key idea of the proof is to consider the decomposition

$\hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi(X), Y) = T_1 + T_2$, where $T_1 = \hat{\rho}_{\text{bsa}}(\phi_N(W), Y) - \rho_{\text{bsa}}(\phi_N(W), Y)$ 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 $n \rightarrow \infty$. T_2 can be shown negligible provided that $\sup_{1 \leq i \leq n} |\phi_{N_i}(W_i) - \phi(X_i)| \leq O_p(P_N)$, where P_N and $\sqrt{n}P_N$ approach 0 as n goes to infinity.

2.6 | Estimation of standard error and CI

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

$$\hat{s}_J^2 = \frac{n-1}{n} \sum_{i=1}^n \left\{ \hat{\rho}_{\text{bsa}}^{(-i)}(\phi_N(W), Y) - \frac{1}{n} \sum_{p=1}^n \hat{\rho}_{\text{bsa}}^{(-p)}(\phi_N(W), Y) \right\}^2. \quad (5)$$

Note that the validity of the jackknife formula (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 2.5 under Theorem 1) are asymptotically negligible given a consistent estimator of $\phi(X)$. Furthermore, other nonparametric 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 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 $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ to asymptotic normality, especially when $\hat{\rho}_{\text{bsa}}(\phi_N(W), Y)$ 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

$$\left[g^{-1}(\tilde{g} - z_{1-\alpha/2} \cdot \tilde{g}' \hat{s}_J), g^{-1}(\tilde{g} + z_{1-\alpha/2} \cdot \tilde{g}' \hat{s}_J) \right], \quad (6)$$

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

3 | ILLUSTRATION OF THE PROPOSED BSA FRAMEWORK

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

3.1 | Three special cases of SFs

Suppose that the functional marker X is ω -times continuously differentiable (see Section 2.2). We denote $X^{(\nu)}$ as its ν th derivative ($0 \leq \nu \leq \omega - 2$), with $X^{(0)} = X$.

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

$$\phi(X) := \phi_{\text{AUC}}^{[\nu]}(X) = \int_{\mathcal{T}} X^{(\nu)}(t) dt.$$

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.

3.1.2 | 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

$$\phi(X) := \phi_{\text{MAG}(t^*)}^{[\nu]}(X) = X^{(\nu)}(t^*).$$

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

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

3.1.3 | 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, that is, $\phi: \mathcal{F}_{\omega} \rightarrow \mathcal{T} \subset \mathbb{R}$:

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

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

$$\phi(X) := \phi_{\text{tMAX}}^{[\nu]}(X) = \operatorname{argsup}_{t \in \mathcal{T}} X^{(\nu)}(t).$$

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

The interpretation of a perfect BSA scenario (ie, $\rho_{\text{bsa}}(\phi(X), Y) = 1$) differs depending on a chosen SF. For instance, if $\phi_{\text{AUC}}(X)$ is adopted, a perfect BSA scenario implies ϕ_{AUC}

$(X_{(*1)}) < \dots < \phi_{AUC}(X_{(*K)})$. 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 SFs.

3.2 | Nonparametric estimation of the special-case SFs

Assume without loss of generality that $\mathcal{T} = [0, 1]$. As illustrated in Model (3), we do not observe the true functional marker X_i ($i = 1, \dots, n$) in practice, but collect its realized values at N_i discrete times points $0 = t_{i1} < t_{i2} < \dots < t_{iN_i} = 1$ 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 (eg, kernel smoothing, spline, moving average, and so on). For instance, a popular approach is to use smoothing splines (eg, 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 nonparametric smoothed estimate of the true underlying function $X_i^{(v)}$ using the following kernel estimator (Gasser and Müller, 1979; 1984; Müller, 1984; 1985):

$$\widehat{W}_i^{[v]}(t) = \frac{1}{b_{N_i}^{v+1}} \sum_{j=1}^{N_i} \int_{d_{ij}-1}^{d_{ij}} K_v\left(\frac{t-u}{b_{N_i}}\right) du \cdot W_i(t_{ij}), \quad (7)$$

where $d_{i,j-1} = (t_{ij} + t_{i,j-1})/2$ ($d_{i0} = 0, d_{N_i} = 1$) is a smoothing parameter (bandwidth) satisfying $b_{N_i} \rightarrow 0, N_i b_{N_i} \rightarrow \infty$ as $N_i \rightarrow \infty$ for all i . K_v is a kernel function of order (v, ω) defined on a compact support $[-1, 1]$ and takes on zero values on the boundary points (see Supporting Information S2). 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 nonparametric estimator $\phi_{N_i}(W_i)$ for each of the three special-case SFs $\phi(X_i)$ based on the Gasser-Müller kernel estimator (7) as following.

