Semi-supervised neighborhoods and localized patient outcome prediction

ALISON E. KOSEL*, PATRICK J. HEAGERTY

*Department of Biostatistics, University of Washington, F-600, Health Sciences Building
1705 NE Pacific Street, Seattle, WA 98195-7232, USA
aekosel@uw.edu*

SUMMARY

Robust statistical methods that can provide patients and their healthcare providers with individual predictions are needed to help guide informed medical decisions. Ideally an individual prediction would display the full range of possible outcomes (full predictive distribution), would be obtained with a user-specified level of precision, and would be minimally reliant on statistical model assumptions. We propose a novel method that satisfies each of these criteria via the semi-supervised creation of an axis-parallel covariate neighborhood constructed around a given point defining the patient of interest. We then provide non-parametric estimates of the outcome distribution for the subset of subjects in this neighborhood, which we refer to as a localized prediction. We implement local prediction methods using dynamic graphical methods which allow the user to vary key options such as the choice of the variables defining the neighborhood, and the size of the neighborhood.

Keywords: Local prediction; Non-parametric; Semi-supervised learning.

1. INTRODUCTION

1.1. *Scientific motivation*

Patients who understand their individualized prognosis are more likely to work collaboratively with their healthcare provider to make treatment choices that are consistent with their goals and specific values. For example, the Seattle Heart Failure Model (SHFM) predicts survival for heart failure patients using traits such as patient sex, age, laboratory measures, and medications (Levy and others, 2006). The SHFM was developed because physicians need to counsel patients about their prognosis in order to guide decisions about medications, devices, or end-of-life care. Similarly, the Framingham Heart Study has developed a prognostic score to assess 10-year risk of cardiovascular disease based on patient sex, age, total and high-density lipoprotein cholesterol, systolic blood pressure, treatment for hypertension, smoking, and diabetes status. The Framingham score has been recommended for use in guiding preventive care decisions (D’Agostino and others, 2008).

Prognostic scores are also used in other medical settings such as organ transplantation. For example, the lung allocation score assigns priority to lung transplant recipients in the United States based on patient

*To whom correspondence should be addressed.

characteristics such as age, clinical status, and specific diagnostic categories which predict both the risk of death without intervention and survival if the patient is transplanted (Egan and others, 2006). In liver disease, the Mayo model for survival among primary biliary cirrhosis patients is based on measurements that are simple to obtain including patient age, total serum bilirubin and serum albumin concentrations, prothrombin time, and severity of edema (Dickson and others, 1989). More recently, the Model for end-stage liver disease assesses chronic liver disease severity to determine priority for liver transplant recipients (Kamath and others, 2001). These select examples show that prognostic models are used across a number of disease settings to both inform patient expectations and to guide clinical decision making.

Our goal is to develop a statistical framework for providing adaptive individual patient predictions that are easily interpretable by both clinicians and patients, and that do not depend on any model assumptions. Non-parametric estimation typically requires large sample sizes. However, we expect the feasibility of such direct estimation approaches will increase with the adoption of electronic health records (EHRs) prompted by recent government initiatives such as the Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted in 2009, which allocates roughly \$30 billion to promote the adoption of health information technology and incentivizes its use (Jha, 2010). Ultimately data quality and generalizability need to be considered with any use of contemporary EHR data (Keiding and Louis, 2016).

1.2. Statistical background and innovation

Understanding how individuals perceive information in order to make decisions is essential for developing effective and appropriate statistical summaries. Psychological research has shown that both the perceived *personal relevance* and the *validity* of information are important aspects that determine the likelihood that information will lead to individual action (Petty and Cacioppo, 1986). Furthermore, select surveys have shown that the likelihood of physicians changing their medical practice on the basis of new clinical evidence depends on the physician’s impression of both the relevance and the reliability of the research results. For example, one aspect that has been shown to impact uptake is the sample size used for a clinical study (Ibrahim and others, 2009), with a larger sample size leading to a greater probability of adopting new interventions. Therefore, we seek a non-parametric prediction method that allows the user to directly control the number of observations within a local neighborhood of subjects. We also believe that it is critically important to disclose the exact region/neighborhood that is used for any individual prediction so that the personal relevance of the support within the data used to generate the prediction can be subjectively judged by the patient and/or provider.

We introduce a simple distance-based method to perform local non-parametric estimation based on a neighborhood that is selected to have a fixed sample size (precision), and which is based on independent restrictions on each covariate of interest. Such a rectangular neighborhood is termed “axis-parallel” and yields a simple, interpretable description for patients, and providers that communicates the data that was actually used to construct local distribution estimates. Our basic goal is to transfer the standard clinical question of making a prediction for an individual subject to the question of prediction for “subjects like the specific individual”, and we seek to return both the desired predictive distribution estimate, and a neighborhood specification that was used to provide the estimate. In summary, our choice is to *transfer* from a specific covariate point to a select subgroup of patients, and to *explicitly return* the subgroup used.

In order to detail our approach, consider an outcome, Y , and a set of covariates, $\mathbf{X} = (X_1, \dots, X_k)$. Typically predictive research goals focus on estimating the outcome distribution for either the entire population of patients (i.e. $F(y)$ marginally) or for a single patient characterized by exact covariates values (i.e. $F(y|\mathbf{X} = \mathbf{x}^*)$). Herein, we choose an intermediate target, where we let $N(\Delta, \mathbf{x}^*)$ be a neighborhood of distance Δ about the covariate vector $\mathbf{x}^*$, and we then define the parameter of interest, $F(y|\mathbf{X} \in N(\Delta, \mathbf{x}^*))$,

based on a localized neighborhood:

$$\{\underline{\mathbf{X}} = \underline{\mathbf{x}}^*\} = N(0, \underline{\mathbf{x}}^*) \subseteq N(\Delta, \underline{\mathbf{x}}^*) \subseteq N(\infty, \underline{\mathbf{x}}^*) = \{\underline{\mathbf{X}} = \underline{\mathbf{x}}, \forall \underline{\mathbf{x}}\}. \quad (1.1)$$

By specifying the distance Δ , we can either focus on a point ($\Delta = 0$), the entire population ($\Delta = +\infty$), or a select subgroup using $0 < \Delta < +\infty$.

Ultimately, we focus on the novel idea of using a fixed precision neighborhood, defined as a region within the covariate space that has a fixed number of points contained within it. By choosing a neighborhood $N(\Delta, \underline{\mathbf{x}}^*)$, such that it contains a desired number of points, m , we are making the decision to accept variable distances to define a neighborhood depending on the density of points around an index point $\underline{\mathbf{x}}^*$ of interest. Finally, by returning both an outcome prediction and the neighborhood used, we clearly inform the user of the data's ability to answer the question of interest: if there are very few data points near a given patient, then the patient and/or clinician will be made aware of this by the associated size of the neighborhood defined by the region in the covariate space used to support and generate the prediction. Although a neighborhood $N(\Delta, \underline{\mathbf{x}}^*)$ could in principle be any shape, we chose to restrict our consideration to axis-parallel neighborhoods, or simple covariate rectangles. We do so for ease of interpretation since axis-parallel boxes can be described by a defined range used for restriction on each co-ordinate of $\underline{\mathbf{X}}$, and are therefore, easily presented and understood by medical professionals and patients.

In order to provide individualized predictions, we seek to estimate an outcome distribution for subgroups of patients who are characterized by their specific baseline clinical or demographic variables. We are interested in the full conditional distribution rather than a summary measure such as the mean, since the distribution can then provide multiple meaningful summaries such as the conditional median value, or the 25th and 75th percentiles. Our premise is that providing details such as the percentage of subjects with very good or very bad outcomes is important for decision makers. In addition, patients can easily understand statements such as: "25% of subjects like you will have an outcome less than or equal to 3 on a 10-point pain scale," and therefore, we focus on determining all percentiles, or equivalently the full conditional distribution. We believe that simple estimates of local means are inadequate to fully inform patient expectations.

Conditional predictions are a focus of many contemporary statistical learning methods that allow great flexibility in the breadth of candidate input predictive information (Hastie and others, 2009). However, most regression and/or learning methods focus on generating a predictive conditional mean or risk prediction, while we seek a valid estimate of the full predictive distributions. Another common limitation to the use of standard predictive methods is that transparency in terms of how well the data support prediction for an individual is not generally an intentional product of the methods. Many clinical risk prediction calculators have been created in recent years, but use of these tools rarely, if ever, gives information back to the user about the underlying covariate data distribution or relevance for the target individual. Our approach is to explicitly produce information on the "support" within the data toward generating a conditional prediction, and we explicitly return information on the characteristics of the patient neighbors who are used to inform individual prediction. Finally, quantile regression methods described by Koenker and Bassett (1978) may be used to obtain any pre-selected percentile, but do not generally provide a full distribution estimate without imposing monotonicity constraints across separate quantile regressions.

