Local Estimation of Patient Prognosis

Alison Kosel

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Patrick Heagerty, Chair

Marco Carone

Robyn McClelland

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Local Estimation of Patient Prognosis

Alison Kosel

Chair of the Supervisory Committee:
Professor Patrick Heagerty
Department of Biostatistics

Statistical methods that can provide patients and their healthcare providers with individual predictions are needed so that informed medical decisions can be made. Ideally an individual prediction would display the full range of possible outcomes (full predictive distribution), would be obtained with a 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 of interest. We then provide non-parametric estimates of the outcome distribution for subjects in this neighborhood, which we refer to as a localized prediction. We implement the local prediction method using dynamic graphical methods that allow the user to vary key options such as the choice of neighborhood variables and the size of the neighborhood. Furthermore, we expand our method to handle multiple treatment groups and longitudinal data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

BALANCING SCORE: A balancing score is a score such that the conditional distribution of covariates is the same for the treatment and control group given that the balancing score is the same in each group [11].

BOLD: The Back pain Outcomes using Longitudinal Data study consists of roughly 5,000 patients age 65 or older who present with a complaint of back pain. The goal is to improve back pain outcomes in these patients by comparing the effectiveness of diagnostic and treatment strategies [20].

BROWNIAN BRIDGE: A Brownian bridge is a Gaussian process $\mathbb{U}$ with continuous sample paths and

1. $\mathbb{U}(0) = \mathbb{U}(1) = 0$;
2. $E[\mathbb{U}(t)] = 0$ for $0 \leq t \leq 1$;
3. $\text{Cov}[\mathbb{U}(s)\mathbb{U}(t)] = s \wedge t - st$ for $0 \leq s, t \leq 1$.

A Brownian bridge is almost surely continuous and nowhere differentiable [38].

BROWNIAN MOTION: Standard Brownian motion is defined as a Gaussian process $\mathbb{S}$ with continuous sample paths and

1. $\mathbb{S}(0) = 0$;
2. $E[\mathbb{S}(t)] = 0$ for $0 \leq t \leq 1$;
3. $\text{Cov}[\mathbb{S}(s)\mathbb{S}(t)] = s \wedge t$ for $0 \leq s, t \leq 1$.

Brownian motion is almost surely continuous and nowhere differentiable [38].

EMPIRICAL BAYES: Empirical Bayes methods are Bayesian methods in which the prior distribution is estimated from the data [33].

GAUSSIAN PROCESS: A Gaussian process is a stochastic process with the property that any finite linear combination of such processes has a Gaussian distribution or is identically zero [38].

NEIGHBORHOOD: A neighborhood $N(\Delta, \underline{\mathbf{x}}^*)$ encompasses all points within a distance Δ of a point $\underline{\mathbf{x}}^*$.

PROPENSITY SCORE: The propensity score is defined as the conditional probability of treatment given the covariates. It is the coarsest possible balancing score [34].

Chapter 1

INTRODUCTION

Patients who understand their individualized prognosis are more likely to work collaboratively with their physician 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 [28]. 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 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 [8]. Prognostic scores are also used in other medical settings such as organ transplantion. For example, the lung allocation score assigns priority to lung transplant recipients in the United States based on patient characteristics such as age, clincal status, and specific diagnostic categories that predict both the risk of death without intervention and survival if the patient is transplanted [9]. 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 [7]. More recently, the Model for End-Stage Liver Disease (MELD) assesses chronic liver disease severity to determine priority for liver transplant recipients [24]. These 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 individual patient predictions that are easily interpretable by both clinicians and patients and that do not depend on any model as-

sumptions. Non-parametric estimation typically requires large sample sizes. We expect the feasibility of such direct estimation approaches will increase with the adoption of electronic health records (EHRs) prompted by recent government initiatives. The Health Information Technology for Economic and Clinical Health (HITECH) Act, enacted in 2009, allocates roughly $30 billion to promote the adoption of health information technology and incentivizes its use [22].

Understanding how individuals perceive information in order to make decisions is essential for developing effective and appropriate statistical summaries. Psychological research has determined 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 [32]. Furthermore, research has 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. One aspect that has been shown to impact uptake is the sample size used for a clinical study, with a larger sample size leading to a greater probability of adopting new interventions [18]. Therefore, we seek a non-parametric method that allows the user to directly control the number of observations within a local neighborhood of subjects, and we think it is important to disclose the exact region/neighborhood that is used for the prediction so that personal relevance can be subjectively judged by the patient and/or provider.

1.1 Current Predictive Methods

We begin with an overview of current methods for prognostic research. Such methods can be roughly divided along two axes: what they estimate and how they estimate it. Estimands range from a single number, such as the mean, to the entire distribution function, while estimators use techniques ranging from parametric to non-parametric. If we are willing to make parametric assumptions, we can easily use regression and splines to estimate the mean of our outcome. Common kernel methods allow us to do the same but are non-parametric. Quantile regression is a parametric method that allows us to estimate the entire distribution of the outcome. However, there is a lack of non-parametric methods that allow estimation of the entire distribution of the outcome, and that is the gap we fill. Suppose we have $Y_i \in \mathbb{R}$ and $\underline{\mathbf{X}}_i \in \mathbb{R}^k$ as random variables measuring the outcome

and covariates, respectively, for each subject $1 \leq i \leq n$, and that we wish to locally estimate $f(\underline{\mathbf{x}}^*) = F(Y|\underline{\mathbf{X}} = \underline{\mathbf{x}}^*)$. Although a method like ordinary least squares regression might be our first thought, there are several other methods available.

1.1.1 Splines

In many cases in which we would like to perform regression, we do not truly believe that $f(\underline{\mathbf{x}})$ is linear in $\underline{\mathbf{X}}$ and we therefore desire a more flexible representation of $\underline{\mathbf{X}}$. We can accomplish this by using piecewise polynomials (i.e., dividing the domain of $\underline{\mathbf{X}}$ into regions and modeling each region with a separate polynomial function). The resulting function can be as simple as a different mean in each region, or we can use higher order polynomials and continuity constraints to obtain smoother functions. In general, an order-M spline has continuous derivatives to order $M - 2$; however, unless one has a particular interest in the derivatives, there is typically no need to go beyond cubic splines, with $M = 4$. For these regression splines, the order of the spline as well as the location and number of the knots must be chosen [14].

Another type of spline is the smoothing spline, which requires the selection of a penalty (typically chosen via cross-validation) and always includes the maximal set of knots; most commonly used is the natural cubic spline with knots at all unique values of $\underline{\mathbf{X}}$, which uniquely minimizes over all $f(\underline{\mathbf{x}})$ with two continuous derivatives the following expression:

$$\sum_{i=1}^{N} [y_i - f(\underline{\mathbf{x}}_i)]^2 + \lambda \int f''(t)^2 dt. \tag{1.1}$$

This is in fact a generalized ridge regression [14]. Interestingly, the smoothing spline estimator is asymptotically equivalent to the local average kernel estimator, which we discuss next [10].

With splines, the undesirable behavior of polynomials near the boundaries of the data is particularly bad; extrapolation is extremely inadvisable. Natural cubic splines take advantage of this limitation to free up four degrees of freedom by constraining the function to be linear beyond the boundary knots. Given the dearth of data in this region, this is often a reasonable assumption, and it allows more degrees of freedom to be spent modeling the interior.

1.1.2 Kernel Methods

We may wish to be still less parametric than splines allow, in which case we can use kernel methods. This broad class of methods, which typically estimate the conditional mean using local information, use a kernel that can be described as

$$K_\lambda(\underline{\mathbf{x}}^*, \underline{\mathbf{x}}) = D\left(\frac{||\underline{\mathbf{x}} - \underline{\mathbf{x}}^*||}{h_\lambda(\underline{\mathbf{x}}^*)}\right) \tag{1.2}$$

where λ is a parameter (typically selected by cross-validation) that dictates the size of the neighborhood, $h_\lambda(\underline{\mathbf{x}}^*)$ is a function that uses λ to determine the width of the neighborhood, and $D(\cdot)$ depends on the kernel in question. Note that larger values of λ lead to larger bias and smaller variance; likewise, smaller values of λ lead to smaller bias and larger variance. Strips of constant width (i.e. $h_\lambda(\underline{\mathbf{x}}^*)$ fixed) lead to fixed bias and variance inversely proportional to local density, while strips of constant sample size lead to fixed variance and bias inversely proportional to local density. Common choices of $D(\cdot)$ include the Gaussian kernel, the Epanechnikov kernel, and the tri-cube kernel.

In order to estimate $f(\underline{\mathbf{x}})$, we can use our kernel to compute a locally weighted average, such as the Nadaraya-Watson average,

$$\hat{f}(\underline{\mathbf{x}}^*) = \frac{\sum_{i=1}^n K_\lambda(\underline{\mathbf{x}}^*, \underline{\mathbf{x}}_i) y_i}{\sum_{i=1}^n K_\lambda(\underline{\mathbf{x}}^*, \underline{\mathbf{x}}_i)}, \tag{1.3}$$

which is local constant estimation. Alternatively, local linear regression can be used to correct the bias often found on the edges of the domain; for each $\underline{\mathbf{x}}^*$ it solves

$$\min_{\alpha(\underline{\mathbf{x}}^*), \beta(\underline{\mathbf{x}}^*)} \sum_{i=1}^n K_\lambda(\underline{\mathbf{x}}^*, \underline{\mathbf{x}}_i)[y_i - \alpha(\underline{\mathbf{x}}^*) - \beta(\underline{\mathbf{x}}^*)\underline{\mathbf{x}}_i]^2 \tag{1.4}$$

and the estimate is then $\hat{f}(\underline{\mathbf{x}}^*) = \hat{\alpha}(\underline{\mathbf{x}}^*) + \hat{\beta}(\underline{\mathbf{x}}^*)\underline{\mathbf{x}}^*$. Essentially, the least squares operation and the selected kernel are combined into a new kernel that corrects the bias exactly to first order. Local polynomial regression expands on this idea by allowing terms of higher order in the regression, i.e.,

$$\min_{\alpha(\underline{\mathbf{x}}^*),\beta_j(\underline{\mathbf{x}}^*),j=1,\ldots,d} \sum_{i=1}^{n} K_\lambda(\underline{\mathbf{x}}^*, \underline{\mathbf{x}}_i) \left[y_i - \alpha(\underline{\mathbf{x}}^*) - \sum_{j=1}^{d} \beta_j(\underline{\mathbf{x}}^*) \underline{\mathbf{x}}_i^j \right]^2 . \tag{1.5}$$

While these higher order fits further reduce bias, they do so at the cost of increased variance [14]. Additionally, nearest neighbors falls into the category of kernel methods, with $\lambda = m$, where m is the size of the neighborhood, and $h_\lambda(\underline{\mathbf{x}}^*) = |\underline{\mathbf{x}}^* - \underline{\mathbf{x}}_{[m]}|$, where $\underline{\mathbf{x}}_{[m]}$ is the mth closest point to $\underline{\mathbf{x}}^*$.

Kernel methods that estimate the entire distribution function do exist. Stüte, for example, proposes a non-parametric local kernel method that does exactly this [42]. However, this method uses standard bandwidth selection techniques that balance bias and variance automatically, bases neighborhood size on the curvature of the function, and does not typically disclose the support of the estimate to the user [1]. Given that our goal is to construct meaningful and interpretable covariate neighborhoods in a semi-supervised manner, this method is insufficient for our needs.

1.1.3 Quantile Regression

Quantile regression, in which the (possibly asymmetrically weighted) absolute value of the residuals is minimized, allows any conditional quantile to be estimated. However, it requires that a separate regression be performed for each quantile and it requires a parametric model [26]. Suppose we wish to find the τth conditional quantile of Y. Then, defining $\rho_\tau(y) = |y(\tau - I_{y<0})|$, we calculate

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(y_i - \underline{\mathbf{x}}_i^T \beta). \tag{1.6}$$

Unlike least-squares regression, quantile regression does not have a closed form solution. However, it enjoys the same equivariance properties as least-squares regression and (additionally) equivariance to monotone transformations [25].

1.1.4 Remarks

Conditional predictions are a focus of many contemporary statistical learning methods that allow great flexibility in the breadth of candidate input predictive information [14]. However, most re-

gression 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 a product of the methods. Many clinical risk prediction calculators have been created in recent years, but these tools rarely give information back to the user about the underlying data's 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 patient neighbors that are used to inform individual prediction.

1.2 Scope of the Dissertation

Our proposed methods are a variation on non-parametric local kernel methods studied by Stüte [42]. 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. By defining our distance function on the basis of covariate relationships with the outcome we are proposing a semi-supervised predictive method. In addition, standard bandwidth selection methods for local estimation typically consider predictive criteria that balance bias and variance while we see 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.

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 select details such as the percentage of subjects with very good or very bad outcomes are important for decision makers. In addition, patients can easily understand statements such as: "Twenty-five percent 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 key percentiles of the full conditional distribution.

In sum, we will

- develop a non-parametric method to estimate the entire empirical distribution of the outcome for a neighborhood about a given patient in an experimental setting;
- extend our methodology to quasi-experimental settings;
- and extend our methodology to longitudinal settings.

We will verify the inference for these estimators using simulation studies and illustrate them using the BOLD data [20], which were collected and managed using REDCap electronic data capture tools hosted at the University of Washington [13].

Chapter 2

LOCALIZED PATIENT OUTCOME PREDICTION

We introduce a simple method to perform local non-parametric estimation based on a neighborhood that is selected to have a fixed sample size (precision) and to be axis-parallel, yielding a simple 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 a distribution estimate and a neighborhood specification that was used to provide the estimate. Our choice is to transfer from a specific covariate point to a select subgroup of patients. In order to understand our idea, consider an outcome, Y, and a set of covariates, $\underline{\mathbf{X}}_1, \ldots, \underline{\mathbf{X}}_n$. Typically one estimates the outcome either for all patients (i.e., $F(y)$) or for a single patient characterized by exact covariates values (i.e., $F(y|\underline{\mathbf{X}} = \underline{\mathbf{x}}^*)$). We choose an intermediate target, where we let $N(\Delta, \underline{\mathbf{x}}^*)$ be a neighborhood of distance Δ about the covariate vector $\underline{\mathbf{x}}^*$, and we can then observe that

$$N(0, \underline{\mathbf{x}}^*) \subset N(\Delta, \underline{\mathbf{x}}^*) \subset N(\infty, \underline{\mathbf{x}}^*). \tag{2.1}$$

By specifying the distance Δ we can 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 neighborhood with 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 and report a variable neighborhood size, Δ, depending on the density of points around an index point $\underline{\mathbf{x}}^*$. By returning both a prediction and the neighborhood used, we clearly present the data's ability to answer the question of interest: if there are very few data points near a given patient, the patient and/or clinician will be made aware of this by the

increased size of the neighborhood.

Although a neighborhood $N(\Delta, \underline{\mathbf{x}}^*)$ could in principle be any shape, we choose to restrict our consideration to axis-parallel neighborhoods, or covariate rectangles. We do so for ease of interpretation; axis-parallel neighborhoods can be described by an independent restriction on each coordinate of $\underline{\mathbf{X}}$, and are therefore easily understood by medical professionals and patients alike. We present a class of options based on the chosen distance metric

$$d(p, A, G, \underline{\mathbf{x}}^*, \underline{\mathbf{x}}) = ||A[G(\underline{\mathbf{x}}) - G(\underline{\mathbf{x}}^*)]||_p. \tag{2.2}$$

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 $p = 2$; L1 distance, with $A = I$, $G = I$, and $p = 1$; and nearest neighbors, with G as the empirical distribution function variable-wise. However, many of these choices are problematic when it comes to interpretability, as we will see shortly.