3.2.1 | AUC-type functionals—This type of SF can be estimated by the following Riemann sum of $\widehat{W}_i^{[v]}$ with respect to the output design points $\{t_{i1}, \dots, t_{iN_i}\}$:

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

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

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

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

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

3.2.3 | Time-specific functionals—Assume that $X_i^{(v)}$ attains a certain threshold value η . Then, its timing can be nonparametrically estimated by its empirical counterpart of $\widehat{W}_i^{[v]}$:

$$\phi_{N_i}(W_i) = \phi_{N_i, \text{TIME}(\eta)}^{[v]}(W_i) = \inf \left\{ t \in \mathcal{T} : \widehat{W}_i^{[v]}(t) = \eta \right\}.$$

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

$$\phi_{N_i}(W_i) = \phi_{N_i, t\text{MAX}}^{[v]}(W_i) = \inf \left\{ t \in \mathcal{T} : \widehat{W}_i^{[v]}(t) = \sup_{u \in \mathcal{T}} \widehat{W}_i^{[v]}(u) \right\}.$$

In Supporting Information S3, we provide and prove a theorem that states the consistency of each of the above estimators. Then by Theorem 1 from Section 2.5, this in turn establishes the consistency and asymptotic normality of the corresponding nonparametric BSA estimator.

4 | STATISTICAL TEST FOR SELECTING A SF

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 testing 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 SF $\phi_1(X)$ leads to a significantly better alignment with an ordinal outcome than that of a competing SF $\phi_2(X)$. For simplicity, let $\rho_{\text{bsa},1}$ and $\rho_{\text{bsa},2}$ denote the true BSA measures based on two different SFs $\phi_1(X)$ 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

$$T_{N,n} = \frac{\sqrt{n}\{g(\hat{\rho}_{\text{bsa},1}) - g(\hat{\rho}_{\text{bsa},2})\}}{\hat{V}_J} = \frac{D_{N,n}}{\hat{V}_J} \xrightarrow{d} N(0,1), \quad (8)$$

where $g(a) = 0.5 \times \ln\{(1+a)(1-a)\}$ (Fisher's Z -transformation) and $\hat{V}_J$ 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 $\hat{D}_{N,n}^{(-i)}$ 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

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

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

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 SFs (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 SF to be analyzed. For the AUC-type SFs, 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 constant and improving degrees of alignment in terms of the crude AUC over the time interval, respectively. Performances based on the magnitude-specific SFs 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 $\text{Cov}(X(s), X(t)) = \exp\{-(s-t)^2\}$, $s, t \in \mathcal{T}$. We consider two scenarios for evaluating the finite-sample performance based on the time-specific SFs. 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 2 illustrates the representative curve sample (one for each ordinal category), the type(s) of SF 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 ($t_0 = 0$ 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 $\mathcal{T}_{\text{grid}} = \{(u-1)/119, u = 1, \dots, 120\}$ separately for each subject.

In order to mimic as closely as possible a real situation, we further contaminate the generated functional markers based on Model (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 (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 Formulas (5) and (6), respectively. Results presented in Table 1 are based on 1000 simulated datasets of size $n = 40$ and 60.

From Table 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-specific and time-specific SFs when the data are highly sparse; see cases (a) and (c). Therefore, when magnitude-specific or time-specific SF is considered, we recommend using functional markers that are collected on at least average of 25 time points to produce reliable 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 SF. 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 4. Specifically, empirical rejection rates of H_0 are obtained under the two selected scenarios from above and are presented in Table 1 of Supporting Information S4. 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 1 is repeated at the level of the first derivative of functional markers, and its results are presented in Table 2 of Supporting Information S4.

6 | RENAL STUDY

In this section, we apply the proposed approaches to the motivating renal study data described in Section 1. In the absence of a gold standard for detection of renal 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.*, 2008). 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 postfurosemide renogram curves to assist readers arrive at correct diagnosis of renal obstruction (Taylor *et al.*, 2008; Bao *et al.*, 2011), and it is of ongoing interest to rigorously establish their connection with the underlying obstruction mechanism to prevent inappropriate patient management and unnecessary surgery.