Our proposed methods are a variation on non-parametric local kernel methods studied by Stütgen (1986b). However, our proposal involves choice of a specific data adaptive distance function to construct meaningful and interpretable covariate neighborhoods that form the basis for estimation. We define a distance function that uses the strength of the covariate relationships with the outcome, and therefore, we are proposing a semi-supervised local predictive method. In addition, standard bandwidth selection methods for local estimation typically consider predictive criteria that balance bias and variance, while we see

Table 1. *A summary of the notation used to characterize the proposed localized prediction*

Description	Notation
Dimension of points	k
Number of points	n
Number of points in neighborhood	$m \leq n$
Points	$\mathbf{X}_1, \dots, \mathbf{X}_n$
Outcomes	$Y_1, \dots, Y_n$
Point of interest	$\underline{\mathbf{x}}^*$
Neighborhood of distance Δ about $\underline{\mathbf{x}}^*$	$N(\Delta, \underline{\mathbf{x}}^*)$
Estimated neighborhood of distance $\hat{\Delta}$ about $\underline{\mathbf{x}}^*$	$N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)$
Vertices defining $N(\Delta, \underline{\mathbf{x}}^*)$	$V(\Delta, \underline{\mathbf{x}}^*) \in \mathbb{R}^k \times \mathbb{R}^k$
Distribution function for outcome in $N(\Delta, \underline{\mathbf{x}}^*)$	$F[y \mathbf{X} \in N(\Delta, \underline{\mathbf{x}}^*)]$

value in direct control of precision (or variance) at a pre-specified level in order to facilitate interpretation, and then we explicitly communicate the transfer of estimation from an index point to a specific patient neighborhood.

As a non-parametric method, we are practically restricted to a low-dimensional set of predictors. However, our estimation can be applied after meaningful covariate dimension reduction. For example, unsupervised methods such as principal components may be adopted to derive a summary score for a related group of clinical measures. Alternatively, supervised methods can generate predictive scores which can then be used to form one element of neighborhood construction. In our illustration in Section 4, we demonstrate use of one such dimension reduction approach.

2. METHODS

In this section, we detail our choice of distance metrics, our approach to obtaining an adaptive neighborhood, and the statistical properties of our local distribution estimator. A summary of key notation can be found below in Table 1.

2.1. Neighborhood definition

Our goal is to create an interpretable localized estimate of the distribution of patient outcomes. We create an estimate based on subjects similar to a given patient, and therefore, rely on covariate distance metrics to define neighbors. We allow the importance of covariates to be determined by adaptive neighborhood scaling, and the final neighborhood can then be used to estimate the full local distribution.

We consider a class of distance options based on the family of metrics defined by key parameters:

$$d(q, A, G, \underline{\mathbf{x}}^*, \underline{\mathbf{x}}) = \|A[G(\underline{\mathbf{x}}) - G(\underline{\mathbf{x}}^*)]\|_q. \quad (2.1)$$

One can in theory choose any metric in this class to create a neighborhood $N(\Delta, \underline{\mathbf{x}}^*)$. Examples of commonly used metrics in this class are Mahalanobis distance, with $A = V^{-\frac{1}{2}}$, $G = I$, and $q = 2$; L1 distance, with $A = I$, $G = I$, and $q = 1$; and nearest neighbors, with G as the function that calculates the element-wise marginal distribution function for each covariate. In the subsections below, we comment on each distance functions element with the goal of identifying a flexible yet interpretable neighborhood specification.

2.2. Choice of q

The choice of the distance parameter q ultimately determines the shape of any neighborhood defined by a restriction to those points within a fixed distance. For example, $q = 2$ corresponds to the L_2 norm and results in a generalized circle. Using $q = 1$ corresponds to L_1 and is well known to return a “diamond” shaped region. Each of these choices defines a region that is not simple to describe to patients or providers in terms of the restriction on values used for each covariate. However, to report back to the user the covariate neighborhood, we could in principle choose any metric and then take the rectangular enclosure of the points defined by that metric. Unfortunately, controlling the size of the local enclosure, or m , in these cases would be difficult, since it is not directly determined by the defining distance, and would potentially vary across the covariate space. Therefore, we prefer to use $q = \infty$, or a scaled version of the supremum norm, L_∞ , dependent on a choice of A , since use of this metric and a distance restriction on points around an index point will directly yield an axis-parallel neighborhood. Such neighborhoods are defined by independent interval restrictions, $[x_j(L), x_j(U)]$ using lower ($x_j(L)$) and upper ($x_j(U)$) limits for each covariate. In this case, the description of a neighborhood is given by a simple table where each covariate is listed and the associated limit values presented. When using the scaled sup-norm metric to define a neighborhood, $N(\Delta, \underline{\mathbf{x}}^*)$, we can control the number of points m exactly by appropriate selection of a distance restriction, and the defining enclosure $V(\Delta, \underline{\mathbf{x}}^*)$ is simple to communicate.

Although the supremum norm creates a square by definition, we choose to shrink this square to a rectangle by taking the furthest actualized pair of distances along each axis instead of considering the neighborhood to extend equal distances in all directions from $\underline{\mathbf{x}}^*$. The difference is likely to be small, but from an interpretation perspective, it is more sensible to consider the neighborhood to be only as big as the points it contains. Note that the resulting rectangle may not be symmetric about $\underline{\mathbf{x}}^*$; it is unlikely to be highly asymmetric, but the square will be shrunk different distances on each side. Also note that, while the square created by the supremum norm is the smallest square centered at $\underline{\mathbf{x}}^*$ containing m points, the resulting rectangle may not be the smallest rectangle containing $\underline{\mathbf{x}}^*$ and m other points. If, for example, the data is particularly sparse on one side of $\underline{\mathbf{x}}^*$, a smaller and highly unbalanced rectangle could be created.

2.3. Choice of A

Although, we have settled on a preferred choice of q for our distance metric, we must still consider the choice of A and G . Note that the supremum norm treats a one unit change in the direction of any co-ordinate as equivalent. However, this is undesirable in many cases due to both different scales of measurement for each covariate, and differential variable importance toward predicting the outcome. For example, one might not wish to equate a 1 mm of mercury change in systolic blood pressure with a one unit change in body mass index. More generally, it would not be desirable to have neighborhoods depend on choice of measurement units for any variable. Therefore, adoption of any distance function that combines information across axes will require consideration of appropriate scaling represented by the matrix A or the function G .

There are two main categories of methods to rescale candidate covariates: outcome-independent methods (unsupervised) and outcome-dependent methods (supervised). In other words, do we choose to rescale based on the covariates alone or do we take their relationship to the outcome into consideration? In the next subsections, we consider using global or local regressions to determine element-wise covariate rescaling specified by the matrix A , and then we consider potential use of transformation functions, G .

2.3.1. Global linear supervised marginal scaling. To correct the potential problem of differing scales among covariates, we propose outcome-based marginal rescaling. To implement this, one simply regresses Y on X_j for each covariate $1 \leq j \leq k$ individually. This produces a $\hat{\beta}[j]$ for each predictor $1 \leq j \leq k$.

To rescale, all x_j and x_j^* are scaled by their regression estimate, and this equates to adopting $A = \text{diag}(\hat{\beta}[j])$. In other words, one rescales each coordinate based on its marginal association with the outcome where all covariate scales are equivalent since distances $d = \beta[j]X_j$ correspond to a common (d -unit) change in the expected outcome. Therefore, coordinates that are strongly associated with the outcome have their distances weighted more heavily, while coordinates that are weakly associated with the outcome have their distances weighted less heavily. Marginal regression coefficient rescaling puts all covariates on an outcome-standardized scale.

While there would potentially be benefits to a regression using all covariates in a multivariate model, there may also be drawbacks. Using marginal scaling allows the user to select different subsets of covariates to form neighborhoods without changing the choice of scaling for each variable that is included. Furthermore, adoption of a multivariate regression for scaling invites issues of interaction to be explicitly considered.

2.3.2. Local linear supervised marginal scaling. Marginal coefficient scaling can be easily relaxed to accommodate possible non-linear associations between a predictor and the outcome. Flexible parametric or smooth non-parametric methods can be used to estimate a general response surface: $E[Y | X_j] = m(X_j, \theta[j])$. In this scenario rescaling at a point of interest $\underline{\mathbf{X}} = \underline{\mathbf{x}}^*$ could then use the derivative at the target point for rescaling: $\widehat{\beta}(\underline{\mathbf{x}}^*)[j] = m'(\underline{\mathbf{x}}^*, \hat{\theta}[j])$. Such a local linear rescaling requires choice of methods to determine the functional form either through spline basis choice (parametric) or bandwidth specification (non-parametric). Inference with parametric spline methods is unchanged from that using linear rescaling, while more general non-parametric rescaling may result in altered asymptotic properties for the resulting localized outcome distribution estimate depending on assumptions regarding the choice of the neighborhood size as a function of sample size. We consider these issues in Section 3.

2.4. Choice of G

In this section, we consider outcome-independent methods that can be used to rescale the covariates into comparable units. One common choice would be to use either element-wise standardization or Mahalanobis distance to convert covariates to a standardized scale. A second option would be to transform covariates into percentiles (i.e. a choice of G). The advantage of percentiles is that a one unit change in any direction results in inclusion of the same marginal fraction of the data, while the advantage of standard deviation units is that the actual distances within each coordinate are maintained. Finally, note that it is possible to combine outcome-based marginal rescaling with standardization, i.e. apply the outcome-based marginal rescaling to data already converted to the scale of percentiles or to standard deviations.

2.5. Algorithm: L_∞ and supervised scaling

In this subsection, we detail one attractive algorithm which involves a simple procedure to find the axis-parallel neighborhood of m points about a target point $\underline{\mathbf{x}}^*$. First, the user inputs the point of interest and the number of points desired in the neighborhood. Then the distance between the point of interest and all other points is calculated, and finally a box containing the desired number of closest points is created. Note that, as described in previous sections, the choices of A and G are left open, but below we detail use of supervised global scaling without use of any additional covariate transformation.

1. Select $\underline{\mathbf{x}}^*$ and m ; let $p = \frac{m}{n}$.
2. Let A be either the identity matrix of size k (no scaling) or, for $1 \leq j \leq k$, regress Y on $\mathbf{X}[j]$, and let the slope / coefficient be $\hat{\beta}[j]$, and A be the diagonal matrix created by $\text{vec}(\hat{\beta}[j])$ (outcome-based

rescaling). Let $G(\mathbf{x})$ be any desired function, such as identity or percentile. For $1 \leq i \leq n$, calculate the resulting distances, $d_i = \|A[G(\mathbf{x}_i) - G(\mathbf{x}^*)]\|_\infty$.