One might be tempted to consider characteristics of the neighborhood such as volume and balance. However, it is clearly seen that volume, defined as the product of the distances along each axis, is a problematic criterion when we consider the case of discrete data, in which there are many ties. For m sufficiently small, we can obtain a volume of zero simply by picking points along the line of one of the coordinates (i.e., a distance of zero along one axis). In other words, we would have optimized our criterion by restricting one covariate to a single point and placing no restrictions on the other covariates, which is presumably less than ideal in practice. While balance in theory sounds like an excellent criterion to optimize, there are again problems in practice. Aside from the problem of defining balance (one can consider how many points lie above and below $\underline{\mathbf{x}}^*$ in each coordinate direction or one can consider the distance away from x^* in each direction the points lie on average), the neighborhood only becomes particularly unbalanced when the data are sparse in a given direction from x^*. Thus, correcting the balance of a neighborhood is sometimes impossible (if $\underline{\mathbf{x}}^*$ is on the edge of the data, for example), or comes at the cost of a dramatic increase in the size of the neighborhood.

2.1 Selection of Distance Function Details

2.1.1 Selection of the Distance Shape or Norm Defined by p

In theory, one can choose any metric and then take the rectangular enclosure of the points chosen by that metric. In fact, Lowsky and colleagues introduce a non-parametric method using nearest neighbors using Mahalanobis distance [29]. However, this method has three main disadvantages when compared to our method: interpretation is more difficult, due to the oblong shape of the distance function; the user has no control over the precision of the estimate, as the number of neighbors is chosen by cross-validation; and the covariates are scaled by their own variance rather than by the strength of their relationship with the outcome (we will come to this shortly). While the use of metrics that do not initially create an axis-parallel neighborhood is possible, specifying m is in these cases difficult. The axis-parallel enclosure of the neighborhood created by these metrics is not the same as the neighborhood created by these metrics; points outside the neighborhood are included in or points inside the neighborhood are excluded from the enclosure of the neighborhood. The L_1 metric, for example, creates a diamond, while the L_2 metric creates a circle; creating an axis-parallel neighborhood out of these would require either taking the largest axis-parallel rectangle contained inside the diamond/circle or the smallest axis-parallel rectangle containing the diamond/circle, thereby including extra points in the corners of a rectangle or removing points along the axes (see Figure 2.1). Clearly neither of these is ideal when controlling precision is one's primary goal. Thus, for the creation of an axis-parallel neighborhood, one metric stands out as the clear winner: the supremum norm, L_∞. When using this metric, $N(\Delta, \underline{\mathbf{x}}^*)$ consists of exactly the points within the initially selected area.

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 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

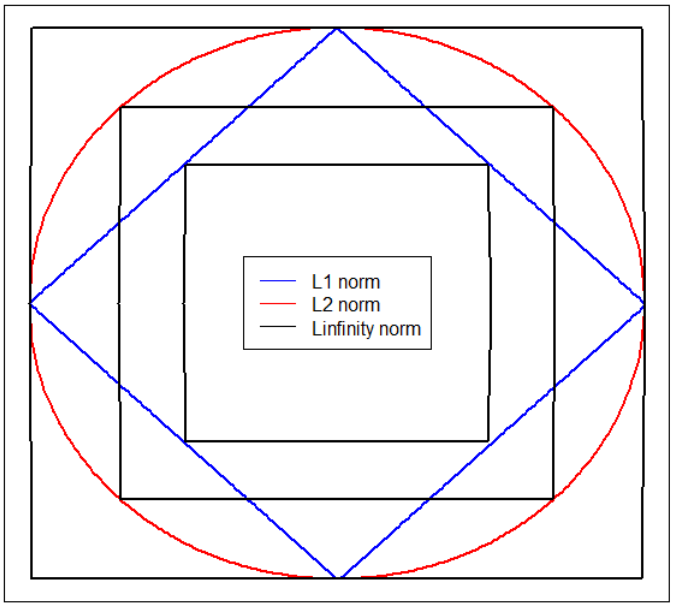

Figure 2.1: **Distance Metric Shapes** In order to create an axis-parallel neighborhood, points would need to be added to the corners or removed from the sides of shapes created by the L_1 and L_2 metrics, whereas the L_∞ metric returns the proper shape automatically.

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 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.

2.1.2 Selection of the Distance Scaling or Matrix A

While we have settled on a choice of p 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 coordinate

as equivalent. However, this is undesirable in most cases. For example, one would generally not wish to equate a one millimeter of mercury change in systolic blood pressure with a one unit change in body mass index. More generally, one would not wish the trade between dimensions to be so heavily dependent on the units of measurement. Unscaled, height would be weighed very differently in meters and inches; clearly this is unreasonable and senseless from a practical perspective. The question, therefore, is not whether to scale the data, but how.

There are two main categories of methods to rescale the data: outcome-independent methods and outcome-dependent methods. In other words, do we rescale based on the covariates alone or do we take their relationship to the outcome into consideration? We consider the former type of method in the next section, in which we consider choice of G, and the latter type of method here. To correct the problem of differing scales among covariates, we propose outcome-based marginal rescaling. To implement this, one simply regresses y on $\underline{\mathbf{x}}_j$ for each $1 \leq j \leq k$ individually. This produces a $\hat{\beta}[j]$ for each $1 \leq j \leq k$, by which one multiplies all $\underline{\mathbf{x}}_j$ and $\underline{\mathbf{x}}_j^*$; this equates to $A = \text{diag}(\hat{\beta})$. In other words, one scales each coordinate based on its marginal association with the outcome. Coordinates that are strongly associated with the outcome have their distances expanded, while coordinates that are weakly associated with the outcome have their distances shrunk. This overcomes the problem of units that is seen in unscaled data and more tightly controls more predictive covariates.

While there would certainly be benefits to a regression using all covariates at once, there would also be drawbacks. Highly collinear covariates would be subject to decreased or uncertain importance, which is undesirable. Surely if two covariates are both important then we wish to match strongly on both of them, regardless of their relationship to one another. In addition, there would be no single relationship between each covariate and the outcome – the relationship would change depending on the other covariates included. With our method, each covariate is assigned a single level of importance and can easily be compared to other covariates one might wish to add to the model.

2.1.3 Selection of the Distance using Covariate Transformations

In this section we consider outcome-independent methods to rescale the data, such as standardization via the use of percentiles or standard deviations, which involve a transformation of the covariates (i.e., a choice of G). The advantage of percentiles is that a one unit change in any direction includes the same amount of data, while the advantage of standard deviation units is that the distances within each coordinate are maintained while those between coordinates are standardized. Additionally, transformations such as the logarithm that are traditionally used in regression can be chosen. 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 standard deviations. Furthermore, we use transformation functions in the expansions to our method.

2.2 Functionals

In addition to estimating $F[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]$, we can also estimate functionals of $F[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]$. Note that our method allows nearly any choice of functional to be estimated because we nonparametrically obtain an estimate of the entire empirical distribution for our axis-parallel neighborhood. Thus, one could select, for example, the quartiles of the distribution, or any other set of quantiles, and obtain them easily and at once.

2.3 Algorithm

Our algorithm details a simple procedure to find the axis-parallel neighborhood of m points about $\underline{\mathbf{x}}^*$. 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 then a neighborhood containing the desired number of closest points is created. The choices of A and G are left to the user.

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 $\underline{\mathbf{X}}[,j]$ on Y, let the slope coefficient be $\hat{\beta}[j]$, and A be the diagonal matrix created by $\hat{\beta}$ (outcome-based rescaling). Let $G(\underline{\mathbf{x}})$ be any desired function, such as identity or percentile. For $1 \leq i \leq n$, calculate $d_i = ||A[G(\underline{\mathbf{x}}_i) - G(\underline{\mathbf{x}}^*)]||_\infty$.

3. Where F_d is the distribution of distances, calculate $q_d = F_d^{-1}(p)$. Discard all $\underline{\mathbf{x}}_i$ such that $d_i > q_d$; call the remaining points $N(\Delta, \underline{\mathbf{x}}^*)$.

4. Find the enclosure of $N(\Delta, \underline{\mathbf{x}}^*)$. For $1 \leq j \leq k$, let $V[j,1] = \min(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$ and $V[j,2] = \max(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$.

2.4 Ties

In discrete data there is the additional problem of ties – many data points may have the exact same value for one or more covariates. In some cases, therefore, obtaining precisely the desired level of precision is impossible. Using the L_∞ metric can produce substantially more points than desired. Subsequently applying a second metric, such as L_1, can reduce the number of points somewhat by trimming the furthest corners of the rectangle. However, in cases of highly discrete data, there are simply too many points with exactly the same coordinates to obtain exactly the number of points requested, and an approximate solution must be accepted.

2.5 Computational Complexity

2.5.1 Single $\underline{x}^*$

Typically we would perform calculations dynamically; that is, we would calculate our estimates upon the user's request. 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 x^* and the n other

points has time complexity $O(n)$. We then must sort the distances. R uses quicksort, which averages $O(n \log n)$ and has a worst case scenario of $O(n^2)$ [39]. Thus, the sorting dominates the cost; overall we average $O(n \log n)$ asymptotically, with a worst case scenario of $O(n^2)$.

2.5.2 All $\underline{\mathbf{x}}$

While referencing an external, static database, we might wish to have all of the 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 as $\underline{\mathbf{x}}^*$ in turn. Segal and Kedem describe an algorithm to obtain the m nearest rectilinear neighbors of a given point; however, they require that $m \geq \frac{n}{2}$ [37].

2.6 Inference

When performing inference, we continue with our transfer from point to neighborhood and consider what happens to our neighborhood asymptotically. We term this a fixed percentage scenario, as we consider the neighborhood of $p \times 100\%$ of the data about $\underline{\mathbf{x}}^*$ as our data accumulates. In other words, we compare our estimate to the results we would obtain if we could locate our neighborhood perfectly and if we knew the outcome in that neighborhood exactly. Letting $\underline{\mathbf{X}} \sim G$, we define $\bar{F}(y|\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)) = \frac{1}{\int g(\underline{\mathbf{x}}) 1_{N(\Delta, \underline{\mathbf{x}}^*)}(\underline{\mathbf{x}}) d\underline{\mathbf{x}}} \int F(y|\underline{\mathbf{x}}) g(\underline{\mathbf{x}}) 1_{N(\Delta, \underline{\mathbf{x}}^*)}(\underline{\mathbf{x}}) d\underline{\mathbf{x}}$. Suppose we consider

$$\sqrt{m}\left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]\right) \tag{2.3}$$

$$= \sqrt{m}\left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] \right. \tag{2.4}$$

$$\left. + \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]\right) \tag{2.5}$$

where we decompose the scaled estimation error into two orthogonal components: that due to uncertainty in outcome (Expression 2.4) and that due to uncertainty in neighborhood location (Expression 2.5). We label the uncertainty in outcome, $\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)]$, as O and the uncertainty in location, $\bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)]$, as L and then the orthogonality of these two components is easily shown:

$$E[OL] = E[E[OL|\underline{\mathbf{X}}]] \tag{2.6}$$
$$= E[E[O|\underline{\mathbf{X}}]L] \tag{2.7}$$
$$= E[E[O|\underline{\mathbf{X}}]]E[L] \tag{2.8}$$
$$= E[O]E[L] \tag{2.9}$$

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 2.4), which, using results from Shorack and Wellner [38], converges to $\mathbb{X}$ as $m \to \infty$, 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 $G_{mi} = F_{mi} \circ \bar{F}_m^{-1}$. 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.

We may then focus on the uncertainty in neighborhood (Expression 2.5), 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 $\underline{\mathbf{x}}^*$,

$$\sqrt{m}\left(\bar{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)]\right) \xrightarrow{d} \tag{2.10}$$
$$N\left(0, \frac{p^2(1-p)}{f_d^2(\Delta)} \times \left\{\frac{d}{d\Delta}\bar{F}[y|x \in N(x^*, \Delta)]\right\}^2\right). \tag{2.11}$$

Therefore,

$$\sqrt{m}\left(\hat{F}[y|\underline{\mathbf{X}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)]\right) \xrightarrow{d} \tag{2.12}$$
$$\mathbb{X} + N\left(0, \frac{p^2(1-p)}{f_d^2(\Delta)} \times \left\{\frac{d}{d\Delta}\bar{F}[y|x \in N(x^*, \Delta)]\right\}^2\right). \tag{2.13}$$

In Appendix A.4 we also provide inference for those who wish to compare our estimate to the truth at the point 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 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 $p_\infty = 0$, $N(\Delta, \underline{\mathbf{x}}^*) = \underline{\mathbf{x}}^*$, and our target is the point rather than the neighborhood.

2.6.1 Effect of Scaling Parameter Estimation

In the above inference, we assume no scaling, or the use of a known scaling parameter. The use of an estimated scaling parameter introduces an additional component to the estimation error because we now consider

$$\sqrt{m}\left(\hat{F}[y|\underline{\mathbf{x}} \in N(\hat{\Delta}_n(\hat{\beta}), \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{x}} \in N(\Delta(\beta), \underline{\mathbf{x}}^*)]\right). \tag{2.14}$$

This additional component can be evaluated using results from van der Vaart and Wellner [43]. Provided that the behavior of the distribution is not pathological, we obtain

$$\sqrt{m}\sqrt{p}\left[\bar{F}(y \mid ||\hat{\beta}(\underline{\mathbf{x}} - \underline{\mathbf{x}}^*)||_\infty \leq \hat{\Delta}_n) - \bar{F}(y \mid ||\beta(\underline{\mathbf{x}} - \underline{\mathbf{x}}^*)||_\infty \leq \hat{\Delta}_n)\right]. \tag{2.15}$$

Note that this term is not orthogonal to the uncertainty in outcome (Expression 2.4). 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, though its exact value can be obtained via the δ-method. 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; $p \to 0$ 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 not be used in the fixed percentage scenario.

2.7 Implementation and Example

2.7.1 Example

We use the Back pain Outcomes using Longitudinal Data (BOLD) project to motivate our research [20]. This data set consists of roughly 5,000 patients age 65 or older who present with a complaint of back pain; the goal is to improve back pain outcomes in these patients by comparing the effectiveness of diagnostic and treatment strategies. Disability is measured using the Roland scale, which asks 24 yes/no questions about how back pain impacts the patient's daily life. Previously, the STarT Back Screening Tool has been used to classify patients with low back pain into three risk groups based on prognostic indicators [16]. We aim to take this a step further by developing a method that provides physicians and patients with a personalized estimate of the patient's expected outcome.

2.7.2 Implementation

We used the package Shiny in R to implement our method in two dimensions [31] [35]. We include a histogram to estimate the empirical distribution, as well as the quartiles and mean of the empirical distribution, both overall and for the neighborhood of interest. For our example, we 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 one year (roland_3) as our outcome. In Figure 2.2, note that baseline disability is much more strongly related than age to disability at one year – the neighborhood is much slimmer on the disability axis than the age axis when scaled. If the data were not scaled, we would obtain a neighborhood with an age range of 71 to 79 and a baseline disability range of 6 to 14, in place of our scaled neighborhood ranging from 67 to 83 in age and 9 to 11 in baseline disability, with an outcome distribution of similar shape but slightly worse overall. In our scaled neighborhood, which is what we would use in practice, we observe a large number of patients who recover completely (i.e. return to a disability score of 0) after one year, and note that the rest 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 cannot adequately inform a clinician or patient's expectations. It is important to be able to clearly visualize the space of possible outcomes.

We can also examine slightly different patients, as seen in Figure 2.3. When we consider a patient with a higher baseline disability score – 15 instead of 10 – we see that the neighborhood remains the same size and that the outcome maintains its bimodal shape, but the second peak is shifted to the right. 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. Even fixing m, we obtain a much larger 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 neighborhood cannot extend in that direction. These are not so much downsides as features of our method. While nearly all methods have difficulty on the boundary of the data, with our method the user is clearly informed when the data about a given patient is sparse and when he might wish to be particularly cautious in his interpretation.