The study was thus designed to assess and improve the diagnostic utility of the baseline and postfurosemide 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 renal 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 “nonobstructed” ($Y = 1$), 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 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 (postfurosemide) renogram curve was obtained with an additional 20 minutes (see the right panel of Figure 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.*'s (2011) 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 ($\phi_{\text{AUC}}^{[1]}$); (c) time to maximum of the crude curve (ϕ_{MAX}); and (d) minimum rate of change ($\phi_{\text{MIN}}^{[1]}$). Furthermore, Eskild-Jensen *et al.* (2004) showed that MAG3 accumulates in the ROI without any excretion for the first 2 to 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 subtime intervals, $\mathcal{T}_1 = [0, 10]$ and $\mathcal{T}_2 = [10, 24]$, dichotomized at the 10-minute milestone. For the postfurosemide renogram curves, Bao *et al.* (2011) suggests the importance of their overall MAG3 intensity in detecting renal 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 (7) of the crude ($\nu = 0$) renogram curves and their first derivatives ($\nu = 1$) using a polynomial kernel of degrees 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 2 presents the BSA estimates between the four selected 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 subtime 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 subtime interval level unveils a noticeably better alignment of the first-derivative AUCs evaluated on $\mathcal{T}_2(\phi_{\text{AUC}, \mathcal{T}_2}^{[1]})$, 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 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 $P < 0.001$). 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 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 presents the BSA estimates between the two selected 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 = 0.57–0.84) and right (estimated BSA = 0.55; 95% CI = 0.37–0.69) kidneys. Maximum values are also well aligned with the expert ratings, but its degree falls short of that of the crude AUCs (both $P < 0.01$).

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

7 | DISCUSSION

In this article, 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 SFs 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 SFs 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 renal 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), $C_{\max}$ (the peak plasma concentration), and $t_{\max}$ (time to reach $C_{\max}$; Craig and Stitzel, 2004). In summary, the nature of a scientific experiment can guide the choice of SFs while providing sensible interpretations for them.

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 SF, 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 4. That is, one can use a Wald-type test statistic (8) based on the two BSA estimates computed for the two gender groups. Recently, Rahman *et al.* (2017) proposed a nonparametric 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 SFs 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 SFs to reduce dimension and maximize prediction performance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research was supported by NIH grants 1R01DK108070-01A1 from the National Institute of Diabetes & Digestive & Kidney Disease and R01 HL 113548 from National Heart, Lung, and Blood Institute. We also thank the coeditor, the associate editor, and an anonymous referee for their constructive comments.

Funding information

National Heart, Lung, and Blood Institute, Grant/Award Number: R01 HL 113548; National Institute of Diabetes and Digestive and Kidney Diseases, Grant/Award Number: 1R01DK108070-01A1

REFERENCES

- Arvesen JN (1969). Jackknifing a U-statistic. *Annals of Mathematical Statistics*, 40, 2076–2100.
- Bao J, Manatunga A, Binongo JNG and Taylor AT (2011). Key variables for interpreting 99m Tc-mercaptoacetyl triglycine diuretic scans: development and validation of a predictive model. *American Journal of Roentgenology*, 197, 325–333. [PubMed: 21785077]
- Bartko JJ (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3–11. [PubMed: 5942109]
- Choudhary PK (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference*, 138, 1102–1115.
- 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. [PubMed: 19673146]
- Craig CR, and Stitzel RE (2004). *Modern Pharmacology with Clinical Applications*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Efron B (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68, 589–599.
- 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, 887–892. [PubMed: 15476528]
- Fleiss J (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378–382.
- 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 HG (1979). Kernel estimates of regression functions In Gasser T, & Rosenblatt M (Eds.), *Smoothing Techniques for Curve Estimation*, Lecture Notes in Math (757). Berlin: Springer-Verlag pp. 23–68.
- Gasser T and Müller HG (1984). Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11, 171–185.
- Green P, Silverman B (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Taylor & Francis.
- Kraemer HC (1980). Extension of the kappa coefficient. *Biometrics*, 36, 207–216. [PubMed: 7190852]
- Li R and Chow M (2005). Evaluation of reproducibility for paired functional data. *Journal of Multivariate Analysis*, 93, 81–101. [PubMed: 19458785]

- Lin L (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255–268. [PubMed: 2720055]
- Müller HG (1984). Smooth optimum kernel estimators of densities, regression curves and modes. *The Annals of Statistics*, 12, 766–774.
- Müller HG (1985). Kernel estimators of zeros and of location and size of extrema of regression functions. *Scandinavian Journal of Statistics*, 12, 221–232.
- 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. [PubMed: 23399559]
- Rahman A, Peng L, Manatunga A and Guo Y (2017). Nonparametric regression method for broad sense agreement. *Journal of Nonparametric Statistics*, 29, 280–300. [PubMed: 28546735]
- Ramsay J, and Silverman B (2005). *Functional Data Analysis*. New York, NY: Springer.
- Rathnayake LN and Choudhary PK (2016). Tolerance bands for functional data. *Biometrics*, 72, 503–512. [PubMed: 26574904]
- Taylor A and Garcia EV (2014). Computer-assisted diagnosis in renal nuclear medicine: rationale, methodology, and interpretative criteria for diuretic renography. *Seminars in Nuclear Medicine*, 44, 146–158. [PubMed: 24484751]
- Taylor A, Manatunga A and Garcia EV (2008). Decision support systems in diuresis renography. *Seminars in nuclear medicine*, 38, 67–81. [PubMed: 18096465]

FIGURE 1.