3. Calculate the quantile of the empirical distribution of the distances, d_i : $Q_d^p = \hat{F}_{d_i}^{-1}(p)$. Discard all $\mathbf{x}_i$ such that $d_i > Q_d^p$; call the remaining points $N(\hat{\Delta}_n, \mathbf{x}^*)$ where $\hat{\Delta}_n = Q_d^p$.
4. Find the enclosure of $N(\hat{\Delta}_n, \mathbf{x}^*)$. For $1 \leq j \leq k$, let $V[j, 1] = \min(\{\mathbf{x}[j] | \mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)\})$ and $V[j, 2] = \max(\{\mathbf{x}[j] | \mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)\})$.
5. Finally, given the identified axis-parallel neighborhood we compute the local empirical distribution function as our final localized outcome prediction:

$$\hat{F}_n[y | \mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)] = \frac{1}{m} \sum_{i=1}^n 1(Y_i \leq y) \cdot 1[\mathbf{X}_i \in N(\hat{\Delta}_n, \mathbf{x}^*)].$$

2.6. Functionals

In addition to estimating the full outcome distribution, $F[y | \mathbf{x} \in N(\Delta, \mathbf{x}^*)]$, we can also estimate any functionals of this distribution such as the mean. Note that our method allows nearly any choice of functional to be estimated because we non-parametrically obtain an estimate of the entire conditional distribution, and therefore simultaneously obtain estimates for any quantiles.

2.7. Ties

With discrete covariate data there is the additional potential issue of ties when constructing fixed size neighborhoods since multiple data points may have the exact same value for one or more covariates at the boundaries of the neighborhood. Therefore, in some cases obtaining precisely the desired level of precision may not be possible unless additional restrictions are used. In particular, since the L_∞ metric results in separate restrictions on each covariate axis any discreteness can produce substantially more points than desired (or fewer depending on use of $<$ or $\leq$ to restrict points). In the case of ties, we recommend applying a second metric, such as L_1 , in order to provide additional restriction and yield a neighborhood closer to the target size of m . In particular, either an additional L_1 or L_2 constraint will result in trimming points that are in the furthest corners of the L_∞ rectangle and will break ties using a secondary distance. However, in the extreme case of highly discrete coordinates, it may be difficult to achieve a neighborhood of m points without arbitrary methods to break ties.

2.8. Computational complexity

2.8.1. Single $\mathbf{x}^*$. In practice, we would typically perform calculations dynamically; that is, we would calculate our neighborhood estimates based upon the user's specific choice. We perform k linear regressions, computed as $(X^T X)^{-1} X^T Y$, with an $n \times 2$ design matrix. Each regression requires calculating $X^T X$, at a cost of $O(n)$; a matrix inversion, at a cost of $o(n)$; calculating $X^T Y$, at a cost of $O(n)$; and, finally, multiplying $(X^T X)^{-1}$ by $X^T Y$, at a cost of $o(n)$. The resulting cost for each regression is $O(n)$, which gives us an asymptotic cost of $O(nk)$ for k regressions. Calculating the distance between $\mathbf{x}^*$ and the n other points has time complexity $O(n)$. We then must sort the distances. For example, "quicksort" is commonly used which averages $O(n \log n)$ and has a worst case scenario of $O(n^2)$ (Skiena, 2009). Thus, the sorting dominates the cost; overall we average $O(n \log n)$ asymptotically, with a worst case scenario of $O(n^2)$.

2.8.2. All $\mathbf{x}$. If individual predictions will be based an external, static database, then we might wish to have all possible calculations already performed and simply awaiting retrieval. Our method asymptotically

results in a cost of $O(n^2 \log n)$ on average and $O(n^3)$ in the worst case when all points are used in turn as $\underline{\mathbf{x}}^*$. Segal and Kedem describe an algorithm to obtain the m nearest rectilinear neighbors of a given point that has lower total cost; however, they require that $m \geq \frac{n}{2}$ (Segal and Kedem, 1998).

3. INFERENCE

In this section, we consider the asymptotic distribution for the local empirical distribution estimator that uses global scaling to determine the neighborhood. When studying inference, we continue with our transfer from a specific point to a localized neighborhood, and we consider a procedure where a neighborhood of $p \times 100$ percent of the data about $\underline{\mathbf{x}}^*$ is used, or $p = m_n/n$. We refer to this as the “fixed percentage” scenario. In this situation the target parameter is defined as:

$$\bar{F}(y|\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)) = E_{\underline{\mathbf{X}}} \{F[y | \underline{\mathbf{X}} = \underline{\mathbf{x}}] \cdot 1[\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]\} / P[\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)].$$

Ultimately we consider four key sources of variability: variation in Y_i for observations in a neighborhood; variation in the points $\underline{\mathbf{X}}_i$ that are in the neighborhood; variation associated with the distance restriction, $\hat{\Delta}_n$; and variation associated with the global scaling parameters, $\hat{\beta}[j]$. Letting $F_{n,i}(y) = P[Y_{n,i} \leq y | \underline{\mathbf{X}}_{n,i} = \underline{\mathbf{x}}_{n,i}]$ be the true conditional distribution function for the i th Y in a sample of n , we define $\tilde{F}_n(y|\underline{\mathbf{X}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)) = \frac{1}{m} \sum_{i=1}^n F_{n,i}(y) \cdot 1_{N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)}(\underline{\mathbf{x}}_{n,i})$. We focus on this local average of distribution functions since we do not necessarily have i.i.d. outcomes within a neighborhood $N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)$.

Suppose we consider

$$\sqrt{m} \left(\hat{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)] \right) \quad (3.1)$$

$$= \sqrt{m} \left(\hat{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \tilde{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] \right) \quad (3.2)$$

$$+ \sqrt{m} \left(\tilde{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] \right) \quad (3.3)$$

$$+ \sqrt{m} \left(\bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)] \right) \quad (3.4)$$

where, we decompose the scaled estimation error into three conditionally independent components: that due to uncertainty in outcome (Expression 3.2), that due to the specific points in the neighborhood (Expression 3), and that due to uncertainty in neighborhood location (Expression 3.4). We label the uncertainty in outcome, $\hat{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \tilde{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)]$, as O and the uncertainty due to specific points and distance, $\tilde{F}_n[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)]$, and $\bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]$, as L and then the orthogonality of these two components is easily shown using standard conditioning arguments:

$$E[OL] = E[E[OL|\underline{\mathbf{X}}]] = E[E[O|\underline{\mathbf{X}}]L] = E[E[O|\underline{\mathbf{X}}]]E[L] = E[O]E[L]$$

where the second equality is due to the fact that L is constant, when $\underline{\mathbf{X}}$ is fixed and the third due to the clear independence of $E[O|\underline{\mathbf{X}}]$ and $E[L]$. We, therefore, may begin by analyzing the uncertainty in outcome (Expression 3.2), which, using results from Shorack and Wellner (2009), converges to $\mathbb{X}$ as $m \rightarrow \infty$, where $\mathbb{X}$ is a normal process with mean zero and a covariance function K that is the limit of $K_m(s, t) = s \wedge t - \frac{1}{m} \sum_{i=1}^m G_{mi}(s)G_{mi}(t)$, where $G_{mi} = F_{mi} \circ \tilde{F}_m^{-1}$. Note that, where $\mathbb{U}$ is a Brownian bridge process, $P(\|\mathbb{X}\| \leq \lambda) \geq P(\|\mathbb{U}\| \leq \lambda)$ for all $\lambda > 0$, with equality when $\underline{\mathbf{X}}$ is iid (i.e. when G_{mi} is the identity function); in essence, $\mathbb{X}$ is a Gaussian process that is less variable than a Brownian bridge.

Expression 3 involves uncertainty associated with specific points, $\mathbf{X}_i$ in the neighborhood defined by $\hat{\Delta}_n$ and therefore can be easily treated with empirical process results for weighted averages. Specifically, the limiting distribution is:

$$\sqrt{m} \left(\tilde{F}_n[y|\mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)] - \bar{F}[y|\mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)] \right) \xrightarrow{d} \sqrt{v_1(y, \Delta)} \cdot Z_1$$

where $Z_1 \sim N(0, 1)$ and the scaling yields the limit of the mean variance of $F_{n,i}(y)$ conditional on $\hat{\Delta}_n$, which we label $v_1(y, \Delta)$.

Finally, we may focus on the uncertainty in neighborhood (Expression 3.4), for which we turn to the Central Limit Theorem (see Appendix A.2 for further details) to obtain, where f_d is the distribution of distances about $\mathbf{x}^*$,

$$\sqrt{m} \left(\bar{F}[y|\mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)] - \bar{F}[y|\mathbf{x} \in N(\Delta, \mathbf{x}^*)] \right) \xrightarrow{d} \quad (3.5)$$

$$\sqrt{v_2(y, p, \Delta)} \cdot Z_2 \quad (3.6)$$

where $Z_2 \sim N(0, 1)$ and $v_2(y, p, \Delta) = \frac{p^2(1-p)}{f_d^2(\Delta)} \times \left\{ \frac{d}{d\Delta} \bar{F}[y|x \in N(x^*, \Delta)] \right\}$.