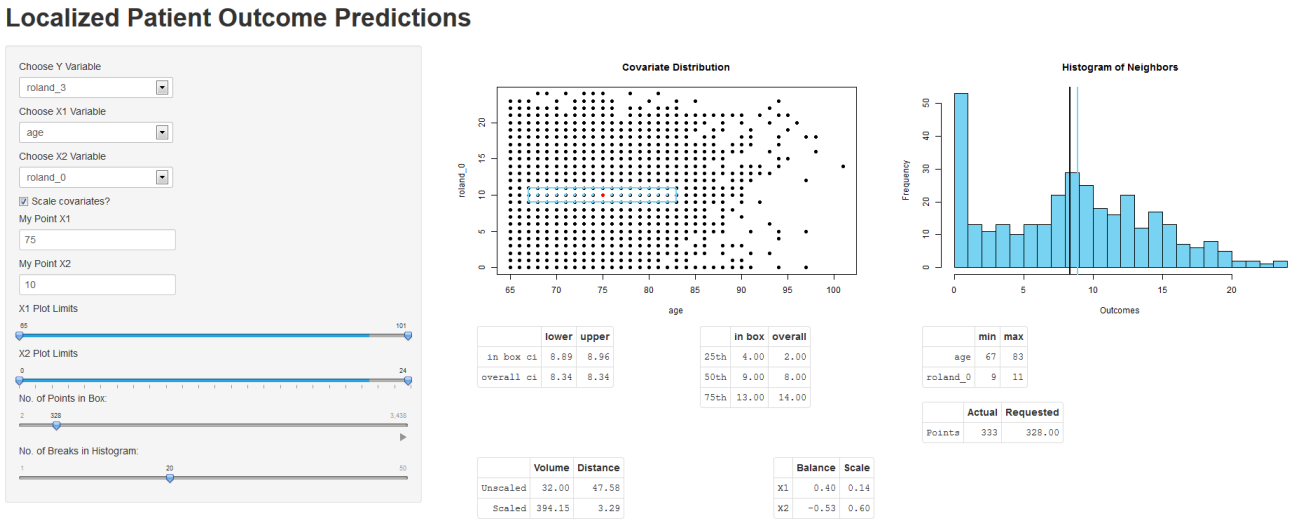

Figure 2.2: **Example of Tool** In this example of our tool, we display the outcome for a 75-year-old patient who has a baseline disability of 10 based on the BOLD data.

2.8 Simulations

We conduct simulations in order to evaluate the adequacy of asymptotic approximations in finite samples as well as the estimation of the uncertainty due to neighborhood location and scaling parameters. Because we recommend fixed percentage inference, this is where we focus our simulations. Because we focus on the entire outcome distribution, our goal is to obtain valid simultaneous confidence intervals for the distribution. We do so for the distribution scaled by its variance (i.e., $t(1-t)$ for $F(t)$, $0 < t < 1$) in order to more equally place error tolerance along the range of the distribution. We first demonstrate our method in the most basic scenario, univariate X, as proof of concept. We then demonstrate our method in the presence of two covariates so that we may include a scaling parameter. We considered simulated data sets of 5000, 10000, and 20000 subjects and neighborhoods containing 5%, 10%, and 20% of the subjects.

2.8.1 Determination of Critical Values

We begin by simulating Brownian motion on the interval [0,1] at points $0 = t_0, t_1, \ldots, t_k, t_{k+1} = 1$, where we define $\Delta t_i \equiv t_i - t_{i-1}$. We first generate k independent standard normal random variables, $Z_1, \ldots, Z_k$, and our path is then generated by the formula

$$\mathbb{S}(t_k) = \sum_{i=1}^{k} \sqrt{\Delta_i} Z_i. \tag{2.16}$$

From our Brownian motion, $\mathbb{S}$, we can generate a Brownian bridge, $\mathbb{U}$, via the following formula [6]:

$$\mathbb{U}(t_k) = \mathbb{S}(t_k) - t_k \mathbb{S}(1). \tag{2.17}$$

We simulated 100,000 such Brownian bridges with 5,000 points each, scaled them by $t_i \times (1 - t_i)$, took the maximum of each bridge, and then set the 95^{th} percentile of the maxima as our critical value for the scaled Brownian bridges in our simulations. Note that even this is an approximation of the true critical value in the cases where additional terms need to be added to the Brownian

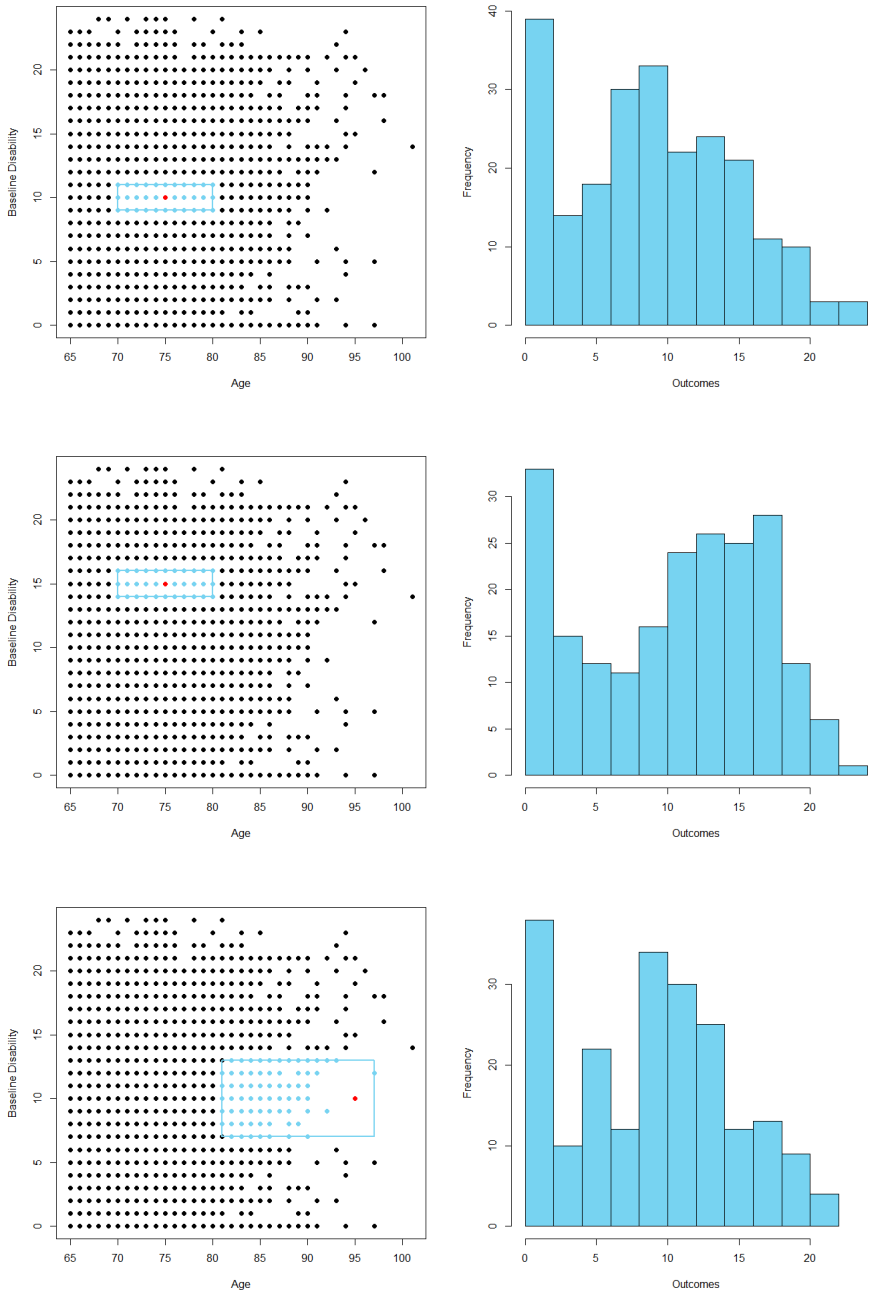

Figure 2.3: **Example Neighborhoods in BOLD Data** In this example of our method, we use the BOLD data to display the change in neighborhood and the change in outcome distribution when a new patient is selected. We begin with a patient who has a baseline disability of 10 and an age of 75, then consider 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.

bridge, as in the case of neighborhood uncertainty, though it appears to be sufficiently accurate in practice.

2.8.2 Calculation of True Parameters

Supposing that $\underline{\mathbf{X}} \sim G$ and $Y|\underline{\mathbf{X}} = \underline{\mathbf{x}} \sim F$, determining the true value of Δ and obtaining the true cdf within $N(\underline{\mathbf{x}}^*, \Delta)$ is quite straightforward. The true distance that the neighborhood extends about $\underline{\mathbf{x}}^*$ is the Δ such that the following constraint is met:

$$p = G(\underline{\mathbf{x}}^* + \beta\Delta) - G(\underline{\mathbf{x}}^* - \beta\Delta). \tag{2.18}$$

We use this constraint to solve for Δ numerically. The marginal distribution of Y within $N(\Delta, \underline{\mathbf{x}}^*)$ can be obtained via numerical integration and is

$$F_{N(\Delta,\underline{\mathbf{x}}^*)}(y) = \frac{1}{\int g(\underline{\mathbf{x}}) 1_{[\underline{\mathbf{x}}^*-\beta\Delta,\underline{\mathbf{x}}^*+\beta\Delta]}(\underline{\mathbf{x}}) d\underline{\mathbf{x}}} \int F(y|\underline{\mathbf{x}}) g(\underline{\mathbf{x}}) 1_{[\underline{\mathbf{x}}^*-\beta\Delta,\underline{\mathbf{x}}^*+\beta\Delta]}(\underline{\mathbf{x}}) d\underline{\mathbf{x}}. \tag{2.19}$$

2.8.3 Calculation of Uncertainty in Neighborhood Location

In our simulations, we evaluated the additional variance component for each scenario at $y = -2, -1, 0, 1, 2, 3$ for $\mathbb{R}^1$ and $y = -1, 0, 1, 2, 3, 4, 5$ for $\mathbb{R}^2$. When considering mixture outcomes in $\mathbb{R}^2$, we calculated the additional component for every integer y value between -5 and 10 for mixture one, -15 and 10 for mixture two, and -25 and 10 for mixture three. We added the extra variance component from the estimated y value closest to the observed y value at each point to the variance in outcome for our simultaneous confidence intervals. While one could in theory evaluate the exact additional variance at every observed y value, this would be quite intensive computationally, and our approach provides sufficiently accurate results.

We estimated f_d from distances based on 10000 simulated $\underline{\mathbf{X}}$ in $\mathbb{R}^1$ and 1000 simulated $\underline{\mathbf{X}}$ in $\mathbb{R}^2$. We plugged these estimates into the formula for σ^2_Δ (see Appendix A.2) for each scenario. To estimate $\frac{\partial}{\partial\Delta}\bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)]$, we obtained the true Δ for an equidistant sequence of 100 values of p between $.025$ and $.225$, calculated $\bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)]$ for each Δ, and then took a left-handed

derivative estimate based on these two sequences. Note that this derivative is identically zero when the distribution is symmetric about x^* (in our case, this occurs at $x^* = 0$ in $\mathbb{R}^1$ and $x^* = (0, 0)$ in $\mathbb{R}^2$). We then combined the above estimates as described in Appendix A.2.

2.8.4 Calculation of Uncertainty due to Parameter Estimation

As when estimating the uncertainty in neighborhood location, we evaluated the additional variance component at $y = -1, 0, 1, 2, 3, 4, 5$ and added the extra variance component from the estimated y value closest to the observed y value. When considering mixture outcomes, we calculated the additional component for every integer y value between -5 and 10 for mixture one, -15 and 10 for mixture two, and -25 and 10 for mixture three.

We simulated 1000 $\underline{\mathbf{X}}$ and their corresponding Y values, from which we estimated σ^2_β. We estimated $\frac{\partial}{\partial\beta}\bar{F}[y|\underline{\mathbf{x}} \in N(\underline{\mathbf{x}}^*, \Delta)]$ by performing a left-handed approximation of the derivative at the true β and then, for $i = 1, 2$, at $\beta_i - .01$, and dividing the difference in each pair by $.01$. Sensitivity analyses showed this method to provide adequate levels of accuracy. The above two terms were then multiplied, as detailed in Appendix A.3.

2.8.5 Simulations when $X \in \mathbb{R}^1$

We begin by considering the simplest scenario: univariate X. We let

$$X \sim G \equiv \mathbf{N}(0, 1) \tag{2.20}$$

$$Y|X = x \sim F \equiv \mathbf{N}(x, 1). \tag{2.21}$$

Our goal is to assess the coverage of our confidence intervals, both simultaneous and at several points across the range of the outcome, and our estimation of Δ. In all scenarios we conducted $10,000$ simulations. We considered $x^* = 0$, $x^* = 1$, and $x^* = 2$ to confirm that our method performs well for patients of interest both at the center and at the edge of the covariate distribution.

In all scenarios our estimates of Δ were unbiased and had quite low variability; as expected, the standard deviations decreased as n increased in each scenario (in the $x^* = 0$ case from .004 to

.003 to .002 for $p = .05$, from .005 to .004 to .003 for $p = .10$, and from .007 to .005 to .004 for $p = .20$, with similar trends for $x^* = 1$ and $x^* = 2$, see Tables 2.1, 2.2, and 2.3 for details). Note that the further from the center of the distribution our chosen point, the larger the Δ – neighbors become sparser and therefore one needs to travel further to obtain a given number of them.

We were almost always within 1-2% of the nominal coverage levels, both for our pointwise confidence intervals at the quartiles of the distribution and for our simultaneous confidence intervals; simultaneous coverage ranged from .938 to .959 for $x^* = 0$, .925 to .954 for $x^* = 1$, and .955 to .964 for $x^* = 2$ (see Tables 2.1, 2.2, and 2.3 for details). We also include a visual of the empirical distributions as compared to the true distributions in Figure 2.4; note the even spread of the uncertainty across the range of Y (due to our scaling of the cdf) and that the estimated curves fall along the desired path. Thus, our inferential methods perform as stated. The empirical distribution of outcomes in the neighborhood is as similar as expected to the true distribution not only overall, but also at points across the distribution.

We next examine our method under more challenging outcome distributions. We followed the same procedures and covariate distribution as in our previous simulations with the following mixture of normal distributions as the outcome:

$$w_i \sim \text{bernoulli}(r) \tag{2.22}$$

$$Y_i \sim X_i + w_i \times N\left(\frac{s}{r}, 4\right) + (1 - w_i) \times N\left(-\frac{s}{1-r}, 1\right) \tag{2.23}$$

We selected three progressively more challenging combinations of r and s: mixture one, .5 and 1; mixture two, .8 and 1.5; and mixture three, .9 and 2. The densities for these distributions can be seen in Figure 2.5. Because the covariate distribution is identical to our previous example, we do not include Δ estimation results in these cases. We also restrict our assessment to simultaneous coverage, as we have already established that coverage is at acceptable levels across the range of outcome values.

When $x^* = 0$, we see very good coverage across the board – in all three examples it generally ranges from .94 to .96, though we dip slightly below that in the $n = 20,000$, $p = .20$ example

x^*	n	p	Δ	$\hat{\Delta}$	CI 25%	CI 50%	CI 75%	CI Simul
0	5000	0.05	0.063	0.063 (0.004)	0.955	0.956	0.954	0.959
0	10000	0.05	0.063	0.063 (0.003)	0.960	0.962	0.956	0.952
0	20000	0.05	0.063	0.063 (0.002)	0.956	0.966	0.960	0.944
0	5000	0.10	0.126	0.126 (0.005)	0.959	0.962	0.960	0.951
0	10000	0.10	0.126	0.126 (0.004)	0.959	0.968	0.956	0.951
0	20000	0.10	0.126	0.126 (0.003)	0.963	0.966	0.963	0.938
0	5000	0.20	0.253	0.253 (0.007)	0.966	0.962	0.966	0.946
0	10000	0.20	0.253	0.253 (0.005)	0.961	0.970	0.965	0.944
0	20000	0.20	0.253	0.253 (0.004)	0.955	0.970	0.958	0.940

Table 2.1: **Simulation Results for $\mathbb{R}^1$ with $x^* = 0$** This table displays the results of our simulations in $\mathbb{R}^1$ with $x^* = 0$. Estimates of Δ are presented as mean (sd), while confidence interval results are empirical coverage of 95% confidence intervals.