Representative renogram curves for three kidneys. The solid lines are from a kidney rated as “nonobstructed” by expert consensus; the dashed lines are from a kidney rated as “equivocal”; and the dotted lines are from a kidney rated as “obstructed.” MAG3, ^{99m}Tc -mercaptoacetyltriglycine

FIGURE 2.

Representative curve sample (three 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 $Y = 3$. BSA, broad sense agreement

TABLE 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) (Cov95)

Scenario	True BSA values	n = 40					n = 60				
		N	EmpBias	EmpSD	EstSE	Cov95	EmpBias	EmpSD	EstSE	Cov95	
1	$\rho_{BSA}(\phi_{AUC}(X), Y) = 0.713$	(a)	-0.001	0.096	0.095	0.938	-0.001	0.076	0.077	0.944	
		(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	$\rho_{BSA}(\phi_{AUC}(X), Y) = 0.425$	(a)	-0.001	0.149	0.151	0.947	-0.006	0.119	0.122	0.961	
		(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	$\rho_{BSA}(\phi_{MAX(1/2)}(X), Y) = 0.682$	(a)	-0.012	0.103	0.106	0.942	-0.014	0.081	0.085	0.959	
		(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	$\rho_{BSA}(\phi_{MAX}(X), Y) = 0.662$	(a)	-0.011	0.108	0.111	0.936	-0.014	0.084	0.089	0.956	
		(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	$\rho_{BSA}(\phi_{TIME(1/2)}(X), Y) = 0.500$	(a)	0.008	0.065	0.060	0.963	0.011	0.050	0.048	0.918	
		(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	$\rho_{BSA}(\phi_{MAX}(X), Y) = 0.500$	(a)	0.020	0.044	0.039	0.978	0.019	0.036	0.033	0.916	
		(b)	0.018	0.024	0.021	0.966	0.019	0.020	0.018	0.973	

Scenario	True BSA values	n = 40					n = 60				
		N	EmpBias	EmpSD	EstSE	Cov95	EmpBias	EmpSD	EstSE	Cov95	
		(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	

Abbreviations: BSA: broad sense agreement.

N denotes the five study designs: (a) unbalanced design with N_j following a Poisson distribution with mean 20; (b) unbalanced design with N_j following a Poisson distribution with mean 40; (c) balanced design with $N_j = 20$; (d) balanced design with $N_j = 40$; and (e) balanced design with $N_j = 60$.

TABLE 2

Estimated BSA measures based on four types of 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 subscan 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)	<0.001
	$\phi_{AUC}^{[1]}$	0.32 (0.15, 0.47)	0.02 (−0.15, 0.18)	0.76 (0.60, 0.86)	<0.001
	ϕ_{MAX}	0.58 (0.38, 0.73)	...	...	...
	$\phi_{MIN}^{[1]}$	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)	<0.001
	$\phi_{AUC}^{[1]}$	0.15 (−0.03, 0.32)	−0.03 (−0.17, 0.12)	0.51 (0.30, 0.67)	<0.001
	ϕ_{MAX}	0.37 (0.16, 0.54)	...	...	...
	$\phi_{MIN}^{[1]}$	0.45 (0.27, 0.60)	...	...	...

Left	ϕ_{MAX}	$\phi_{MIN}^{[1]}$
	0.045	0.314
Right	ϕ_{MAX}	$\phi_{MIN}^{[1]}$
	0.178	0.544

* *P*-values from testing equality of BSA measures (vs. other type SFs)
Abbreviations: BSA: broad sense agreement; CI: confidence interval; SF: summary functional.

TABLE 3

Estimated BSA measures based on two types of SFs and results of hypothesis tests comparing their BSA values (*P*-value) for postfurosemide renogram data

Kidney	Estimated BSA (95% CI)		<i>P</i> -value
	Φ_{AUC}	Φ_{MAX}	
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

Abbreviations: BSA: broad sense agreement; CI: confidence interval; SF: summary functional.