Therefore,

$$\sqrt{m} \left(\hat{F}_n[y|\mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)] - \bar{F}[y|\mathbf{x} \in N(\Delta, \mathbf{x}^*)] \right) \xrightarrow{d} \quad (3.7)$$

$$\mathbb{X} + \sqrt{v_1(y, \Delta)} \cdot Z_1 + \sqrt{v_2(y, p, \Delta)} \cdot Z_2. \quad (3.8)$$

In Appendix A.4 we also provide inference for those who wish to compare our estimate to the point-specific target parameter at the point $\mathbf{x}^*$ and obtain an unbiased estimator thereof. In this case, our method can still be used and viewed as a data-adaptive smoothing method. We term this case the smoothing scenario, where $m = O(n^\alpha)$, $0 < \alpha < 1$. It is important to note that p_n varies in this scenario, as m grows at a slower rate than n ; because $p_\infty = 0$, $N(\Delta, \mathbf{x}^*) = \mathbf{x}^*$, and our target is the point rather than the neighborhood.

In the above inference, we assume no scaling, or the use of a known scaling parameter. The use of an estimated scaling parameter introduces additional variation to the estimator, and we, therefore, also consider

$$\sqrt{m} \left(\hat{F}_n[y|\mathbf{x} \in N(\hat{\Delta}_n(\hat{\beta}_n), \mathbf{x}^*)] - \bar{F}[y|\mathbf{x} \in N(\Delta(\beta), \mathbf{x}^*)] \right). \quad (3.9)$$

This additional component can be evaluated using results from [van der Vaart and Wellner \(2007\)](#). Provided that the behavior of $\hat{\beta}_n$ satisfies standard assumptions for regression estimators, we obtain

$$\sqrt{m}\sqrt{p} \left[\bar{F}(y \mid \|\hat{\beta}_n(\mathbf{x} - \mathbf{x}^*)\|_\infty \leq \hat{\Delta}_n) - \bar{F}(y \mid \|\beta(\mathbf{x} - \mathbf{x}^*)\|_\infty \leq \hat{\Delta}_n) \right] \xrightarrow{d} \sqrt{v_3(y, \Delta)} \cdot Z_3 \quad (3.10)$$

where $Z_3 \sim N(0, 1)$ and $v_3(y, \Delta)$ can be obtained via the delta method (see Appendix).

Note that this regression term is not orthogonal to the uncertainty in outcome (Expression 3.2). We focus here only on the order of this term, as its exact value will vary depending on the method of the parameter's estimation. We have four possible scenarios: local and global scaling parameter estimation in both the smoothing and the fixed percentage situations. In the smoothing situation under global scaling estimation, the additional uncertainty from estimating β is asymptotically negligible; $p \rightarrow 0$ and the rest

of the uncertainty term converges at a rate of $\sqrt{m}$. In the smoothing scenario under local scaling estimation and the fixed percentage scenario under global scaling estimation, the additional uncertainty from scaling estimation is of the same order as the Gaussian process. Local scaling would generally not be used in the fixed percentage scenario since the variation due to rescaling would dominate asymptotically.

3.1. *Evaluation of asymptotic approximations*

We have conducted extensive simulations to evaluate the coverage properties of confidence bands and intervals based on the asymptotic approximations to our localized patient outcome distribution estimators. Specifically, we have considered $\mathbf{X} \in \mathbb{R}^1$ and $\mathbf{X} \in \mathbb{R}^2$, with both unscaled and scaled estimators. For distributional shapes we considered a simple normal model, and more complex mixture distributions. Coverage of the localized cumulative distribution function parameter, $\bar{F}[y|\mathbf{x} \in N(\Delta(\beta), \mathbf{x}^*)]$, was assessed for both simultaneous confidence bands across all y , and pointwise for select values of y . In addition, we considered prediction for $\mathbf{x}^*$ at the center and at the edge of the distribution of $\mathbf{X}$. In summary, our estimation of Δ proved to be unbiased and our coverage was almost always within 1-2% of nominal levels, regardless of the outcome distribution as would be expected for non-parametric estimators. Accounting for estimation of scaling parameters also yielded nominal coverage levels. Detailed simulation settings and associated results may be found in [supplementary material](#) available at *Biostatistics* online.

4. IMPLEMENTATION AND ILLUSTRATION

4.1. *Implementation using dynamic graphics*

We have used the Shiny package in R ([R Core Team, 2012](#), [Chang and others, 2017](#)) to implement our method and to allow the user to dynamically select multiple key parameters. Specifically, we query the user regarding the covariate characteristics of the index subject for whom a prediction is desired, and we require specification of the desired sample size. We then return the covariate restrictions that define the neighborhood and produce an estimated cumulative distribution function or a simple histogram to estimate the localized outcome distribution. We also provide a graph showing the neighborhood that was used and the relative density of observations for the entire data set in order to inform the user of the characteristics of the neighborhood that was used to support the local prediction (see [Figure 1](#) for an example discussed below). In addition, we automatically provide summary statistics for the outcome such as the quartiles and mean of the empirical distribution, displaying both the full sample summary (overall) and the summary for the neighborhood of interest. By providing dynamic computational tools we allow the user to easily evaluate the impact of changing the target sample size, or of changing the scaling methods used to construct the neighborhood. In particular, we allow choice of global linear scaling of the covariate axes, or local linear scaling to allow the relative importance of covariates to potentially change depending on the index point selected for the local prediction. In summary, our dynamic graphical tool allows simple web-based exploration of predictive distributions in a user-selected dataset and focuses on returning both the characteristics of the neighborhood that was used and the associated non-parametric full predictive distribution for the outcome of interest.

Our current implementation in R has a number of key features that facilitate adoption by others, and there are some select limitations that we hope to address in future releases. First, the localized patient outcome prediction (LPOP) tool easily allows a user-loaded dataset, and automatically populates each of the variable choice drop-down menus with the available variables. This feature allows the user to easily change the predictors of interest and to explore predictive distributions as a function of alternative variable sets. Furthermore, the choice of neighborhood size is a simple slider and different sizes can be explored with the neighborhood and predictive distribution plots updated immediately. Finally, the tool allows the user to choose details of neighborhood construction such as using no covariate scaling or either global or

Localized Patient Outcome Predictions

Fig. 1. In this example of our dynamic graphical tool, we display the outcome for a patient aged 75 who has a baseline disability of 10.

local scaling options. Although our current dynamic graphical implementation presents the neighborhood for only two covariates, this tool could easily be extended to additional covariates. Rather than showing the bivariate or multivariate neighborhood, we would present the marginal density of each covariate, and we would show the index subject and the associated neighborhood boundaries for each covariate. Because our neighborhood is defined as axis-parallel restrictions, these univariate displays would provide a sufficient description of the neighborhood. The code to implement our method is available at the Github account <https://github.com/aekosel/lpop>.

4.2. Illustration: longitudinal cohort study and prognosis

Low back pain is the leading cause of years lost to disability according to the 2013 Global Burden of Disease (Vos and others, 2013). Contemporary treatment strategies now use select clinical tools such as the STarT Back Screening Tool to classify patients with low back pain into three risk groups based on key prognostic indicators (Hill and others, 2008). Using data from a recent large-scale longitudinal cohort study, we aim to provide physicians and patients with a personalized estimate of the patient's expected disability outcome over time. Specifically, we use the Back pain Outcomes using Longitudinal Data (BOLD) (Jarvik and others, 2012) which consists of approximately 5000 patients age 65 or older who present with a complaint of back pain. Standard of care is given to patients in this observational cohort, and longitudinal measures of pain and function are recorded for a 2 year follow-up period. In this

article, we focus on using baseline data to predict patient disability at 1 year after their index primary care visit.

For our example, we first select a patient who is 75 years old (age variable) and has a baseline disability (roland_0 variable) of 10, on a scale from 0 to 24 points, with a higher score indicating increasing severity. We consider the patient's disability at 1 year (roland_3) as our outcome. We specify a neighborhood size of $m = 446$ which is 10% of the total sample size. Figure 1 shows results for this subject and shows that the data-adaptive covariate neighborhood construction suggests that baseline disability is much more strongly related than age to disability at 1 year—the selected neighborhood is much narrower on the disability axis than the age axis when dynamically scaled. In comparison, if the covariates were not scaled, we would obtain a neighborhood with an age range of 71–79 and a baseline disability range of 6–14, in place of our scaled neighborhood ranging from 67 to 83 in age and 9 to 11 in baseline disability. Furthermore, we can compare data-adaptive (supervised) scaling to more traditional unsupervised covariate scaling strategies such as use of Mahalanobis distance. In this example, we find that the correlation between age and baseline disability is only $r = 0.067$ suggesting that Mahalanobis-based covariate ellipses would have minimal rotation. In addition, the standard deviation in age is 6.74 and the standard deviation for baseline disability is 6.36. Therefore, with Mahalanobis scaling one age unit is nearly equal to one baseline disability unit. However, when we use the 1-year disability outcome to construct semi-supervised neighborhoods we find that the linear regression coefficient of age is 0.12 while the linear regression coefficient for baseline disability is 0.62. Therefore, a one unit change in baseline disability is associated with a nearly five-fold greater impact on the mean outcome than a one-unit change in age. Consequently, our semi-supervised neighborhood is much tighter in the tolerance for baseline disability as compared with age, and such a scaling trade-off is not captured via typical unsupervised scaling approaches such as Mahalanobis distance.

In the globally scaled neighborhood, which is what we would use in practice, we see clear value in providing the complete predictive distribution. Importantly, we observe a large number of patients who recover completely (i.e. return to a disability score of 0) after 1 year, and note that the remainder of the patients form a roughly normal distribution around the chosen patient's original disability score. In cases such as this, with a bimodal outcome distribution, a single summary measure such as the mean or median cannot adequately inform a clinician or patient's expectations. It is important to be able to clearly visualize the space of possible outcomes.