(Table 2.4). When $x^* = 1$, coverage remains between 94% and 96% at all times (Table 2.5). When $x^* = 2$, we obtain slight overcoverage; coverage ranges from roughly .95 to .97 (Table 2.6). A visual confirmation of our coverage results can be seen in Figure 2.6, where we display the first 100 empirical cdfs for each mixture as compared to the true cdf for that mixture. It is clear that our method performs well even under quite trying outcome distributions.

2.8.6 Simulations when $\underline{\boldsymbol{X}} \in \mathbb{R}^2$

In this scenario we let

$$\underline{\mathbf{X}} \sim G \equiv \mathrm{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) \tag{2.24}$$

$$Y|\underline{\mathbf{X}} = \underline{\mathbf{x}} \sim F \equiv \mathrm{N}\left(\begin{pmatrix}1 & 2\end{pmatrix} \cdot \underline{\mathbf{x}}, 1\right). \tag{2.25}$$

x^*	n	p	Δ	$\hat{\Delta}$	CI 25%	CI 50%	CI 75%	CI Simul
1	5000	0.05	0.103	0.103 (0.006)	0.953	0.958	0.958	0.954
1	10000	0.05	0.103	0.103 (0.004)	0.961	0.955	0.960	0.953
1	20000	0.05	0.103	0.103 (0.003)	0.959	0.957	0.960	0.949
1	5000	0.10	0.207	0.207 (0.009)	0.956	0.962	0.958	0.950
1	10000	0.10	0.207	0.207 (0.006)	0.962	0.969	0.958	0.947
1	20000	0.10	0.207	0.207 (0.004)	0.966	0.966	0.958	0.943
1	5000	0.20	0.413	0.413 (0.012)	0.950	0.963	0.953	0.941
1	10000	0.20	0.413	0.413 (0.008)	0.962	0.963	0.960	0.939
1	20000	0.20	0.413	0.413 (0.006)	0.968	0.967	0.970	0.925

Table 2.2: **Simulation Results for $\mathbb{R}^1$ with $x^* = 1$** This table displays the results of our simulations in $\mathbb{R}^1$ with $x^* = 1$. Estimates of Δ are presented as mean (sd), while confidence interval results are empirical coverage of 95% confidence intervals.

x^*	n	p	Δ	$\hat{\Delta}$	CI 25%	CI 50%	CI 75%	CI Simul
2	5000	0.05	0.425	0.424 (0.022)	0.955	0.945	0.952	0.964
2	10000	0.05	0.425	0.425 (0.016)	0.952	0.953	0.948	0.958
2	20000	0.05	0.425	0.425 (0.011)	0.951	0.962	0.965	0.959
2	5000	0.10	0.736	0.735 (0.022)	0.953	0.952	0.955	0.964
2	10000	0.10	0.736	0.736 (0.016)	0.950	0.959	0.950	0.960
2	20000	0.10	0.736	0.736 (0.011)	0.963	0.963	0.962	0.956
2	5000	0.20	1.161	1.161 (0.020)	0.946	0.946	0.944	0.961
2	10000	0.20	1.161	1.161 (0.014)	0.949	0.953	0.956	0.959
2	20000	0.20	1.161	1.161 (0.010)	0.950	0.962	0.960	0.955

Table 2.3: **Simulation Results for $\mathbb{R}^1$ with $x^* = 2$** This table displays the results of our simulations in $\mathbb{R}^1$ with $x^* = 2$. Estimates of Δ are presented as mean (sd), while confidence interval results are empirical coverage of 95% confidence intervals.

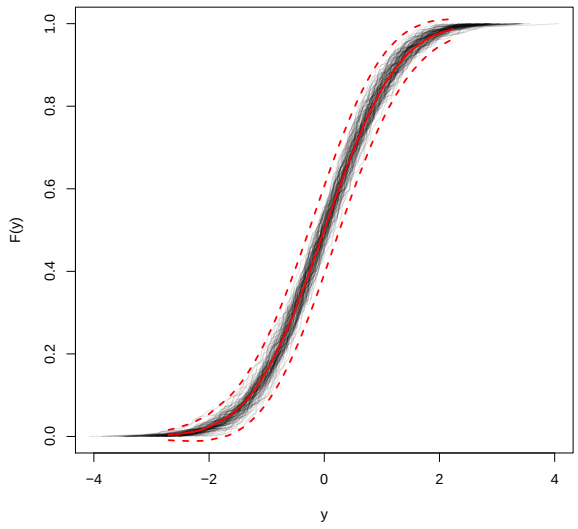

Figure 2.4: **Empirical Outcome Distributions** Here we display the first 100 empirical cdfs (the remainder are redacted for clarity) from the $x^* = 0$, $n = 5000$, $p = .05$ case with normal outcome, along with the true cdf and its 95% simultaneous confidence bounds. Results from other scenarios are similar.

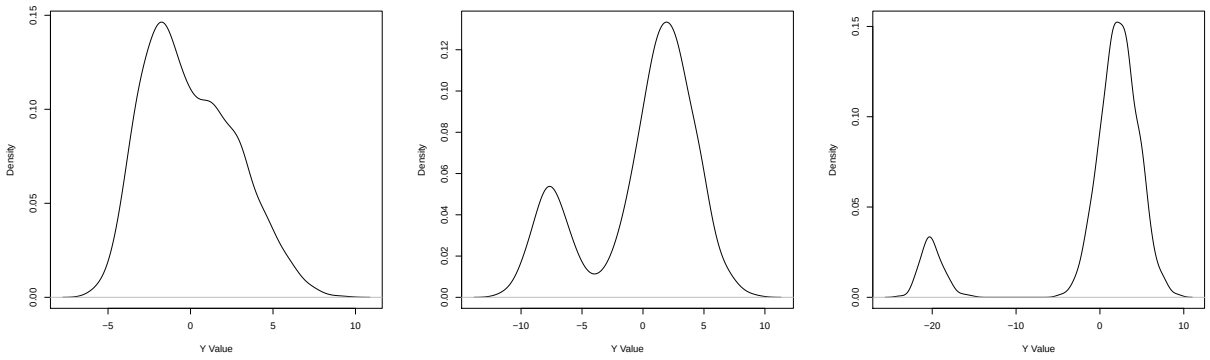

Figure 2.5: **Mixture Outcome Densities** Here we display the outcome densities for, from left to right, outcome mixtures one, two, and three.

x^*	n	p	Mix 1	Mix 2	Mix 3
0	5000	0.05	0.958	0.957	0.955
0	10000	0.05	0.956	0.949	0.952
0	20000	0.05	0.948	0.944	0.945
0	5000	0.10	0.951	0.955	0.953
0	10000	0.10	0.949	0.952	0.949
0	20000	0.10	0.944	0.941	0.943
0	5000	0.20	0.948	0.948	0.943
0	10000	0.20	0.944	0.940	0.941
0	20000	0.20	0.934	0.936	0.937

Table 2.4: **Simulation Results for Mixture Distributions for $\mathbb{R}^1$ with $x^* = 0$** This table displays the results of our simulations with normal mixture distributions as the outcome in $\mathbb{R}^1$ with $x^* = 0$. Confidence interval results are empirical coverage of 95% simultaneous confidence intervals.

x^*	n	p	Mix 1	Mix 2	Mix 3
1	5000	0.05	0.954	0.958	0.959
1	10000	0.05	0.952	0.949	0.946
1	20000	0.05	0.949	0.943	0.943
1	5000	0.10	0.952	0.954	0.954
1	10000	0.10	0.953	0.945	0.947
1	20000	0.10	0.946	0.944	0.940
1	5000	0.20	0.955	0.950	0.953
1	10000	0.20	0.953	0.947	0.951
1	20000	0.20	0.953	0.946	0.950

Table 2.5: **Simulation Results for Mixture Distributions for $\mathbb{R}^1$ with $x^* = 1$** This table displays the results of our simulations with normal mixture distributions as the outcome in $\mathbb{R}^1$ with $x^* = 1$. Confidence interval results are empirical coverage of 95% simultaneous confidence intervals.

x^*	n	p	Mix 1	Mix 2	Mix 3
2	5000	0.05	0.958	0.965	0.964
2	10000	0.05	0.961	0.956	0.956
2	20000	0.05	0.958	0.960	0.961
2	5000	0.10	0.967	0.965	0.964
2	10000	0.10	0.968	0.960	0.961
2	20000	0.10	0.963	0.957	0.957
2	5000	0.20	0.971	0.958	0.963
2	10000	0.20	0.971	0.957	0.962
2	20000	0.20	0.970	0.958	0.960

Table 2.6: **Simulation Results for Mixture Distributions for $\mathbb{R}^1$ with $x^* = 2$** This table displays the results of our simulations with normal mixture distributions as the outcome in $\mathbb{R}^1$ with $x^* = 2$. Confidence interval results are empirical coverage of 95% simultaneous confidence intervals.

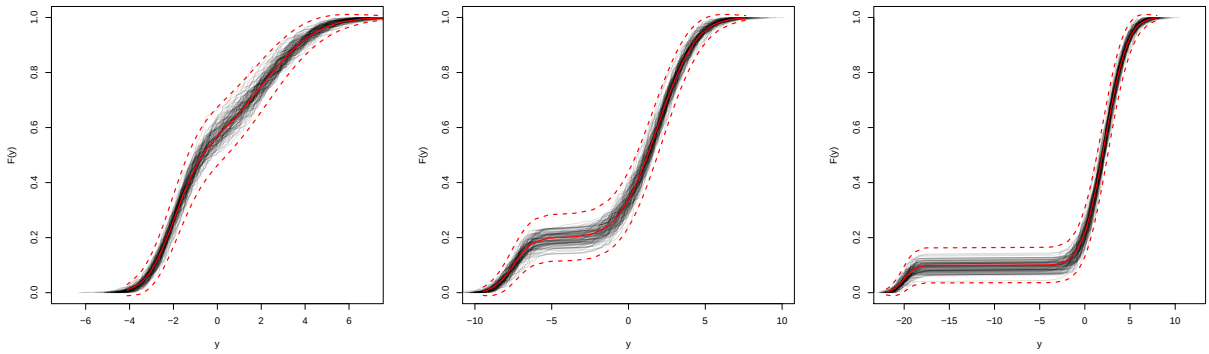

Figure 2.6: **Empirical Outcome Distributions** Here we display the first 100 empirical cdfs for, from left to right, mixture distributions one, two, and three (the remainder are redacted for clarity) from the $x^* = 0$, $n = 5000$, $p = .05$ case, along with the true cdf and the associated 95% simultaneous confidence bounds. Results from other scenarios are similar.

Having already confirmed that our method produces accurate coverage across the range of Y, our goal here is twofold: to demonstrate that we estimate Δ well in the presence of multivariate covariates and that we obtain good overall coverage in the presence of estimated parameters. In all scenarios we conducted $1,000$ simulations. We considered $x^* = (0,0)$ and $x^* = (1,1)$ to assess our method for a subject at the center of the data and for a more unusual subject.

Our estimation of Δ was again both precise and unbiased, in both the scaled and unscaled cases. As in $\mathbb{R}^1$, the point further away from the center of the distribution had a larger Δ, roughly half again the size in some cases (.284 versus .462 for $p = .05$ in the unscaled scenario). Our coverage was again extremely close to nominal levels, ranging from .93 to .95 for the unscaled and .94 to .97 for the scaled $x^* = (0,0)$ and from .94 to .96 for the unscaled and .95 to .97 for the scaled $x^* = (1,1)$ (see Tables 2.7, 2.8, 2.9, and 2.10 for details). Thus, we have confirmed that our method performs well in the presence of multivariate covariates and estimated parameters.

As in $\mathbb{R}^1$, we also examined our method under more challenging outcome distributions. We followed the same procedures and covariate distribution as in our previous simulations with the following mixture of normal distributions as the outcome:

$$w_i \sim \text{bernoulli}(r) \tag{2.26}$$

$$Y_i \sim \underline{\mathbf{X}}_i[1] + 2\underline{\mathbf{X}}_i[2] + w_i \times N\left(\frac{s}{r}, 4\right) + (1 - w_i) \times N\left(-\frac{s}{1-r}, 1\right). \tag{2.27}$$

We selected three progressively more challenging combinations of r and s: mixture one, .5 and 1; mixture two, .8 and 1.5; and mixture three, .9 and 2. We considered only coverage, as the Δ estimation is precisely the same as before. (In the unscaled scenario, Δ is related to $\underline{\mathbf{X}}$ only. In the scaled scenario, the relationship between $\underline{\mathbf{X}}$ and Y remains of the same strength.) In all scenarios our coverage hovers right around .95, ranging from .93 to .96 in each of the unscaled scenarios and .94 to .97 in each of the scaled scenarios (see Tables 2.11, 2.12, 2.13 and 2.14). Our method continues to perform well under difficult outcome distributions.

x^*	n	p	Δ	$\hat{\Delta}$	CI Simul
(0, 0)	5000	0.05	0.284	0.284 (0.009)	0.95
(0, 0)	10000	0.05	0.284	0.284 (0.006)	0.95
(0, 0)	20000	0.05	0.284	0.284 (0.004)	0.95
(0, 0)	5000	0.10	0.407	0.407 (0.009)	0.95
(0, 0)	10000	0.10	0.407	0.407 (0.006)	0.94
(0, 0)	20000	0.10	0.407	0.407 (0.005)	0.94
(0, 0)	5000	0.20	0.594	0.594 (0.009)	0.94
(0, 0)	10000	0.20	0.594	0.593 (0.006)	0.94
(0, 0)	20000	0.20	0.594	0.594 (0.005)	0.93

Table 2.7: **Unscaled Simulation Results for $\mathbb{R}^2$ with $x^* = (0, 0)$: Normal Outcome** This table displays the results of our simulations in $\mathbb{R}^2$ with $x^* = (0, 0)$ and no scaling.

x^*	n	p	Δ	$\hat{\Delta}$	CI Simul
(1, 1)	5000	0.05	0.462	0.462 (0.014)	0.96
(1, 1)	10000	0.05	0.462	0.463 (0.010)	0.95
(1, 1)	20000	0.05	0.462	0.462 (0.007)	0.95
(1, 1)	5000	0.10	0.655	0.656 (0.014)	0.95
(1, 1)	10000	0.10	0.655	0.655 (0.010)	0.95
(1, 1)	20000	0.10	0.655	0.655 (0.007)	0.96
(1, 1)	5000	0.20	0.934	0.934 (0.013)	0.96
(1, 1)	10000	0.20	0.934	0.934 (0.010)	0.96
(1, 1)	20000	0.20	0.934	0.934 (0.007)	0.94

Table 2.8: **Unscaled Simulation Results for $\mathbb{R}^2$ with $x^* = (1, 1)$: Normal Outcome** This table displays the results of our simulations in $\mathbb{R}^2$ with $x^* = (1, 1)$ and no scaling.

x^*	n	p	Δ	$\hat{\Delta}$	CI Simul
(0, 0)	5000	0.05	0.403	0.403 (0.015)	0.96
(0, 0)	10000	0.05	0.403	0.403 (0.011)	0.96
(0, 0)	20000	0.05	0.403	0.403 (0.007)	0.94
(0, 0)	5000	0.10	0.580	0.580 (0.017)	0.97
(0, 0)	10000	0.10	0.580	0.579 (0.012)	0.96
(0, 0)	20000	0.10	0.580	0.580 (0.009)	0.97
(0, 0)	5000	0.20	0.852	0.852 (0.021)	0.95
(0, 0)	10000	0.20	0.852	0.851 (0.015)	0.97
(0, 0)	20000	0.20	0.852	0.853 (0.010)	0.97