Our non-parametric methodology directly provides distributional estimates, $\hat{F}_n[y \mid \mathbf{x} \in N(\hat{\Delta}_n, \mathbf{x}^*)]$, but can also provide measures of uncertainty for standard summaries of this predictive distribution. For example, with our index subject characterized by $\mathbf{x}^* = (\text{age} = 75, \text{baseline disability} = 10)$ the estimated predictive 25th percentile has a 95% confidence interval of (3.09, 5.67) while the 50th and 75th percentiles have confidence intervals of 8.38–10.14 and 12.37–13.93. Herein, we use our asymptotic results establishing normality and then adopt bootstrap methods for computational convenience. Furthermore, if other predictive summaries are of interest such as the predictive mean, or prediction error then these can also be directly derived from the localized predictive distribution estimate.

We can also examine slightly different patients, as seen in Figure 2. When we consider a patient with a higher baseline disability score—15 instead of 10—we see that when the neighborhood remains approximately the same size (volume) and that the outcome maintains its bimodal shape, but the second peak is shifted to the right indicating greater disability. We can also examine a patient on the edge of the data; for example, a patient who still has a baseline disability score of 10, but who is age 95 instead of 75. When keeping the sample size at 200 similar to earlier estimates, we now obtain a much larger volume neighborhood, though our outcome distribution still has the same bimodal shape. Instead of an age range of 16 years, we obtain a range of 19 years; in place of a disability range of 2 points, we obtain a range of 6 points. Note also the marked asymmetry of the neighborhood. There simply do not exist patients with a similar disability score who are older than the chosen patient, and so the box cannot extend in that direction. We feel that these neighborhood characteristics are not deficiencies of our approach but rather

Fig. 2. In this example of our method, we display the change in neighborhood and the change in outcome distribution when a new patient is selected. We begin with a patient who has a baseline disability of 10 and an age of 75, then consider a patient of the same age with a baseline disability of 15, and then consider again a patient with a baseline disability of 10 but this time an age of 95. In all three cases we see a bimodal outcome distribution, though the neighborhood for the patient on the edge of the data is substantially larger than the other two neighborhoods.

Localized Patient Outcome Predictions

Fig. 3. In this example of our dynamic graphical tool, we display the estimated 1-year disability outcome distribution using two predictors: baseline disability; and a linear combination of other prognostic variables including age, baseline back pain numerical rating scale, and leg pain numerical rating. Herein, we address the potential for use of multiple predictors through supervised dimension reduction where an outcome mean regression is used to derive a predicted score based on a subset of covariates. The patient-centered interpretation is that we present estimates “for 200 patients similar to you using your baseline disability score of 15, and your predicted mean outcome of 12 based on age and baseline back and leg pain ratings.”

they represent features of our method since the neighborhood information provides important disclosure regarding the inherent limitations of the data. While nearly all local statistical methods have difficulty on the boundary of the data, with our approach the user is clearly informed when the data about a given patient is sparse and when he might wish to be cautious in his interpretation.

Finally, with BOLD we consider application of localized estimation with multiple predictive covariates. One strategy is to derive a meaningful summary score and then to use that variable as a component of neighborhood construction. To illustrate the approach we consider use of baseline disability as the primary predictor, and then combine age, with both baseline back pain numerical rating and leg pain numerical rating into a derived variable. Using age and the pain scores we can use regression to derive a linear combination that predicts 1-year disability. In Figure 3, we show results for a subject with a baseline disability equal to 15 and an age and pain derived predicted mean outcome of 12. Careful choice of derived dimension reduction can retain patient-centered meaning and here we would present the estimate as representing the 1-year outcomes for “200 patients similar to you using your baseline disability score of 15, and your predicted mean outcome of 12 based on age and baseline back and leg pain ratings.”

Table 2. Analysis results for $n = 4459$ subjects from the BOLD cohort study

Methods	Subject = (75, 10)			Subject = (75, 15)			Subject = (95, 10)		
	Estimate (SE)	95% CI		Estimate (SE)	95% CI		Estimate (SE)	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
25th Percentile									
LPOP ($m = 200$)	5 / 5.32 (1.05)	3.25	7.40	7 / 7.06 (0.93)	5.24	8.87	5 / 5.52 (0.78)	3.99	7.06
LPOP ($m = 400$)	5 / 4.55 (0.72)	3.13	5.97	6 / 6.44 (0.81)	4.84	8.04	5 / 5.07 (0.49)	4.12	6.02
RQ additive	4.40 (0.50)	3.42	5.38	7.65 (0.64)	6.39	8.91	4.40 (0.50)	3.42	5.38
RQ interaction	5.52 (0.56)	4.28	6.50	8.24 (0.76)	6.79	9.75	3.33 (2.79)	−1.69	9.28
50th Percentile									
LPOP ($m = 200$)	10 / 9.53 (0.67)	8.21	10.85	12 / 12.19 (0.66)	10.89	13.49	10 / 9.94 (0.68)	8.61	11.28
LPOP ($m = 400$)	9 / 9.40 (0.50)	8.42	10.37	12 / 12.31 (0.52)	11.28	13.34	10 / 9.77 (0.47)	8.84	10.70
RQ additive	9.17 (0.28)	8.61	9.72	12.50 (0.34)	11.84	13.16	9.67 (0.61)	8.46	10.87
RQ interaction	10.06 (0.38)	9.30	10.81	12.82 (0.42)	11.99	13.54	7.84 (1.83)	4.25	11.45
75th Percentile									
LPOP ($m = 200$)	13 / 13.25 (0.61)	12.05	14.45	16 / 16.18 (0.58)	15.03	17.32	13 / 13.13 (0.49)	12.18	14.09
LPOP ($m = 400$)	13 / 13.31 (0.48)	12.38	14.24	16 / 16.35 (0.48)	15.41	17.30	13 / 13.04 (0.27)	12.51	13.57
RQ additive	12.81 (0.27)	12.29	13.33	16.68 (0.25)	16.19	17.18	13.40 (0.54)	12.34	14.46
RQ interaction	13.35 (0.38)	12.61	14.09	16.38 (0.36)	15.67	17.10	12.70 (1.66)	9.46	15.95

We show estimation results for select quantiles (25th, 50th, and 75th) and compare both our proposed method labeled LPOP and regression quantile methods labeled RQ. We provide details for three subjects characterized by (age, baseline, and disability) as shown in Figure 2. For LPOP, we consider neighborhoods of size $m = 200$ ($\sim 5\%$) and $m = 400$ ($\sim 10\%$), while for RQ, we consider models that use linear splines for both age and baseline disability either adopting an additive structure or including interactions between age and baseline disability to allow more parametric flexibility in order to compare to the non-parametric LPOP estimates. For each scenario, we present the point estimate, SE estimate, and 95% CI. For each method the quantile estimate is presented in bold, and for the LPOP estimates we also present the mean estimate from 1000 bootstrap replicates.

4.3. Illustration: comparison with regression quantile methods

Lastly, we use the BOLD data to compare our localized strategy to alternative parametric strategies. Specifically, in Table 2 we compare our LPOP quantile estimates using neighborhoods of $m = 200$ and $m = 400$ to quantile estimates based on flexible parametric regression using M-estimator methods implemented in the R package `quantreg` (Koenker, 2005). We separately estimate quantile regression functions for the 25th, 50th, and 75th percentiles. In order to use regression quantile methods with continuous predictors we used linear splines with knots at the three quartiles for both age and baseline disability. An additive model therefore uses 9 total parameters, and a more flexible bivariate predictive surface approach includes the additional interactions between the two predictors for a total of 25 parameters. Table 2 captures key expected trade-offs between non-parametric and regression-based parametric approaches. First, the LPOP estimates with $m = 200$ generally have the largest standard error since they are non-parametric and local. Second, there is generally good agreement between the flexible parametric estimates and the LPOP quantile estimates. The key exception to these patterns are estimates for an individual at the extreme of the covariate distribution such as the subject with age = 95. For this index individual, we find large uncertainty associated with the regression estimates and this is contrasted with the clear neighborhood transfer that the LPOP methods intentionally disclose (Figure 2). Finally, comparison with regression strategies highlight the intrinsic challenge of distributional estimation with moderate dimension predictors since flexible multivariate surfaces may generate high-dimensional bases.

5. DISCUSSION

By focusing on semi-supervised axis-parallel neighborhoods, our proposed methods provide an easily calculated estimate of an individual patient's prognosis that is based on a subgroup of subjects chosen to be interpretable by both clinicians and patients. The increasing amount of EHR data that is becoming accessible to inform clinical predictions will allow patients and providers to obtain desired levels of precision despite the non-parametric nature of our method. We recognize that, when analyzing biomedical data, ensuring adequate access control to protected health information is an important concern and therefore would propose that only those who already have access to the database be given access to the tool. The major limitation of our work is that it is suitable for low-dimensional problems only. Due to the sparsity of data as dimension increases, neighborhoods quickly become so large as to be useless. However, marginal scaling and clinician guidance allow us to narrow down the set of candidate predictors.

In the future, we hope to extend our work to comparative prediction, so that patients may obtain an estimate of their prognosis under a variety of alternative treatment options. In addition, we recognize that patients often have multiple longitudinal measurements rather than a single cross-sectional measurement. Therefore, creating methods that can include individual patient histories or trajectories as input predictive data will be an important extension to consider. The inclusion of patient trajectories would require appropriate low dimensional summary measures for data that may be irregularly measured over time. Finally, although we focus on quantitative outcome measures the core methods can be easily extended to survival endpoints.

SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This research was partially supported by National Institutes of Health (R01 HL072966 and UL1 TR002319).

APPENDIX

A. Mathematical details

A.1. Construction of empirical processes

Suppose $V_1, \dots, V_n \sim \text{Uniform}(0, 1)$. Then their empirical distribution function (edf) is

$$\mathbb{G}_n(t) \equiv \frac{1}{n} \sum_{i=1}^n 1_{[0,t]}(V_i), \quad (\text{A.1})$$

where we note that $n\mathbb{G}_n(t) \sim \text{Binomial}(n, t)$ and hence

$$E[\mathbb{G}_n(t)] = t \quad (\text{A.2})$$

$$nCov[\mathbb{G}_n(s), \mathbb{G}_n(t)] = s \wedge t - st. \quad (\text{A.3})$$

The uniform empirical process is then

$$\mathbb{U}_n(t) \equiv \sqrt{n}[\mathbb{G}_n(t) - t]. \quad (\text{A.4})$$

We let $\mathbb{U}$ be a Brownian Bridge and, by the Central Limit Theorem,

$$\mathbb{U}_n \xrightarrow{d} \mathbb{U}. \quad (\text{A.5})$$

Now suppose we have $X_1, \dots, X_n \sim F$ for some distribution function F . Then their edf is

$$\mathbb{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, x]}(X_i) \quad (\text{A.6})$$

and their empirical process is $\sqrt{n}[\mathbb{F}_n(x) - F(x)]$. If we define $X_i^* = F^{-1}(V_i)$ for $i = 1, \dots, n$, where V_i is as above, then $X_1^*, \dots, X_n^* \sim F$ and we have

$$\mathbb{F}_n^*(x) = \mathbb{F}_n(x^*) \quad (\text{A.7})$$

$$= \mathbb{G}_n[F(x^*)] \quad (\text{A.8})$$

$$= \frac{1}{n} \sum_{i=1}^n 1_{[0, x]}[F(X_i^*)] \quad (\text{A.9})$$

and, noting that $G[F(x)] = F(x)$,

$$\sqrt{n}[\mathbb{F}_n^*(x) - F(x)] = \mathbb{U}_n[F(x)]. \quad (\text{A.10})$$

Therefore,

$$\sqrt{n}[\mathbb{F}_n(x) - F(x)] \xrightarrow{d} \mathbb{U}[F(x)]. \quad (\text{A.11})$$

Because $\mathbb{U}$ has a zero mean and known variance, we can use these results to determine how close the empirical cdf should be to the true cdf (i.e. the confidence interval associated with the cdf is based on the variance of $\mathbb{U}$).

Suppose we have independent but not identically distributed random variables, i.e. $X_1, \dots, X_n$ where $X_i \sim F_i$ for $1 \leq i \leq n$. Then our definition of $\mathbb{F}_n$ remains the same, but we introduce the average df

$$\bar{\mathbb{F}}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n F_i(x) \quad (\text{A.12})$$

and rewrite our empirical process as

$$\sqrt{n}[\mathbb{F}_n(x) - \bar{\mathbb{F}}_n(x)] = \mathbb{X}_n[\bar{\mathbb{F}}_n(x)] \quad (\text{A.13})$$

where, letting $\mathbb{G}_n(t) \equiv \frac{1}{n} \sum_{i=1}^n 1_{[0, t]}[F_i(X_i)]$,

$$\mathbb{X}_n(t) \equiv \sqrt{n}[\mathbb{G}_n(t) - t]. \quad (\text{A.14})$$

Letting $G_i(t) = F_i(t) \circ \bar{F}_n^{-1}(t)$, we define $K_n(s, t) \equiv s \wedge t - \sum_{i=1}^n G_i(s)G_i(t)$. That $K_n(s, t) \rightarrow K(s, t)$ for some $K(s, t)$ is necessary and sufficient for $\mathbb{X}_n \Rightarrow \mathbb{X}$ where $\mathbb{X}$ is a normal process with zero means and covariance function K . Note that, for all $\lambda \geq 0$, $P(\|\mathbb{X}\| \leq \lambda) \geq P(\|\mathbb{U}\| \leq \lambda)$ [Shorack and Wellner \(2009\)](#).

A.2. Uncertainty in neighborhood location

For the uncertainty in neighborhood location, we turn to the δ -method [Casella and Berger \(2002\)](#). We assume that $\bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)]$ is differentiable with respect to Δ and that its second derivative with respect to Δ is bounded. We then use a Taylor expansion [Casella and Berger \(2002\)](#) to obtain

$$\sqrt{n} \left\{ \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \hat{\Delta}_n)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)] \right\} = \sqrt{n} \left\{ \frac{\partial}{\partial \Delta} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)] \times (\hat{\Delta}_n - \Delta) + \varepsilon \right\} \quad (\text{A.15})$$

where $\varepsilon = O(n^{-1})$ and hence determine that the asymptotic variance of our expression is

$$\left\{ \frac{\partial}{\partial \Delta} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)] \right\}^2 \sigma_{\Delta}^2 \quad (\text{A.16})$$

where σ_{Δ}^2 is the asymptotic variance of $\sqrt{n}(\hat{\Delta}_n - \Delta)$. Because Δ is simply a quantile of the distance, we may use the formula for the asymptotic variance of a sample quantile [Shorack and Wellner \(2009\)](#) to obtain

$$\sigma_{\Delta}^2 = \frac{p(1-p)}{f_d^2(F_d^{-1}(p))} \quad (\text{A.17})$$

$$= \frac{p(1-p)}{f_d^2(\Delta)}, \quad (\text{A.18})$$

where the second equality is due to the fact that $F_d^{-1}(p) = \Delta$. Combining the above results, our variance is

$$\frac{p(1-p)}{f_d^2(\Delta)} \times \left\{ \frac{\partial}{\partial \Delta} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)] \right\}^2. \quad (\text{A.19})$$

Because the uncertainty due to location is orthogonal to the uncertainty in outcome, we may add the estimation error from each term to obtain the overall estimation error in the fixed percentage scenario. Note that this term is $\sqrt{n}$ while the uncertainty in outcome is $\sqrt{m}$; taking advantage of the fact that $\sqrt{n} = \sqrt{\frac{m}{p}}$, we multiply our variance by p (i.e. multiply our entire expression by $\sqrt{p}$) to obtain a distribution of

$$N \left(0, \frac{p^2(1-p)}{f_d^2(\Delta)} \times \left\{ \frac{\partial}{\partial \Delta} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)] \right\}^2 \right). \quad (\text{A.20})$$

A.3. Uncertainty Due to Estimation of Parameters

Scaling parameter estimation introduces an extra component to the estimation error. We turn to van der Vaart and Wellner [van der Vaart and Wellner \(2007\)](#) to evaluate this component and we obtain, where V_{β}

is the covariance matrix of the scaling parameters,

$$\begin{bmatrix} \frac{\partial}{\partial \beta_1} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta(\beta))] & \cdots & \frac{\partial}{\partial \beta_k} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta(\beta))] \end{bmatrix} V_\beta \begin{bmatrix} \frac{\partial}{\partial \beta_1} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta(\beta))] \\ \vdots \\ \frac{\partial}{\partial \beta_k} \bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta(\beta))] \end{bmatrix}. \quad (\text{A.21})$$

For the uncertainty due to estimation of functions of $\underline{\mathbf{x}}$ (i.e. $\hat{G}(\cdot)$), we turn to the same results to obtain, where V_G is the covariance matrix of the functions,

$$\begin{bmatrix} \frac{\partial}{\partial G_1} \bar{F}[y|G(\underline{\mathbf{x}}) \in N(\underline{\mathbf{x}}^*, \Delta)] & \cdots & \frac{\partial}{\partial G_k} \bar{F}[y|G(\underline{\mathbf{x}}) \in N(\underline{\mathbf{x}}^*, \Delta)] \end{bmatrix} V_G \begin{bmatrix} \frac{\partial}{\partial G_1} \bar{F}[y|G(\underline{\mathbf{x}}) \in N(\underline{\mathbf{x}}^*, \Delta)] \\ \vdots \\ \frac{\partial}{\partial G_k} \bar{F}[y|G(\underline{\mathbf{x}}) \in N(\underline{\mathbf{x}}^*, \Delta)] \end{bmatrix}. \quad (\text{A.22})$$

Note that these terms are correlated with the uncertainty in outcome and hence consideration of the covariance between the terms is also necessary.