Table 2.9: **Scaled Simulation Results for $\mathbb{R}^2$ with $x^* = (0, 0)$: Normal Outcome** This table displays the results of our simulations in $\mathbb{R}^2$ with $x^* = (0, 0)$ and global scaling.

x^*	n	p	Δ	$\hat{\Delta}$	CI Simul
(1, 1)	5000	0.05	0.654	0.654 (0.022)	0.97
(1, 1)	10000	0.05	0.654	0.654 (0.015)	0.96
(1, 1)	20000	0.05	0.654	0.655 (0.011)	0.95
(1, 1)	5000	0.10	0.929	0.929 (0.024)	0.96
(1, 1)	10000	0.10	0.929	0.930 (0.017)	0.97
(1, 1)	20000	0.10	0.929	0.929 (0.012)	0.96
(1, 1)	5000	0.20	1.335	1.333 (0.028)	0.96
(1, 1)	10000	0.20	1.335	1.336 (0.021)	0.97
(1, 1)	20000	0.20	1.335	1.334 (0.014)	0.96

Table 2.10: **Scaled Simulation Results for $\mathbb{R}^2$ with $x^* = (1, 1)$: Normal Outcome** This table displays the results of our simulations in $\mathbb{R}^2$ with $x^* = (1, 1)$ and global scaling.

x^*	n	p	Mix 1	Mix 2	Mix 3
(0, 0)	5000	0.05	0.96	0.96	0.96
(0, 0)	10000	0.05	0.95	0.95	0.93
(0, 0)	20000	0.05	0.95	0.94	0.95
(0, 0)	5000	0.10	0.96	0.95	0.96
(0, 0)	10000	0.10	0.95	0.94	0.95
(0, 0)	20000	0.10	0.94	0.94	0.93
(0, 0)	5000	0.20	0.95	0.95	0.95
(0, 0)	10000	0.20	0.93	0.94	0.93
(0, 0)	20000	0.20	0.94	0.93	0.94

Table 2.11: **Unscaled Simulation Results for $\mathbb{R}^2$ with $x^* = (0, 0)$: Mixture Outcomes** This table displays the coverage of our 95% confidence intervals in $\mathbb{R}^2$ with $x^* = (0, 0)$ and no scaling when using mixture distributions as the outcome.

2.9 Discussion

By focusing on axis-parallel neighborhoods, our proposed method provides 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. Our method performs well across the range of patient outcomes and in the presence of estimated parameters. The large amount of EHR data available allows us to obtain reasonable levels of precision despite the non-parametric nature of our method. 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 predictors.

x^*	n	p	Mix 1	Mix 2	Mix 3
(1, 1)	5000	0.05	0.96	0.96	0.96
(1, 1)	10000	0.05	0.96	0.94	0.94
(1, 1)	20000	0.05	0.94	0.94	0.95
(1, 1)	5000	0.10	0.94	0.95	0.95
(1, 1)	10000	0.10	0.95	0.94	0.95
(1, 1)	20000	0.10	0.94	0.93	0.96
(1, 1)	5000	0.20	0.95	0.95	0.94
(1, 1)	10000	0.20	0.94	0.95	0.94
(1, 1)	20000	0.20	0.93	0.93	0.94

Table 2.12: **Unscaled Simulation Results for $\mathbb{R}^2$ with $x^* = (1, 1)$: Mixture Outcomes** This table displays the coverage of our 95% confidence intervals in $\mathbb{R}^2$ with $x^* = (1, 1)$ and no scaling when using mixture distributions as the outcome.

x^*	n	p	Mix 1	Mix 2	Mix 3
(0, 0)	5000	0.05	0.95	0.96	0.96
(0, 0)	10000	0.05	0.97	0.97	0.96
(0, 0)	20000	0.05	0.95	0.95	0.96
(0, 0)	5000	0.10	0.96	0.96	0.96
(0, 0)	10000	0.10	0.95	0.94	0.95
(0, 0)	20000	0.10	0.95	0.95	0.95
(0, 0)	5000	0.20	0.96	0.96	0.95
(0, 0)	10000	0.20	0.96	0.96	0.96
(0, 0)	20000	0.20	0.95	0.96	0.96

Table 2.13: **Scaled Simulation Results for $\mathbb{R}^2$ with $x^* = (0, 0)$: Mixture Outcomes** This table displays the coverage of our 95% confidence intervals in $\mathbb{R}^2$ with $x^* = (0, 0)$ and global scaling when using mixture distributions as the outcome.

x^*	n	p	Mix 1	Mix 2	Mix 3
(1, 1)	5000	0.05	0.95	0.96	0.96
(1, 1)	10000	0.05	0.97	0.96	0.96
(1, 1)	20000	0.05	0.95	0.95	0.95
(1, 1)	5000	0.10	0.95	0.95	0.94
(1, 1)	10000	0.10	0.95	0.96	0.96
(1, 1)	20000	0.10	0.96	0.95	0.96
(1, 1)	5000	0.20	0.96	0.96	0.96
(1, 1)	10000	0.20	0.97	0.95	0.95
(1, 1)	20000	0.20	0.96	0.97	0.96

Table 2.14: **Scaled Simulation Results for $\mathbb{R}^2$ with $x^* = (1, 1)$: Mixture Outcomes** This table displays the coverage of our 95% confidence intervals in $\mathbb{R}^2$ with $x^* = (1, 1)$ and global scaling when using mixture distributions as the outcome.

Chapter 3

MULTIPLE TREATMENTS

Frequently we not only observe a patient but also make decisions about how best to treat him, and thus we require the ability to compare available treatments. We therefore extend our method by considering a scenario in which patients are subject to multiple possible treatments. For the sake of simplicity, we consider a two treatment scenario throughout this chapter, but the methodology could easily be extended to handle more treatment options. Our method allows the user to make multiple subgroup-specific treatment effect comparisons based on parallel predictions and summarize them in the manner desired, whether that is the proportion of patients who achieve meaningful improvement, the mean difference, or something else entirely. However, we face two kinds of selection bias: that due to choice of neighborhood and that due to choice of treatment. If we select our two neighborhoods poorly, the patients in each neighborhood will be very different, and the assessment of outcomes will be biased. If we face confounding, as is generally the case outside of randomized controlled trials, we can also face bias in which patients are in each treatment group. Our goal is to provide a method and the associated inference for fairly comparing these patient populations.

3.1 Reconciling Multiple Neighborhoods

Our first challenge when considering multiple treatments is to define a target neighborhood that maintains comparability of the patient populations. Because of our choice of the supremum norm as our metric, the smaller of the two neighborhoods is enclosed in the larger provided that we select common scaling parameters from the population as a whole, and thus we need only concern ourselves with neighborhood size rather than any horizontal or vertical translations of the neighborhood. The major concern in this setting is determining the size of the common neighborhood.

We have several options: controlling minimum precision (i.e., letting the final size be when the sparser of the two neighborhoods reaches $n_0 = m$ points so that each of the two estimates has at least the precision requested by the user), controlling overall precision (i.e., letting the final size be when the two neighborhoods reach $n_0 + n_1 = 2m$ points together), or controlling the precision of the difference between groups (i.e., $\frac{1}{m} = \frac{1}{n_0} + \frac{1}{n_1}$). It is impossible to control the number of points in both neighborhoods at once; if this is attempted, the neighborhoods do not cover the same covariate ranges, and we lose comparability. Thus, some difference in the number of points in the two neighborhoods is to be expected. The only question is the most intuitive way to allow the user to specify the size of the neighborhood.

If we allow the user to control the combined number of points, there is no way to ensure that he obtains at least as many points as he would like in the sparser neighborhood, which can contain dramatically fewer points than he imagined. We demonstrate this in example data in which one neighborhood is much more dispersed than the other (see Figure 3.1). Here the user requests 200 total points and is left with one neighborhood of only 26 points. If we allow him to specify the number of points in the sparser neighborhood, then he obtains 100 points there, as he was likely hoping for. The denser neighborhood will of course contain more points – 373 in this case – but we expect users to be more concerned with having difficulty meeting their minimum desired levels of precision than having more points than expected, and hence controlling minimum precision is the preferred option. Controlling the precision of the mean difference creates a similar result as specifying the minimum precision, but it is more difficult to maintain continuity with the selection process in our one treatment framework and less intuitive for a user who is interested in the number of patients in the neighborhood.

3.2 Control of Confounding

In addition to a new neighborhood size criterion, our method requires additional help when we are faced with the challenge of confounding. Estimating the treatment effect in quasi-experimental studies is difficult because patient characteristics can influence treatment assignment and we wish to estimate the treatment effect in the counterfactual situation in which both treated and control

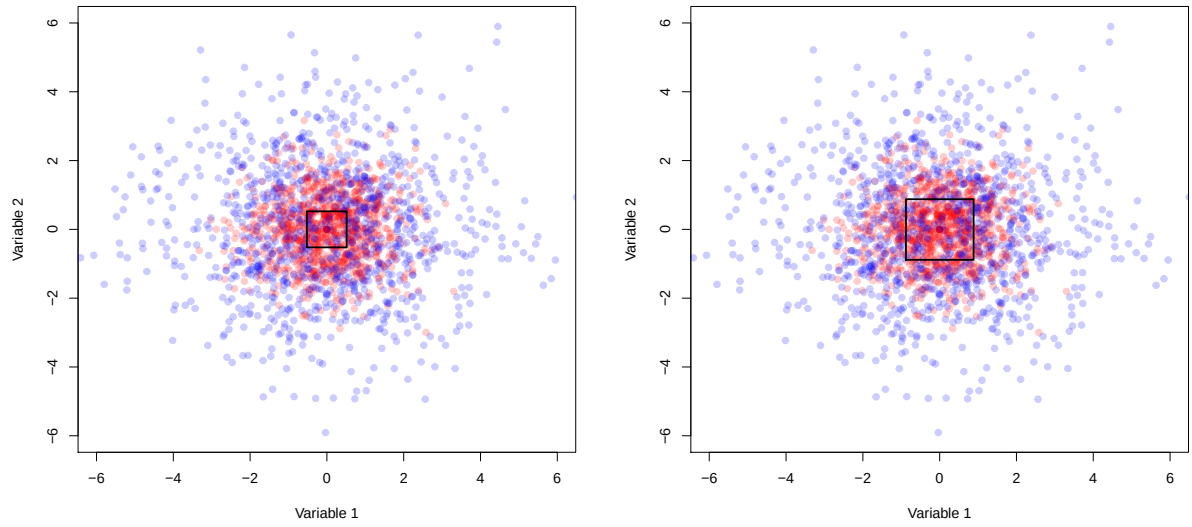

Figure 3.1: **Reconciling Neighborhoods** Here we consider two groups of 1000 subjects, red and blue, where blue subjects have much more variable covariates. When we aim for 200 total subjects, as in the left picture, we obtain only 26 blue subjects in our neighborhood (and 174 red subjects). When we require each neighborhood to have at least 100 subjects, we obtain 100 blue subjects and 373 red subjects.

patients are otherwise comparable. Therefore, we require a method that will correct for selection bias.

Suppose we have for patients $i = 1, \ldots, n$ the set of variables $(\underline{\mathbf{X}}_i, W_i, Y_{0i}, Y_{1i})$, where $\underline{\mathbf{X}}_i$ represents subject i's covariates, W_i whether he receives the control ($W_i = 0$) or treatment ($W_i = 1$), Y_{0i} his outcome under the control, and Y_{1i} his outcome under the treatment. Note that we observe either his outcome under the control or his outcome under the treatment – one of these outcomes is counterfactual. We can express his outcome under the Neyman-Rubin counterfactual framework as

$$Y_i = W_i Y_{1i} + (1 - W_i) Y_{0i}. \tag{3.1}$$

This framework relies on two main assumptions: the ignorable treatment assignment assumption and the stable unit treatment value assumption (SUTVA). The ignorable treatment assignment assumption states that treatment assignment and outcome are independent when we condition on the covariates, i.e.

$$(Y_0, Y_1) \perp W | \underline{\mathbf{X}}. \tag{3.2}$$

SUTVA states that $Y_{W_i i}$ does not depend on either the mechanism used to select W_i or on W_j for any $j = 1, \ldots n$, $j \neq i$. We can then define the average treatment effect as

$$ATE = E(Y_1|W = 1) - E(Y_0|W = 0). \tag{3.3}$$

and the average treatment effect among the treated as

$$ATT = E(Y_1|W = 1) - E(Y_0|W = 1). \tag{3.4}$$

3.2.1 Propensity Score

In quasi-experimental studies, it is crucial that we balance the data (i.e., consider patients in each group with comparable covariates) to obtain a valid, unbiased estimate of the ATE [11]. Balancing

scores – scores such that the conditional distribution of covariates is the same for the treatment and control group given that the balancing score is the same – allow us to make the treatment and control groups comparable. In other words, the covariates are independent of treatment when the balancing scores are equivalent and therefore, given our assumptions, all balancing scores allow us to estimate the average treatment effect. An obvious (and rather useless) balancing score would be the covariates themselves; given that the covariates are identical for the treated and control groups, the covariates are independent of treatment assignment. However, a much coarser and more useful balancing score exists: the propensity score. The propensity score is defined as

$$e(\underline{\mathbf{x}}_i) = P(W_i = 1|\underline{\mathbf{X}}_i = \underline{\mathbf{x}}_i), \tag{3.5}$$

i.e., the probability of being treated given one's covariates, and it is the coarsest possible balancing score. In other words, the propensity score gives us enough information to balance the data and no more; this is particularly useful in high-dimensional situations when the treatment and control groups do not already have comparable covariates and balancing the covariates directly is infeasible [34]. Given our assumptions, the expected difference in observed response to treatment at $e(\underline{\mathbf{x}}_i)$ is equal to the ATE at $e(\underline{\mathbf{x}}_i)$, which links the propensity score to the counterfactual framework [11].

Estimation

In nonrandomized experiments, the propensity score is typically unknown, but it can be estimated, and the estimated scores can be used to produce sample balance. Estimated propensity scores in fact perform better than true propensity scores in practice – while the true propensity score corrects for systematic bias only, the estimated propensity score also corrects for imbalance that results simply from chance [23]. However, it is important to note that the propensity score cannot balance unobserved covariates.

When using propensity scores, we begin by estimating the conditional probability of receiving treatment. This is typically done via logistic regression, though there is no definitive approach to creating a propensity score model. Our goal is to choose the observed covariates that affect selec-

tion bias and to specify their functional form. It is suggested that all covariates thought to be related to the outcome should be included in the propensity score model [4]; more generally, excluding an important confounder is very costly in terms of bias, while including a variable unassociated with the outcome only slightly increases variance, and therefore a liberal approach to variable inclusion is recommended [40]. However, matching is easier and variance smaller if only potential (associated with the outcome) or true (associated with the treatment and the outcome) confounders are included, as opposed to all covariates or all covariates associated with treatment assignment, so some care must be taken [3].

More sophisticated methods for propensity score development have recently been proposed. McCaffrey and colleagues suggest generalized boosted models (GBMs) to overcome the problem of large numbers of covariates with unknown functional forms [30]. Unlike standard logistic regression, GBMs are non-parametric, automated, and data-adaptive. Although selecting all covariates thought to be important can work when the number of covariates is small, this approach will often exhaust the regression's degrees of freedom in higher-dimensional scenarios; it is precisely in these scenarios that boosting has been shown to perform well with respect to prediction error. GBMs add many simple functions (in this case, regression tree models), which lack smoothness and are poor approximations to the function of interest, to obtain a good estimate of a smooth function of many covariates. To model the log-odds of treatment assignment, McCaffrey and colleagues initially set this value to the constant baseline log-odds of treatment for the entire sample and then improve the fit in successive steps, where improvement is measured by increasing the Bernoulli log-likelihood for the (log-odds of the) propensity score model (i.e., maximizing the likelihood). These improvements could in theory take any form, but here they are regression trees that model the residuals of the current model. McCaffrey and colleagues also offer recommendations for the tuning parameters necessary for this algorithm.