A.4. Smoothing scenario

In the smoothing scenario, where we consider $\Delta_n \rightarrow 0$ as $n \rightarrow \infty$, we turn to results from Stütte to handle the uncertainty in outcome (Expression 3.2) Stütte (1986a); see Appendix A.5 for a more detailed explanation of the mathematics. Assuming that $np_n^3 \rightarrow \infty$, we obtain

$$\sqrt{m} \left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] \right) \xrightarrow{d} \mathbb{Q} \quad (\text{A.23})$$

where $\mathbb{Q}$ is a centered Gaussian process with covariance

$$\text{cov}(\mathbb{Q}(s), \mathbb{Q}(t)) = P(Y \leq s \wedge t | \underline{\mathbf{X}} = \underline{\mathbf{x}}^*) - P(Y \leq s | \underline{\mathbf{X}} = \underline{\mathbf{x}}^*) P(Y \leq t | \underline{\mathbf{X}} = \underline{\mathbf{x}}^*). \quad (\text{A.24})$$

The uncertainty in neighborhood is in this case of order $\sqrt{n}$ and hence for this term we rewrite $\sqrt{m}$ as $\sqrt{\frac{m}{n}} \sqrt{n}$. Because $\frac{m}{n} \rightarrow 0$ and

$$\sqrt{n} \left(\bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)] \right) \quad (\text{A.25})$$

converges to a finite distribution, the entire term is asymptotically negligible and we obtain

$$\sqrt{m} \left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)] \right) \xrightarrow{d} \mathbb{Q}, \quad (\text{A.26})$$

which, because $\Delta = 0$, is equivalent to

$$\sqrt{m} \left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - F[y|\underline{\mathbf{x}} = \underline{\mathbf{x}}^*] \right) \xrightarrow{d} \mathbb{Q}. \quad (\text{A.27})$$

A.5. Uncertainty in outcome: a detailed explanation

When evaluating the asymptotic distribution of our estimator, we need to consider the variability in outcome using results from Stütte (1986a), who tells us that the result is a Gaussian process. He first proves the result pointwise and then expands his results to a curve. For ease of explanation, the proof is presented for univariate X , though the results are easily expanded to multivariate $\mathbf{X}$. We let $(X_1, Y_1), (X_2, Y_2), \dots$ be independent random observations with the same distribution function H as $(X, Y) \in \mathbb{R}^2$. Suppose $E(Y^2) < \infty$ and let $\{p_n\}$ will be a sequence of bandwidths converging to zero such that $np_n^3 \rightarrow \infty$ as $n \rightarrow \infty$. We define

$$m_n(x^*) = p_n^{-1} \int yK \left(\frac{F_n(x^*) - F_n(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.28})$$

and

$$\bar{m}_n(x^*) = p_n^{-1} \int yK \left(\frac{F(x^*) - F(x)}{p_n} \right) H(dx, dy) \quad (\text{A.29})$$

where $K(x) = 1_{[-1,0]}(x)$ is a one-sided kernel. Our initial goal is to prove that, where

$$\sigma_0^2 = \text{Var}(Y|X = x^*) \int K^2(u) du, \quad (\text{A.30})$$

we obtain

$$\sqrt{np_n}[m_n(x^*) - \bar{m}_n(x^*)] \xrightarrow{d} N(0, \sigma_0^2) \quad (\text{A.31})$$

for μ -almost all $x^* \in \mathbb{R}$. In other words, we prove our desired result for any given point. We will subsequently expand our proof to the curve as a whole. Our first step is to obtain, via a Taylor expansion,

$$m_n(x^*) = p_n^{-1} \int yK \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.32})$$

$$+ p_n^{-2} \int y[F_n(x^*) - F_n(x) - F(x^*) + F(x)]K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.33})$$

$$+ p_n^{-3} \int y[F_n(x^*) - F_n(x) - F(x^*) + F(x)]^2 \frac{K''(\Lambda)}{2} H_n(dx, dy) \quad (\text{A.34})$$

$$\equiv I_1 + I_2 + I_3, \quad (\text{A.35})$$

where Λ is between $p_n^{-1}[F_n(x^*) - F_n(x)]$ and $p_n^{-1}[F(x^*) - F(x)]$. We will prove that I_3 is negligible and rewrite I_2 in a different form that is asymptotically equivalent. These two changes will allow us to rewrite $m_n(x^*)$ as a whole in such a way that we can prove our theorem. We begin by showing that

$$\sqrt{np_n}I_3 \xrightarrow{p} 0 \quad (\text{A.36})$$

as $n \rightarrow \infty$. Because K vanishes outside of some finite interval, our expansion of $m_n(x^*)$ holds true with integration restricted to those x for which $|F_n(x^*) - F_n(x)| < p_n$. We know that K'' is bounded and that $\limsup_{n \rightarrow \infty} \int |y|H_n(dx, dy) < \infty$ with probability one. By previous results from Stute, we also know that,

over the values of X in question, $\sup \sqrt{np_n^{-1}}[F_n(x^*) - F_n(x) - F(x^*) + F(x)]$ is stochastically bounded as $n \rightarrow \infty$. We combine these facts to obtain our desired result.

Our next (and more involved) step is to rewrite I_2 into a more tractable expression—its asymptotic equivalent,

$$-\sqrt{np_n p_n^{-1}} m(x^*) \int K \left(\frac{F(x^*) - F(x)}{p_n} \right) [F_n(dx) - F(dx)]. \quad (\text{A.37})$$

We will move from y to $m(x)$, then from $H_n(dx, dy)$ to $F(dx)$, then from $m(x)$ to $m(x^*)$, and finally move from the derivative of the kernel to the kernel itself. We begin by defining

$$Z_n \equiv n^{-1} p_n^{-3/2} \sum_{i=1}^n [Y_i - m(X_i)] \cdot [\alpha_n(x^*) - \alpha_n(X_i)] K' \left(\frac{F(x^*) - F(X_i)}{p_n} \right) \quad (\text{A.38})$$

where $\alpha_n(x) = \sqrt{n}[F_n(x) - F(x)]$ denotes the empirical process pertaining to $X_1, \dots, X_n$. We let $\mathcal{F}$ be the σ -field generated by the X -data. Upon proving that $Z_n \rightarrow 0$ as $n \rightarrow \infty$, we obtain

$$(np_n)^{1/2} I_2 = (np_n^3)^{-1/2} \int y [F_n(x^*) - F_n(x) - F(x^*) + F(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.39})$$

$$= (np_n^3)^{-1/2} \int [y - m(x)] [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.40})$$

$$+ (np_n^3)^{-1/2} \int m(x) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.41})$$

$$\asymp (np_n^3)^{-1/2} \int m(x) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy). \quad (\text{A.42})$$

Our next goal is to move from $H_n(dx, dy)$ to $F(dx)$. We define

$$k(x, y) = m(x) K' \left(\frac{F(x^*) - F(x)}{p_n} \right) \{1_{(-\infty, x^*]}(y) - 1_{(-\infty, x]}(y)\} \quad (\text{A.43})$$

with corresponding von Mises statistic

$$T_n = n \int k(x, y) [F_n(dy) - F(dy)] [F_n(dx) - F(dx)] \quad (\text{A.44})$$

$$= \int k(x, y) \alpha_n(dy) \alpha_n(dx). \quad (\text{A.45})$$

By results from previous works, $E[T_n^2] = O(1)$ as $n \rightarrow \infty$ and hence

$$(np_n^3)^{-1/2} T_n = p_n^{-3/2} \int k(x, y) \alpha_n(dy) [F_n(dx) - F(dx)] \quad (\text{A.46})$$

$$\xrightarrow{P} 0 \quad (\text{A.47})$$

because $T(n)$ is bounded and $(np_n^3)^{-1/2} \rightarrow 0$. Thus,

$$(np_n)^{1/2} I_2 \asymp p_n^{-3/2} \int m(x) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.48})$$

$$= p_n^{-3/2} \int k(x, y) \alpha_n(dy) [F_n(dx) - F(dx)] + a_n^{-3/2} \int k(x, y) \alpha_n(dy) F(dx) \quad (\text{A.49})$$

$$\asymp p_n^{-3/2} \int k(x, y) \alpha_n(dy) F(dx) \quad (\text{A.50})$$

where, we jump from the second to the third equality by using the previous result. We then integrate the quantity in the last equality with respect to y to obtain

$$(np_n)^{1/2} I_2 \asymp p_n^{-3/2} \int m(x) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx). \quad (\text{A.51})$$

We will now move from $m(x)$ to $m(x^*)$. By results from previous works,

$$p_n^{-3/2} \int |m(x) - m(x^*)| |\alpha_n(x^*) - \alpha_n(x)| \left| K' \left(\frac{F(x^*) - F(x)}{p_n} \right) \right| F(dx) \xrightarrow{p} 0 \quad (\text{A.52})$$

as $n \rightarrow \infty$ and hence

$$(np_n)^{1/2} I_2 \asymp p_n^{-3/2} \int m(x) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) \quad (\text{A.53})$$

$$= p_n^{-3/2} \int [m(x) - m(x^*)] [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) \quad (\text{A.54})$$

$$+ p_n^{-3/2} \int m(x^*) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) \quad (\text{A.55})$$

$$\asymp p_n^{-3/2} \int m(x^*) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx). \quad (\text{A.56})$$

We then move on to our final goal of converting the derivative of the kernel to the kernel itself. We first note that, because the kernel vanishes outside a bounded region, $\int K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) = 0$. Therefore, recalling that $m(x^*)$ and $\alpha_n(x^*)$ are constant with respect to x and can be extracted from the integral,

$$(np_n)^{1/2} I_2 \asymp p_n^{-3/2} \int m(x^*) [\alpha_n(x^*) - \alpha_n(x)] K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) \quad (\text{A.57})$$

$$= -p_n^{-3/2} m(x^*) \int \alpha_n(x) K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx). \quad (\text{A.58})$$