Imai and Ratkovic suggest a parametric method they term the covariate balancing propensity score [19]. They take advantage of the propensity score's two definitions: the conditional probability of treatment assignment, which is typically used in propensity score estimation, and the fact that the propensity score is a balancing score. Adding this second trait to their estimation process

minimizes the effect of model misspecification by forcing the moments of the covariates in the treatment and control groups to balance.

Once the propensity score is constructed, it must be assessed to determine whether the model has been adequately specified. This requires comparing the covariate distributions among treated and control subjects, beginning with the means and medians of continuous covariates and the distribution of categorical covariates. This is done via the standardized difference, which compares differences in means in units of pooled standard deviation. This measure is not affected by sample size and allows the relative balance of covariates to be assessed. Typically a standardized difference less than .1 is taken to be negligible, though there is no universal standard. Subsequent examination of higher order moments and covariate interactions among treated and untreated subjects is strongly suggested. Additionally, graphical displays such as box plots and density plots can be used. This is generally an iterative process that is performed until the propensity score model leads to acceptable levels of balance. Testing for the statistical significance of covariate differences between treated and untreated subjects is not recommended. These tests are confounded with sample size and their reference to a superpopulation is inappropriate (balance is a characteristic of a particular sample). Also not recommended is the use of the c-statistic, which studies have shown provides no information about how well the propensity score model has been specified [3].

Use in Practice

Regardless of how our estimated propensity score is obtained, it is typically used in one of four manners: stratification, matching, covariate adjustment, and inverse probability of treatment weighting (IPTW) [2].

In stratification, subjects are ranked according to their propensity score and then split into mutually exclusive subsets based on predefined thresholds. Stratification in practice is most commonly based on propensity score quintiles; this type of stratification reduces roughly 90% of selection bias [2]. An increased number of strata leads to an increase in bias reduction, but there are diminishing returns as more strata are added [3].

Stratification can be thought of as a sort of meta-analysis: we estimate the treatment effect

for each stratum and then pool the estimates to obtain an overall estimate. Weighting based on the number of subjects in each stratum produces an estimate of the ATE, while weighting based on the number of treated subjects in each stratum produces an estimate of the ATT. A pooled variance estimate can likewise be obtained by pooling the individual variance estimates [3]. As with matching, stratification can be combined with additional regression adjustment to account for residual differences within a particular stratum.

Matching on the propensity score allows for estimation of the ATT, and treatment effects can be reported using the same metrics as in controlled trials. It is important to account for matching when estimating the variance because matched subjects share an inherent similarity to their partners; simulation studies have shown that this adjustment leads to more accurate results. Examples of such methods are the paired t-test for continuous outcomes and McNemar's test for dichotomous outcomes. In addition, matching requires several further decisions to be made: whether to match with or without replacement, whether to use greedy or optimal matching, whether to specify a caliper width to cap the distance between matches, and how many untreated subjects to match to each treated subject [3].

Matching with and without replacement is self-explanatory, though it is worth noting that matching with replacement requires additional adjustments to the variance estimate to account for untreated subjects appearing in multiple matched sets [3]. Greedy matching takes each treated subject in turn and assigns him the best untreated match available, regardless of whether that untreated subject might be a better match for a subsequent treated subject. Optimal matching aims to minimize the total within-pair difference of the PS. Greedy matching is generally sufficient in practice [3]. In many cases one will wish to forbid unsuitable matches even if they are the best matches available; in other words, it is better to let some treated subjects go without matches rather than to give them poor matches. This is done by imposing a caliper that specifies the maximum distance between acceptable matches. In practice, a wide range of caliper widths has been used, though theoretical studies tend to suggest .2 standard deviations of the logit of the propensity score as a reasonable choice [3].

Most often one untreated subject is matched to each treated subject, though if additional un-

treated subjects are available, matched sets may be larger. When matching multiple untreated subjects to one treated subject, letting the number of untreated subjects vary in matched sets provides better bias reduction [3]. The most common type of matching in practice is nearest neighbors without replacement within a specified caliper of propensity score [2]. In addition, propensity score matching can be combined with additional matching on prognostic factors or regression adjustment.

Covariate adjustment, in which the outcome variable is regressed on treatment status and propensity score, is the most commonly used method of accounting for the propensity score [2]. It requires the additional assumption that the outcome model is correctly specified, though there is some evidence to suggest that if one uses robust variance estimators and a correctly specified propensity score, one is relatively safe from biased hypothesis tests caused by misspecification of the outcome model when using canonical GLMs [45]. Note also that this is the only one of the four methods that does not separate design and analysis of the study [3].

IPTW, a type of marginal structural model, is an indirect method of regression adjustment that can help handle sparse data. It re-weights the subjects to break the relationship between covariates and treatment assignment; in other words, it creates a sample in which baseline covariates and treatment are independent, similar to the way in which survey sampling weights can be used to create a more representative sample of a population. The weights can become quite unstable when probability of treatment is very low, and the use of stabilizing weights has been proposed to handle these situations [3].

Multiple studies have suggested that matching provides better bias reduction than stratification or covariate adjustment, while IPTW performs similarly to matching [3]. In other words, matching and IPTW most effectively reduce systematic differences between treatment groups [2]. However, covariate adjustment and IPTW may be more sensitive to accurate modeling of the propensity score than matching and stratification [3].

Use of the propensity score provides a number of advantages over direct regression adjustment for baseline covariates. Propensity score methods allow for estimation of the marginal effect of treatment, they allow increased flexibility when modeling rare outcomes, and they separate design

and analysis. With propensity score it is also simpler to determine whether our model is correctly specified and the comparability of treated and control subjects [3].

Combining Propensity and Prognostic Scores

While the propensity score collapses the covariates into a measure summarizing their relationship with the treatment, a prognostic score collapses the covariates into a measure summarizing their relationship with outcome. The outcome under a given treatment and the covariates are independent given a fixed value of prognostic score, or, where ψ is our prognostic score,

$$Y_0 \perp \underline{\mathbf{X}}|\psi, \tag{3.6}$$

whereas for the propensity score,

$$(Y_0, Y_1) \perp W|e. \tag{3.7}$$

Like the propensity score, the prognostic score is a balancing score, but rather than minimizing covariate differences with respect to treatment assignment, the prognostic score attempts to minimize covariate differences with respect to potential outcome under a given treatment, and hence tends to balance covariates that are strongly associated with the outcome [12]. It is important to note that prognostic score modeling requires the outcome for only a single treatment group and hence there is little potential for unconscious investigator bias in favor of a treatment effect. Studies suggest that joint use of the propensity score and the prognostic score provides superior bias reduction to using either alone [27] [36].

3.3 Practical Considerations

We control confounding via a modified design-based use of global propensity score. This allows us to reduce confounders to a single axis of our neighborhood, which, given that we require low-dimensional covariates, is crucial. Alternatively, one could place a caliper on propensity score and restrict it to values within a certain distance of the patient of interest's score. However, we

believe our semi-supervised approach is superior in practice because it allows the data to determine the weight given to the propensity score based on the severity of confounding. Additionally, the weight adaptively changes the range of propensity scores allowed in the patient's neighborhood based on the number of neighbors requested. Creating a new caliper by hand for each change in neighborhood size would be difficult, and creating only one caliper for all neighborhood sizes would be either unnecessarily liberal or restrictive. Additionally, the creation of any caliper would require either a standardized cutoff that does not take advantage of information in the data about the severity of confounding or it would require additional time and research on the part of the user.

We scale based on the pooled data in order to maintain the comparability of neighborhoods. Likewise, we force Δ for the control and treatment neighborhoods to be the same. We do so by controlling minimum precision; i.e., the sparser neighborhood will contain the requested number of points, while the denser neighborhood will contain as many points as are within Δ calculated from the sparser neighborhood. We select this approach because our main concern when controlling precision is giving users an estimate that is based on a sufficient number of patients. If we chose to control the number of subjects in the denser neighborhood, or overall, the sparser neighborhood would quite possibly be too small for the user's comfort.

3.4 New Algorithm

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 $\underline{\mathbf{X}}[, j]$ on Y, let the slope coefficient be $\hat{\beta}[j]$, and A be the diagonal matrix created by $\hat{\beta}$ (outcome-based rescaling). Let $G(\underline{\mathbf{x}})$ be the propensity score, with the additional elements of this vector filled with any desired function of the covariates, such as identity or percentile. For $1 \leq i \leq n$, calculate $d_i = ||A[G(\underline{\mathbf{x}}_i) - G(\underline{\mathbf{x}}^*)]||_\infty$.

3. Where F_d is the distribution of distances, calculate $q_{d0} = F^{-1}_{tx=0,d}(p)$ and $q_{d1} = F^{-1}_{tx=1,d}(p)$. Let $q_d = \max(q_{d0}, q_{d1})$. Discard all $\underline{\mathbf{x}}_i$ such that $d_i > q_d$; call the remaining points $N(\Delta, \underline{\mathbf{x}}^*)$.

4. Find the enclosure of $N(\Delta, \underline{\mathbf{x}}^*)$. For $1 \leq j \leq k$, let $V[j,1] = \min(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$ and $V[j,2] = \max(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$.

3.5 Mathematical Treatment of Score Uncertainty

We begin with our previous asymptotic result,

$$\sqrt{m}\left(\hat{F}[y|\underline{\mathbf{X}} \in N(\hat{\Delta}_n, \underline{\mathbf{x}}^*)] - \bar{F}[y|\underline{\mathbf{X}} \in N(\Delta, \underline{\mathbf{x}}^*)]\right) \tag{3.8}$$

$$\xrightarrow{d} \mathbb{X} + N\left(0, \frac{p^2(1-p)}{f_d^2(\Delta)} \times \left\{\frac{d}{d\Delta}\bar{F}[y|x \in N(x^*, \Delta)]\right\}^2\right). \tag{3.9}$$

As the propensity score is a function of the covariates and the estimated propensity score is therefore an estimated function of the covariates, we now have both $\hat{\beta}$ and $\hat{G}$ to contend with, and we wish to consider

$$\hat{F}[y|\hat{G}(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\hat{\beta}), \underline{\mathbf{x}}^*)] - \bar{F}[y|G(\underline{\mathbf{X}}) \in N(\Delta(\beta), \underline{\mathbf{x}}^*)], \tag{3.10}$$

which can be rewritten as

$$\hat{F}[y|\hat{G}(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\hat{\beta}), \underline{\mathbf{x}}^*)] - \hat{F}[y|\hat{G}(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\beta), \underline{\mathbf{x}}^*)] \tag{3.11}$$

$$+\hat{F}[y|\hat{G}(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\beta), \underline{\mathbf{x}}^*)] - \hat{F}[y|G(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\beta), \underline{\mathbf{x}}^*)] \tag{3.12}$$

$$+\hat{F}[y|G(\underline{\mathbf{X}}) \in N(\hat{\Delta}_n(\beta), \underline{\mathbf{x}}^*)] - \bar{F}[y|G(\underline{\mathbf{X}}) \in N(\Delta(\beta), \underline{\mathbf{x}}^*)]. \tag{3.13}$$

We then have one term in which β is moving from its estimated to true value and one term in which $G(\cdot)$ is moving from its estimated to its true value. The estimation of these functions introduces additional uncertainty, uncorrelated with the previous terms, that can be evaluated using results from van der Vaart and Wellner, provided that the behavior of the distribution is not pathological [43]. We obtain

$$\sqrt{m}\sqrt{p}\left[\bar{F}(y \mid ||\hat{\beta}[\hat{G}(\underline{\mathbf{x}}) - \hat{G}(\underline{\mathbf{x}}^*)]||_\infty \leq \hat{\Delta}_n) - \bar{F}(y \mid ||\beta[\hat{G}(\underline{\mathbf{x}}) - \hat{G}(\underline{\mathbf{x}}^*)]||_\infty \leq \hat{\Delta}_n)\right] \tag{3.14}$$

and

$$\sqrt{m}\sqrt{p}\left[\bar{F}(y \mid ||\beta[\hat{G}(\underline{\mathbf{x}}) - \hat{G}(\underline{\mathbf{x}}^*)]||_\infty \leq \hat{\Delta}_n) - \bar{F}(y \mid ||\beta[G(\underline{\mathbf{x}}) - G(\underline{\mathbf{x}}^*)]||_\infty \leq \hat{\Delta}_n)\right]. \quad (3.15)$$

Again we focus only on the order of these terms, as their exact value will vary depending on the method of propensity score and scaling estimation. In the fixed percentage scenario, the additional uncertainty is of the same order as the Gaussian process and its exact value under any given method of estimation can be determined using the functional δ-method. This method can be used so long as the functional we use is Hadamard differentiable at our distribution.

3.6 Illustration

In this section we use simulated data to demonstrate the value of the propensity score for our method. We first consider the effect of using the propensity score rather than a single confounder as one of our covariates. Because our method is suitable to low-dimensional data only, using all confounders separately is not practical; this would in many cases consume all available dimensions and then some, while preventing us from including other important covariates. Thus, we consider only methods of controlling confounding that result in a one-dimensional score or covariate, and we demonstrate that the propensity score is superior given this constraint. We then examine the effect of the propensity score's predictive strength on the neighborhood shape, which demonstrates the adaptive scaling discussed previously.

Where Y is our outcome, Z is our treatment, $\underline{\mathbf{X}}$ are our k confounders, P is a precision variable, and ε is our error term, suppose we have

$$Z \sim \text{binomial}(n, \text{expit}[\gamma\underline{\mathbf{X}}]) \quad (3.16)$$

$$Y \sim \beta\underline{\mathbf{X}} + P + Z + \varepsilon. \quad (3.17)$$

In essence, we have a set of confounders, $\underline{\mathbf{X}}$, that impact Y and Z to varying degrees, and we must account for them when comparing treated and untreated subjects. Note that the treatment effect is uniformly 1. In our example we let

$$\underline{\mathbf{X}}_j \sim N(0,1) \quad \forall\, 1 \leq j \leq k \tag{3.18}$$

$$P \sim N(0,1) \tag{3.19}$$

$$\varepsilon \sim N(0,1). \tag{3.20}$$

We assign β as $\frac{b}{k}(1,\ldots,k)$ and γ as $\frac{g}{k}(k,\ldots,1)$ for scalars b, g. We consider 10,000 subjects, $k = 5$ confounders, and we let $b = g = 1$. We examine a roughly average patient, one who has an estimated propensity score of .5, a precision variable value of 0, and a first confounder value of 0. We create two neighborhoods of 100 patients: one based on the precision variable and the propensity score, the other based on the precision variable and the first confounder. This allows us to compare the efficacy of the propensity score and the most important confounder in reducing confounding (see Figure 3.2 for a visual examination of the difference). We obtain a mean difference between treated and untreated group of 1.23 when we use the precision variable and the propensity score as opposed to 2.25 when we use the precision variable and one confounder; recall that the true mean difference between groups is 1. Thus, the propensity score does a dramatically better job of controlling confounding than a single confounder despite being of the same dimension.

We then reduce g to .25 while holding all other parameters fixed and again create a neighborhood of 100 patients around a patient with a propensity score of .5 and a precision variable of 0. Here we obtain a mean difference of 1.17; again, the propensity score reduces confounding (see Figure 3.3). Note that the neighborhood is wider on propensity score (i.e., propensity score is weighted less relative to the other covariate) when propensity score is less predictive, an example of our adaptive scaling. Our method automatically places a tighter limit on a strongly predictive propensity score than on one that is weakly predictive.

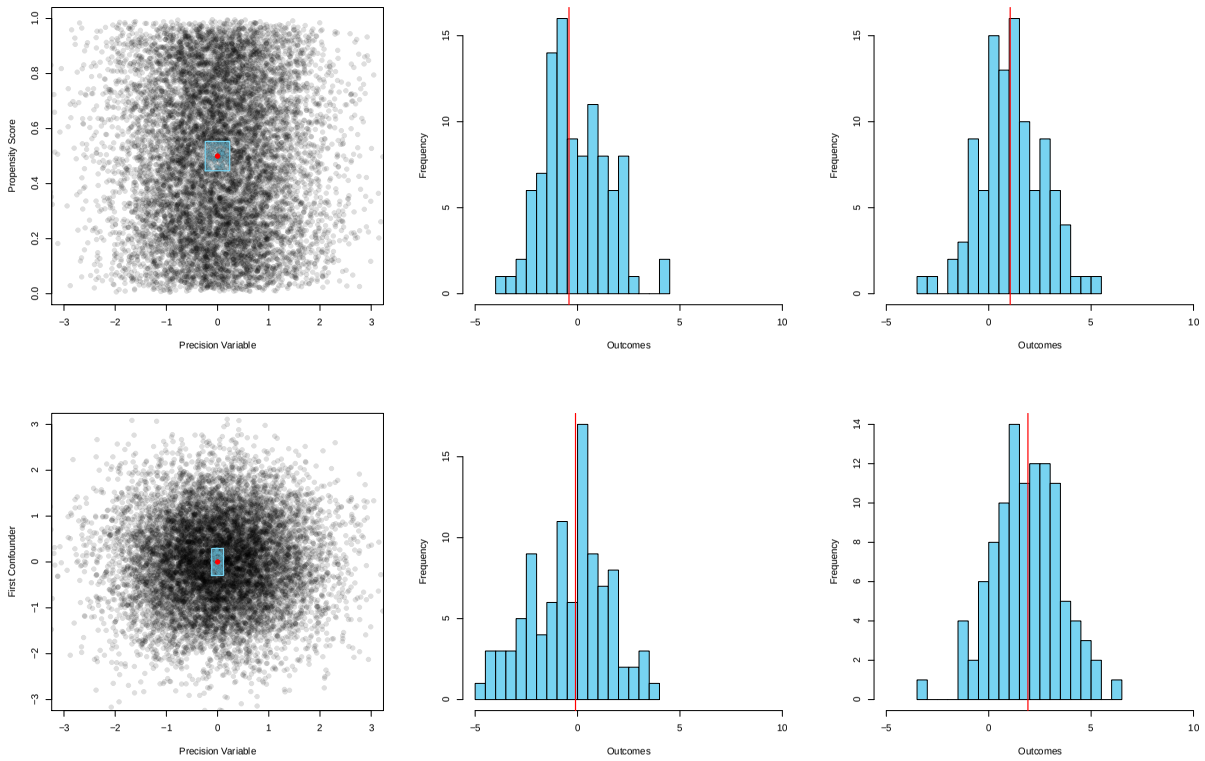

Figure 3.2: **Propensity Score Illustration: Propensity Score vs Confounder** We consider a patient with an estimated propensity score of .5, a first confounder value of 0, and a precision variable value of 0. We examine the effects of using the most predictive confounder (bottom) rather than the propensity score (top). Both examples display the covariate neighborhoods on the left, the outcomes with no treatment in the center, and the outcomes with treatment on the right. Red lines indicate the median outcome value in a given neighborhood. Observe that the neighborhood gives much less weight to the confounder than to the propensity score, as the confounder is much less predictive of the outcome.

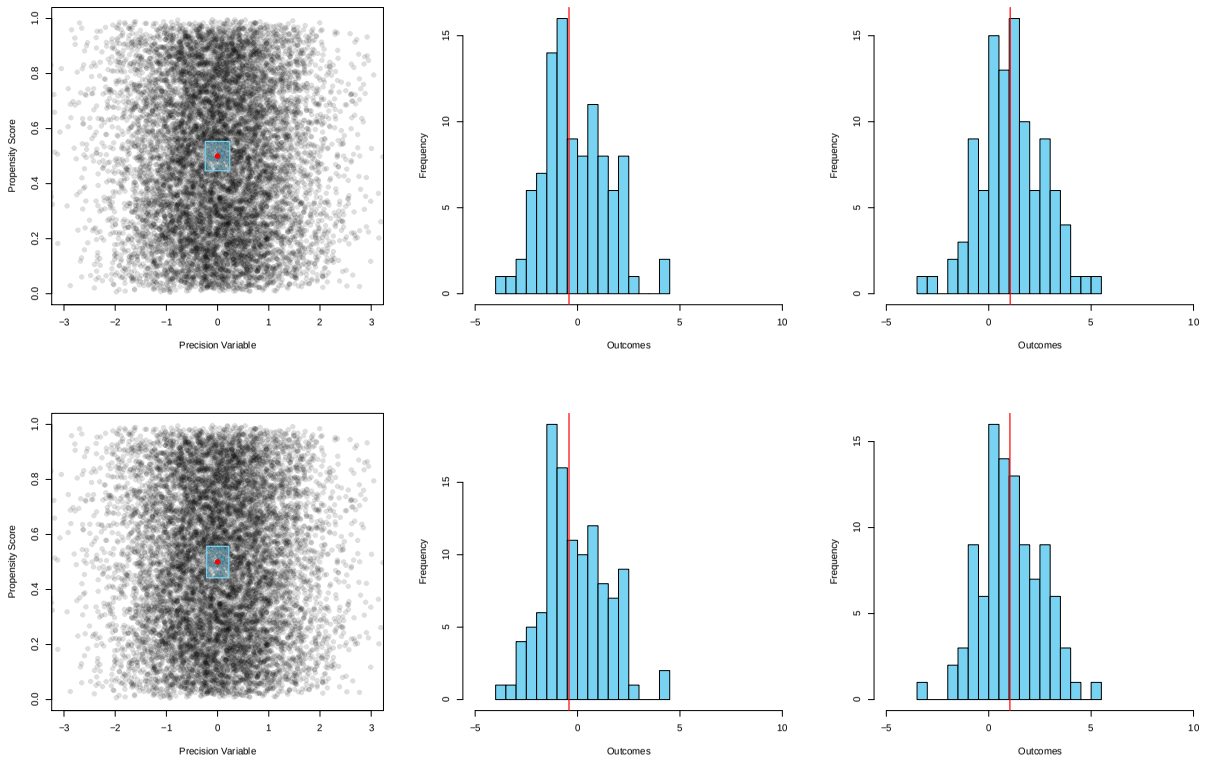

Figure 3.3: **Propensity Score Illustration: Strength of Propensity Score** We consider a patient with an estimated propensity score of .5 and and a precision variable value of 0. We examine the impact of a less predictive propensity score (top, strongly predictive propensity score with $g = 1$; bottom, weakly predictive propensity score with $g = .25$). Both examples display the covariate neighborhoods on the left, the outcomes with no treatment in the center, and the outcomes with treatment on the right. Red lines indicate the median outcome value in a given neighborhood. Observe that the strongly predictive propensity score is more tightly controlled than the weakly predictive propensity score.

3.7 Example

We again turn to the BOLD data set, and here we consider patients who either did or did not receive lumbar spine imaging within six weeks of their initial visit as our treated and control groups respectively [21]. If diagnostic imagining helps patients achieve a better outcome, it is important that they receive it as soon as possible; on the other hand, if diagnostic imaging provides no benefit, then there is no reason for patients to endure an unnecessary procedure and providers to endure an unnecessary expense. Most studies have focused on younger adults and, due to the lower prevalence of serious underlying conditions in this demographic, are not necessarily applicable to older adults. We first create a propensity score to balance the baseline covariates in the two groups and then display our results.

3.7.1 Construction of Propensity Score

In order to render subjects who did and did not receive lumbar spine imaging comparable, we constructed a propensity score including the following variables: age, sex, race, baseline disability, baseline back pain, and baseline leg pain. Baseline disability and baseline back pain were modeled with cubic splines, while the remaining continuous variables were modeled linearly. Note that baseline disability and pain levels initially differ substantially between the two groups (see Table 3.1); the differences in groups become negligible once we apply the propensity score (see Tables 3.2 and 3.3).

3.7.2 Visualization

We again used the package Shiny in R to implement our method in two dimensions [31] [35]. We include a histogram to estimate the empirical distribution in each group, as well as the quartiles and mean of the empirical distribution, both overall and for the neighborhood of interest for each group. We allow the user to determine whether he would like to view all subjects or only subjects in a particular group when examining the neighborhood.

As an example, we selected two 75-year-old patients, one with a propensity score of .35, the

		No Image ($n = 3432$)	Image ($n = 1730$)	Pooled ($n = 5162$)
		Mean	Mean	SD
Female		0.24	0.12	–
Race				
	Black	0.16	0.15	–
	Native American	0.00	0.00	–
	Native Islander	0.00	0.00	–
	Asian	0.04	0.04	–
	White	0.74	0.75	–
	Other	0.04	0.04	–
	Mixed	0.02	0.01	–
Age		73.58	74.18	6.86
Disability		8.74	11.15	6.43
Back Pain		4.74	5.61	2.79
Leg Pain		3.18	3.98	3.31

Table 3.1: **Baseline Patient Characteristics** This table displays the differences in baseline characteristics between subjects who did and did not receive lumbar spine imaging.

	Age	Disability	Back Pain	Leg Pain
Quintile 1	-0.05	-0.09	-0.12	-0.08
Quintile 2	-0.05	-0.08	0.03	-0.06
Quintile 3	0.00	0.04	-0.00	0.03
Quintile 4	0.01	-0.01	0.01	-0.02
Quintile 5	0.03	-0.06	-0.08	-0.01

Table 3.2: **Standardized Differences of Covariates by Propensity Score Quintile** We display the standardized differences between groups for each variable as split by quintiles of propensity score. All covariates appear balanced.

	Female	Black	N. American	N. Islander	Asian	White	Other	Mixed
Quintile 1 - NI	0.42	0.10	0.01	0.01	0.03	0.80	0.03	0.02
Quintile 1 - NI	0.40	0.18	0.02	0.02	0.04	0.69	0.04	0.03
Quintile 2 - NI	0.35	0.24	0.01	0.01	0.04	0.64	0.06	0.02
Quintile 2 - I	0.31	0.23	0.01	0.00	0.03	0.62	0.07	0.04
Quintile 3 - NI	0.33	0.16	0.00	0.00	0.04	0.73	0.05	0.02
Quintile 3 - I	0.35	0.14	0.00	0.00	0.03	0.80	0.02	0.01
Quintile 4 - NI	0.35	0.14	0.00	0.00	0.05	0.76	0.04	0.01
Quintile 4 - I	0.34	0.13	0.00	0.00	0.05	0.77	0.04	0.01
Quintile 5 - NI	0.31	0.13	0.00	0.00	0.04	0.78	0.05	0.01
Quintile 5 - I	0.35	0.11	0.00	0.00	0.05	0.78	0.06	0.00

Table 3.3: **Patient Sex and Race by Propensity Score Quintile** We display the proportion of subjects who are female and who belong to each race by quintiles of propensity score. NI are subjects who received no image while I are subjects who received an image. Balance appears to be quite good overall.

median among our subjects, and one with a higher propensity score of .40. We scaled globally and obtained parameters of .12 for age and 32.64 for propensity score. Note that age is not only more weakly associated with outcome than propensity score but also has a much larger range, hence the discrepancy in scaling parameters. We see that the patient with a higher chance of being treated has a slightly worse expected outcome, which is not surprising – doctors are more likely to intervene when patients are not doing well (Figure 3.4). We also see that patients who receive imaging do not seem to have dramatically better outcomes in either case; the shape of the outcome distributions and the value of any particular quantile are similar. Our method easily allows us to examine differences in the two groups both across the range of outcomes and across the range of data.

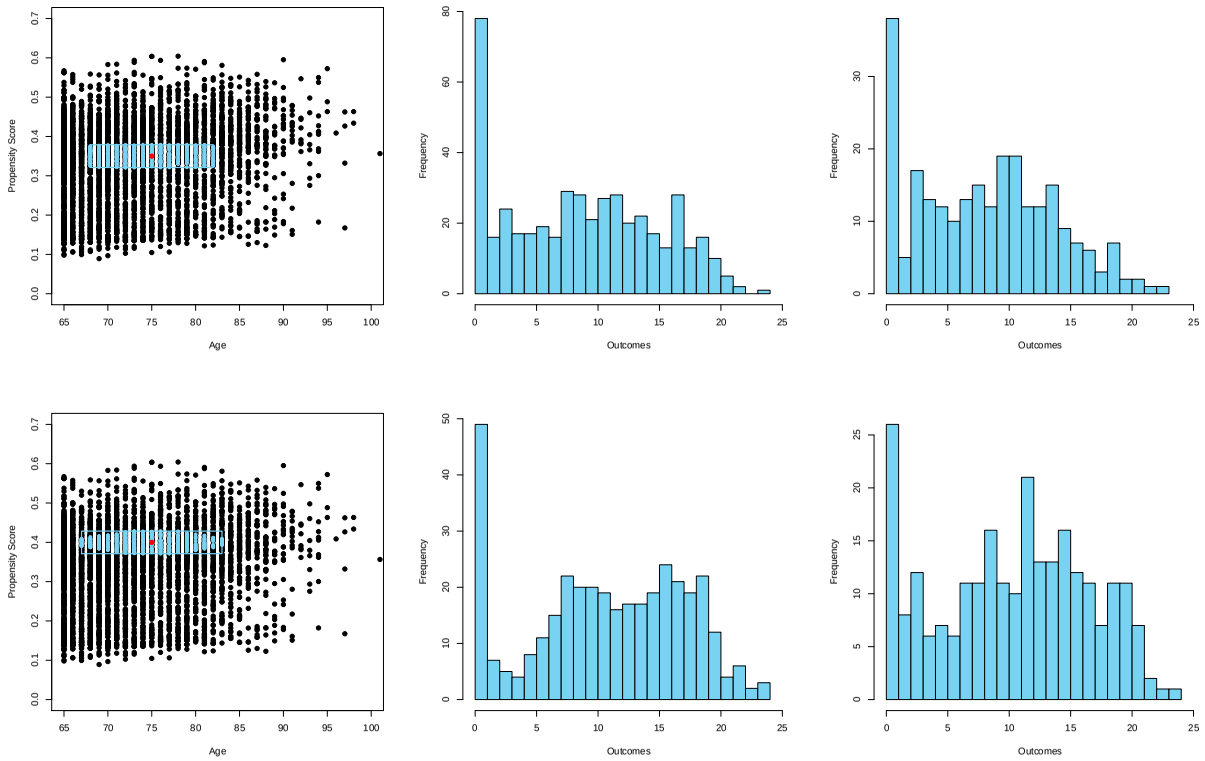

Figure 3.4: **Propensity Score Example in BOLD Data** We consider two 75-year-old patients, with their covariate neighborhoods on the left, their outcomes with no imaging in the center, and their outcomes with imaging on the right. The neighborhood for the patient with a propensity score of .35 extends from age 68 to 82 and propensity scores .32 to .38. There are 467 untreated and 250 treated patients in the neighborhood. The neighborhood for the patient with a propensity score of .40 extends from age 67 to 83 and propensity scores .37 to .43. There are 362 untreated and 250 treated patients in the neighborhood.

3.8 Discussion

We expand our method to handle multiple treatments by creating a single neighborhood based on the sparsest treatment group. By using the propensity score as an axis in our neighborhood, we allow the comparison of treatment groups even in the presence of selection bias. Our method automatically weights the propensity score based on its predictive value, sparing the user from creating a caliper for each situation. In the future, we hope to examine a local propensity score, much in the same way we consider local scaling. This would allow for finer control over relevant neighbors while discarding less useful information from more distant patients.