We then apply integration by parts, i.e. $\int u dv = uv - \int v du$. We assign $u = \alpha_n(x)$ and $dv = K' \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx)$. Therefore,

$$uv = -p_n \alpha_n(x) K' \left(\frac{F(x^*) - F(x)}{p_n} \right) \quad (\text{A.59})$$

$$= 0 \quad (\text{A.60})$$

because $\alpha_n(x) = 0$ when evaluated at the limits of integration—here, the lower and upper bound of the region outside which the kernel vanishes—due to the fact that the empirical distribution of F and F itself are equivalent at 0 and 1. Thus, only the $\int v du$ term remains, and we obtain

$$(np_n)^{1/2} I_2 = -p_n^{-1/2} m(x^*) \int K \left(\frac{F(x^*) - F(x)}{p_n} \right) \alpha_n(x) \quad (\text{A.61})$$

as desired. We are now in a position to substitute I_1 and our rewritten I_2 for $m_n(x^*)$ in the left-hand side of our theorem to obtain I_4 , which we prove converges in distribution to the desired quantity. Observe that

$$I_4 \equiv \left(\frac{n}{p_n} \right)^{1/2} \int [y - m(x^*)] K \left(\frac{F(x^*) - F(x)}{p_n} \right) [H_n(dx, dy) - H(dx, dy)] \quad (\text{A.62})$$

$$= -(np_n)^{1/2} p_n^{-1} m(x^*) \int K \left(\frac{F(x^*) - F(x)}{p_n} \right) [H_n(dx, dy) - H(dx, dy)] \quad (\text{A.63})$$

$$+ (np_n)^{1/2} p_n^{-1} \int y K \left(\frac{F(x^*) - F(x)}{p_n} \right) H_n(dx, dy) \quad (\text{A.64})$$

$$- (np_n)^{1/2} p_n^{-1} \int y K \left(\frac{F(x^*) - F(x)}{p_n} \right) H(dx, dy) \quad (\text{A.65})$$

$$= -(np_n)^{1/2} p_n^{-1/2} m(x^*) \int K \left(\frac{F(x^*) - F(x)}{p_n} \right) \alpha_n(dx) + I_1 - \bar{m}_n(x^*) \quad (\text{A.66})$$

$$\asymp \sqrt{np_n} (I_2 + I_1 - \bar{m}_n(x^*)) \quad (\text{A.67})$$

$$\asymp \sqrt{np_n} (m_n(x^*) - \bar{m}_n(x^*)). \quad (\text{A.68})$$

We then note that, for each n , I_4 is a standardized sum of i.i.d. random variables with (using $\text{Var}(x) = E[x^2] - E[x]^2$)

$$\text{Var}(I_4) = p_n^{-1} \left\{ \int (y - m(x^*))^2 K^2 \left(\frac{F(x^*) - F(x)}{p_n} \right) H(dx, dy) - \right. \quad (\text{A.69})$$

$$\left. \left[\int (y - m(x^*)) K \left(\frac{F(x^*) - F(x)}{p_n} \right) H(dx, dy) \right]^2 \right\} \quad (\text{A.70})$$

$$= p_n^{-1} \left\{ \int h(x) K^2 \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) - \right. \quad (\text{A.71})$$

$$\left. \left[\int (m(x) - m(x^*)) K \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx) \right]^2 \right\} \quad (\text{A.72})$$

where $h(x) = E[(Y - m(x^*))^2 | X = x]$. We, then, use results from previous works to show that, asymptotically, we can substitute $h(x^*)$ for $h(x)$ and that the second term is negligible, and hence $\text{Var}(I_4) \rightarrow h(x^*) \int K^2(u) du$ for μ -almost all $x^* \in \mathbb{R}$. Thus, it suffices to show that the array defining the I_4 's satisfies the Lindeberg condition for μ -almost all x^* ; in other words, to prove that, as $n \rightarrow \infty$, and for all $\delta > 0$

$$p_n^{-1} \int_{|y - m(x^*)| \geq \delta \sqrt{np_n}} (y - m(x^*))^2 K^2 \left(\frac{F(x^*) - F(x)}{p_n} \right) H(dx, dy) \rightarrow 0. \quad (\text{A.73})$$

Because $np_n \rightarrow \infty$, the above will follow if, with $a > 0$ and

$$h_a(x) = E[(Y - m(x^*))^2 1_{|Y - m(x^*)| > a} | X = x] \quad (\text{A.74})$$

we can make the following expression arbitrarily small if a large enough a is chosen:

$$\limsup_{n \rightarrow \infty} p_n^{-1} \int h_a(x) K^2 \left(\frac{F(x^*) - F(x)}{p_n} \right) F(dx). \quad (\text{A.75})$$

From standard results in differentiation theory, the above is equivalent to $h_a(x^*) \int K^2(u) du$ for μ -almost all $x^* \in \mathbb{R}$, and hence may be made small by letting $a \rightarrow \infty$. In order to guarantee that m is equicontinuous in a neighborhood about d_0 , and thus that the standardized process m_n has continuous sample paths, we assume that

$$\sup_{||t-s|| \leq \delta} |m(t|x) - m(s|x)| = o((\ln \delta^{-1})^{-1}) \quad \text{as } \delta \rightarrow 0. \quad (\text{A.76})$$

To derive distributional results for m_n , we must also assume that H has uniform marginals. Then for Lebesgue-almost all $0 < x^* < 1$

$$(np_n)^{1/2} [m_n(\cdot|x^*) - \tilde{m}_n(\cdot|x^*)] \xrightarrow{d} \mathbb{Q} \equiv \mathbb{Q}(x^*). \quad (\text{A.77})$$

Here $\mathbb{Q}$ is a centered Gaussian process on $[0, 1]$ with continuous sample paths vanishing at the lower boundary of $[0, 1]$ and covariance

$$\text{cov}(\mathbb{Q}(y_1), \mathbb{Q}(y_2)) = [m(y_1 \wedge y_2|x^*) - m(y_1|x^*)m(y_2|x^*)] \int K^2(u) du. \quad (\text{A.78})$$

REFERENCES

- CASELLA, G. AND BERGER, R. L. (2002). *Statistical Inference*, Volume 2. Pacific Grove CA: Duxbury.
- Chang, W., Cheng, J., Allaire, J. J., Xie, Y. and McPherson, J. (2017). shiny: Web Application Framework for R. R package version 1.0.5. <https://CRAN.R-project.org/package=shiny>.
- D'AGOSTINO, R. B., VASAN, R. S., PENCINA, M. J., WOLF, P. A., COBAIN, M., MASSARO, J. M. AND KANNEL, W. B. (2008). General cardiovascular risk profile for use in primary care the Framingham heart study. *Circulation* **117**, 743–753.
- DICKSON, E. R., GRAMBSCH, P. M., FLEMING, T. R., FISHER, L.D. AND LANGWORTHY, A. (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* **10**, 1–7.
- EGAN, T. M., MURRAY, S., BUSTAMI, R. T., SHEARON, T. H., MCCULLOUGH, K. P., EDWARDS, L. B., COKE, M. A., GARRITY, E. R., SWEET, S. C., HEINEY, D. A. and others. (2006). Development of the new lung allocation system in the united states. *American Journal of Transplantation* **6**, 1212–1227.
- HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. H. (2009). *The Elements of Statistical Learning*. New York: Springer.
- HILL, J. C., DUNN, K. M., LEWIS, M., MULLIS, R., MAIN, C. J., FOSTER, N. E. AND HAY, E. M. (2008). A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Care & Research* **59**, 632–641.
- IBRAHIM, K. A., PANETH, N., LAGAMMA, E. AND REED, P. L. (2009). Clinician opinion to design clinical trials that change standards-of-cares. *Pediatric Research*.

- JARVIK, J. G., COMSTOCK, B. A., BRESNAHAN, B. W., NEDELJKOVIC, S., NERENZ, D. R., BAUER, Z., AVINS, A. L., JAMES, K., TURNER, J. A., HEAGERTY, P. J. *and others*. (2012). Study protocol: The back pain outcomes using longitudinal data (BOLD) registry. *BMC Musculoskeletal Disorders* **13**, 64.
- JHA, A. K. (2010). Meaningful use of electronic health records: the road ahead. *JAMA* **304**, 1709–1710.
- KAMATH, P. S., WIESNER, R. H., MALINCHOC, M., KREMERS, W., THERNEAU, T. M., KOSBERG, C. L., D’AMICO, G., DICKSON, E. R. AND KIM, W. R. (2001). A model to predict survival in patients with end-stage liver disease. *Hepatology* **33**, 464–470.
- KEIDING, N. AND LOUIS, T. A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**, 319–376.
- KOENKER, R. (2005). *Quantile Regression*, Number 38. Cambridge, England: Cambridge University Press.
- KOENKER, R. AND BASSETT, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, **46**, 33–50.
- LEVY, W. C., MOZAFFARIAN, D., LINKER, D. T., SUTRADHAR, S. C., ANKER, S. D., CROPP, A. B., ANAND, I., MAGGIONI, A., BURTON, P., SULLIVAN, M. D. *and others*. (2006). The Seattle heart failure model prediction of survival in heart failure. *Circulation* **113**, 1424–1433.
- PETTY, R. E. AND CACIOPPO, J. T. (1986). *The Elaboration Likelihood Model of Persuasion*. New York, NY: Springer.
- R CORE TEAM. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- SEGAL, M. AND KEDEM, K. (1998). Geometric applications of posets. *Computational Geometry* **11**, 143–156.
- SHORACK, G. R. AND WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics*, Volume 59. Philadelphia, PA: SIAM.
- SKIENA, S. S. (2009). *The Algorithm Design Manual*. New York, NY: Springer.
- STÜTE, W. (1986a). Conditional empirical processes. *The Annals of Statistics* **14**, 638–647.
- STÜTE, W. (1986b). On almost sure convergence of conditional empirical distribution functions. *The Annals of Probability* **14**, 891–901.
- VAN DER VAART, A. W. AND WELLNER, J. A. (2007). Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, 234–252.
- VOS, T., FLAXMAN, A. D., NAGHAVI, M., LOZANO, R., MICHAUD, C., EZZATI, M., SHIBUYA, K., SALOMON, J. A., ABDALLA, S., ABOYANS, V. *and others*. (2013). Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet* **380**, 2163–2196.

[Received April 2, 2017; revised March 12, 2018; accepted for publication March 18, 2018]