Chapter 4

LONGITUDINAL DATA AND LOCALIZED PREDICTION

Follow-up of patients at multiple time points is common and desirable when the goal is to measure the impact of a change in variables over time or to assess the risk of events at varying time points in the future. Because longitudinal data gives us a dynamic assessment of patient prognosis, we can better create an individualized treatment plan [44]. However, longitudinal data typically presents three major challenges: it is high-dimensional, it is heterogeneous, and it is strongly prognostic. The high dimension of histories is in our case a particular problem, as our method is by nature suited to low-dimensional data. It is crucial that we reduce the dimension of these histories while maintaining the most important information contained within them. Heterogeneity means that the extent of history for each patient is different – some patients visit frequently, while others have more infrequent visits or have been seen less recently. We need a way to account for the extent of the history when summarizing it. Finally, we need to extract as much information as possible from the histories, as they generally provide quite valuable prognostic information.

4.1 History Features as Prognostic Variables

We have several options when attempting to extract useful information from patient histories. To make these options concrete, suppose we have two patients, one of whom has had his disability level evaluated ten times after his baseline visit and the other of whom has been evaluated five times after his baseline visit (see Figure 4.1). Although these patients both have a baseline disability of five, their histories diverge dramatically from there. The first patient appears to improve over the the first few visits but ends up worse than when he started. The second patient also has a downward trajectory and then makes no more visits. Additionally, he misses the fourth visit. How do we summarize the histories of these patients?

Our first option is a user-defined summary. This can be as simple as the last observation, or the mean of the observations, or perhaps the linear trend of the observations, and all of these can return different results, as you can see with our example patients in Figure 4.2. The last observation and mean observation provide a simple way of reducing the dimension of history, but they tell us nothing about the trajectory, and the last observation in particular is quite unstable. While the linear model tells us about the trajectory, it (along with the previous two methods) does not help with the difference in number of visits – none of the summaries accounts for us having twice as much information on the first patient as the second, and for the inherent instability in the trajectories of patients for whom we have less information. Empirical Bayes methods, which we describe in detail in the next section, allow us to account for the heterogeneity in visit timing and frequency by pulling patients about whom we know less closer to the overall average in the data.

Although Empirical Bayes allows us to account for heterogeneity, it does not tell us which measurements to include – we have the option of a local (recent) or global (complete) history. Supposing we wanted to only consider the patient's most recent five visits, our first patient's trajectory would become much steeper (see Figure 4.3). How much of the history to use is not only a statistical question but also a scientific one that must be answered on a case by case basis. In some cases years of history may be useful, while in others measurements only months old may be irrelevant. Heagerty and Comstock suggest distributed lag regression methods that can be used to derive predictive summaries of an individual's history, although the dimension of potential models will grow exponentially as the dimension of the history increases and it is difficult to apply when histories are of variable length [15].

4.1.1 Empirical Bayes Methods

Generalized linear mixed models include normally distributed random effects, which allow us to account for the dependence of within-subject measurements, in addition to the fixed effects found in generalized linear models. These random effects are the best linear unbiased predictors of patient-specific effects regardless of whether distributional assumptions are met [10]. Because the random effects are random variables, using Bayesian methods to estimate them is quite natural.

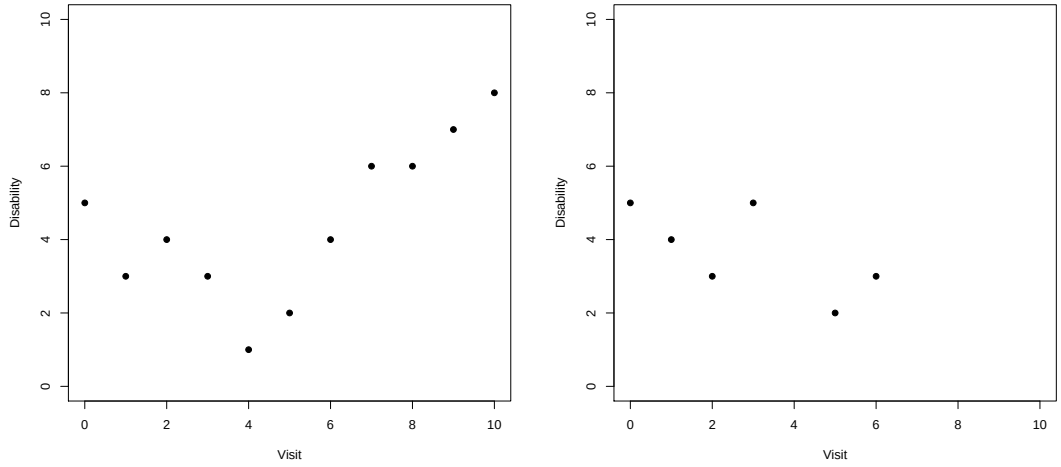

Figure 4.1: **Example Patient Histories** We have two patients, one of whom has been evaluated ten times and one only five.

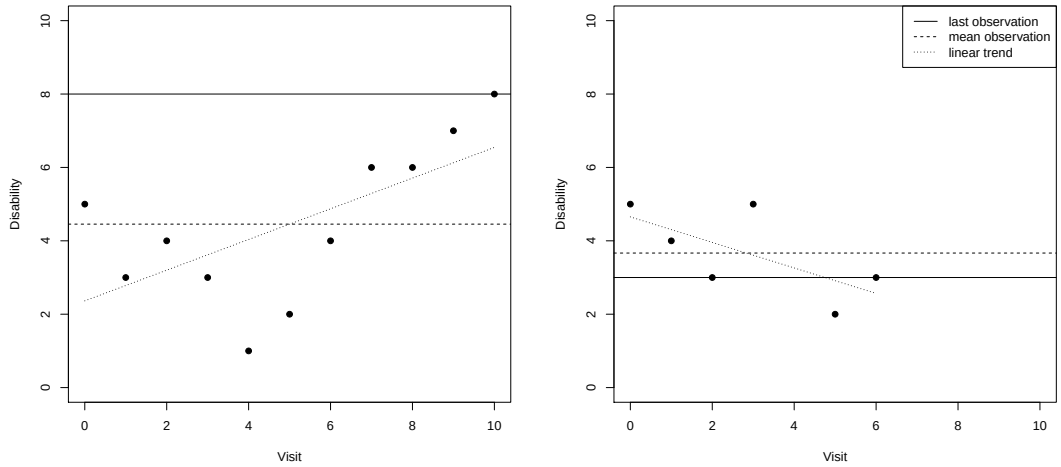

Figure 4.2: **User Defined Summaries of History** The patient histories have been summarized in three ways: by their last observation, by their mean observation, and by their linear trend.

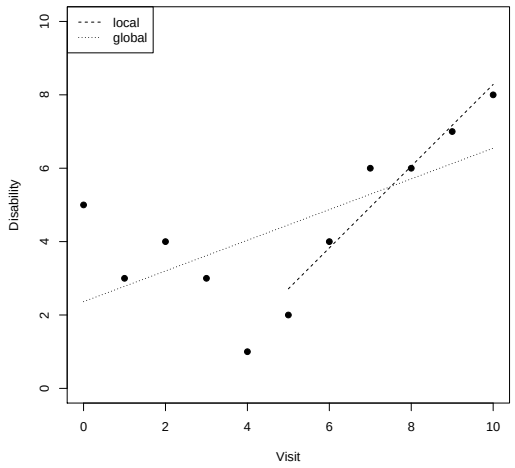

Figure 4.3: **Local versus Global History** We consider the patient's trajectory both overall and over his last five visits.

Supposing we use the maximum likelihood estimates for the unknown parameters when we estimate the random effects, our estimate of the random effects is an Empirical Bayes estimate [46]. We may then select the subject-specific parameter (i.e., random effect) of our choosing to summarize a given patient. In other words, Empirical Bayes methods allow us to pre-specify and estimate a chosen feature of each patient's history to use as a low-dimensional summary of that patient's past. The patient's individual summaries are shrunk to the overall data, resulting in stable estimates that take advantage of all of the data available to estimate our parameter of interest. Thus, Empirical Bayes methods provide a simple, effective way to estimate subject-specific parameters.

Suppose for the sake of simplicity we consider a linear mixed-effects model,

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{4.1}$$

with Y_i the response vector for subject $i \in 1, \ldots, n$, X_i and Z_i matrices of known covariates, β a p-dimensional vector of fixed effects, b_i a q-dimensional vector of random effects, and ε_i the vector

of residuals, that satisfies

$$b_i \sim N(0, D) \tag{4.2}$$

$$\varepsilon_i \sim N(0, \Sigma_i) \tag{4.3}$$

with $b_1, \ldots, b_n, \varepsilon_1, \ldots, \varepsilon_n$ independent and Σ_i dependent on i only through its dimension. We estimate b_i (note that only for linear mixed models does a closed-form expression exist) as

$$\hat{b}_i = E[b_i | Y_i = y_i] \tag{4.4}$$

$$= DZ_i^T V_i^{-1}(y_i - X_i\beta), \tag{4.5}$$

noting that $\hat{b}_i$ is inversely proportional to V_i. It has covariance matrix

$$\text{var}(\hat{b}_i) = DZ_i^T \left\{ V_i^{-1} - V_i^{-1} X_i \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} X_i^T V_i^{-1} \right\} Z_i D. \tag{4.6}$$

Although functional principal component methods are slightly more flexible in that they could be used to relax the parametric mean and to empirically identify functions of the history that appear most strongly predictive of the outcome of interest, they are much more difficult to interpret in a clinical setting, and thus we do not suggest their use for our purposes [17].

4.2 Algorithm for Dimension Reduction

Our algorithm is easily modified to summarize patient histories before finding the axis-parallel neighborhood of m points about $\underline{\mathbf{x}}^*$. As before, the user inputs the point of interest and the number of points desired in the neighborhood, but the user also selects the desired feature of history to be summarized. Then the distance between the point of interest and all other points in the neighborhood is calculated and a box containing the desired number of closest points is created.

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 $\underline{\mathbf{X}}[,j]$ on Y, let the slope coefficient be $\hat{\beta}[j]$, and A be the diagonal matrix created by $\hat{\beta}$ (outcome-based rescaling). Let $G(\underline{\mathbf{x}})$ be the random effects from the chosen model, with the additional elements of this vector filled with any desired function of the covariates, such as identity or percentile. For $1 \leq i \leq n$, calculate $d_i = ||A[G(\underline{\mathbf{x}}_i) - G(\underline{\mathbf{x}}^*)]||_\infty$.

3. Where F_d is the distribution of distances, calculate $q_d = F_d^{-1}(p)$. Discard all $\underline{\mathbf{x}}_i$ such that $d_i > q_d$; call the remaining points $N(\Delta, \underline{\mathbf{x}}^*)$.

4. Find the enclosure of $N(\Delta, \underline{\mathbf{x}}^*)$. For $1 \leq j \leq k$, let $V[j,1] = \min(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$ and $V[j,2] = \max(\{\underline{\mathbf{x}}[j] | \underline{\mathbf{x}} \in N(\Delta, \underline{\mathbf{x}}^*)\})$.

4.3 Mathematical Treatment of Score Uncertainty

As our score is again simply an estimated function of the covariates, the treatment of uncertainty due to estimation of patient trajectories is handled in the same way as uncertainty due to estimation of the propensity score (see Section 3.5). In this scenario $\hat{G}(X)$ is as defined in Equation 4.4 and G represents the true individual summary, while other terms retain their previous values. The exact value of the additional uncertainty will vary depending on the history features summarized and can be determined using the functional δ-method, though this would generally be quite complex in practice and obtaining a tractable expression would be difficult.

4.4 Illustration

We use simulated data to evaluate the performance of Empirical Bayes methods and user-defined summaries. We assign each individual patient a random intercept and slope, distributed normally with zero mean and variance 1 for intercept and $\frac{1}{10}$ for slope. Each patient's covariates follow his line with an error term $\varepsilon \sim N(0,1)$. His outcomes are his covariate value with an additional error term of the same magnitude added. We let half of our 10,000 patients have 5 visits and half of our patients have 10 visits prior to their outcome being assessed; these patients are otherwise

identical. We then compare the neighborhood that a patient (here with a slope of 1 and intercept of .1) obtains in three situations: with his true intercept and slope, with his intercept and slope estimated via Empirical Bayes, and with his intercept and slope estimated via a separate linear model for each patient (see Figure 4.5).

It is striking to note that roughly 70% of the patients in the neighborhood created by the linear model are from the half of patients with fewer visits, as opposed to roughly 60% in the Empirical Bayes neighborhood and 50% in the true neighborhood (see Figure 4.4). The patients with fewer visits have more variable individual summaries and thus are more likely to be captured in the neighborhood of a patient far from the center of the data by random error alone. Empirical Bayes estimates, which are shrunk based on the amount of information we have on a given patient, are not subject to this problem, and return similar results to the true values. This means that Empirical Bayes not only more accurately captures neighbors but also more accurately estimates the outcome. The mean in the Empirical Bayes neighborhood is .97, which is much closer to the true neighborhood mean of .91 than is the individual summary estimate of .68. Although in our case patients seen frequently and infrequently are identical by design, this poses a serious concern in situations where patients who are seen frequently differ from patients who are seen infrequently, as is often true in practice. This can be seen if we consider a scenario identical to our previous scenario but for patients who are seen more often having a boost in their final outcome. In this situation we obtain a mean outcome for our patient with a slope of 1 and an intercept of .1 of 1.60 in the box created with the true individual summaries, 1.58 with the Empirical Bayes individual summaries, and 1.40 with the linear model summaries. The difference between the linear model summaries and the other two summaries is likely due to the fact that again patients with fewer visits are oversampled.

4.5 Example

We return to the BOLD data and this time consider the Roland score at times 0, 1, and 2 as the history for the Roland score at time 3 in a linear mixed model with intercept and visit as fixed effects and a random intercept and slope for each subject. We thus obtain two sets of individual

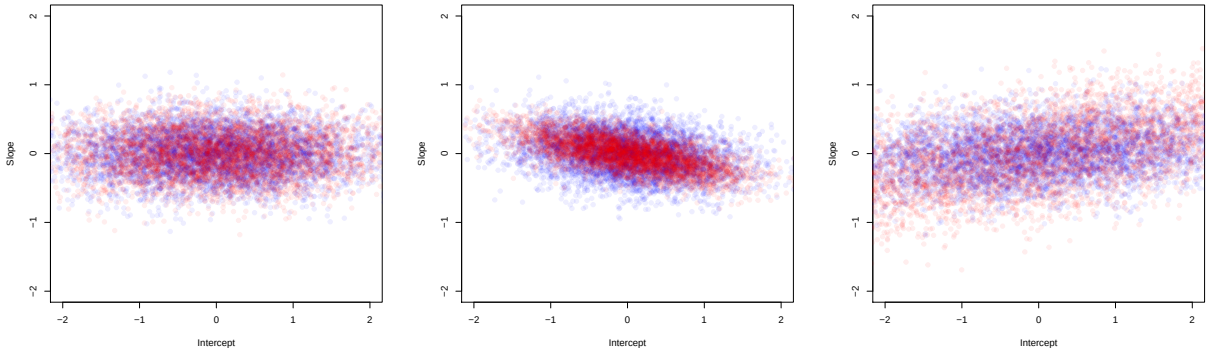

Figure 4.4: **Variability of Summaries of Individual Histories** These plots display the intercept and slope from the true data, the Empirical Bayes summaries, and the individual summaries respectively. Red patients have a low number of visits, while blue patients have a high number of visits. Patients are spread equally in truth, while Empirical Bayes shrinks low visit patients more than high visit patients, and low visit patients display more variability than high visit patients under individual summaries.

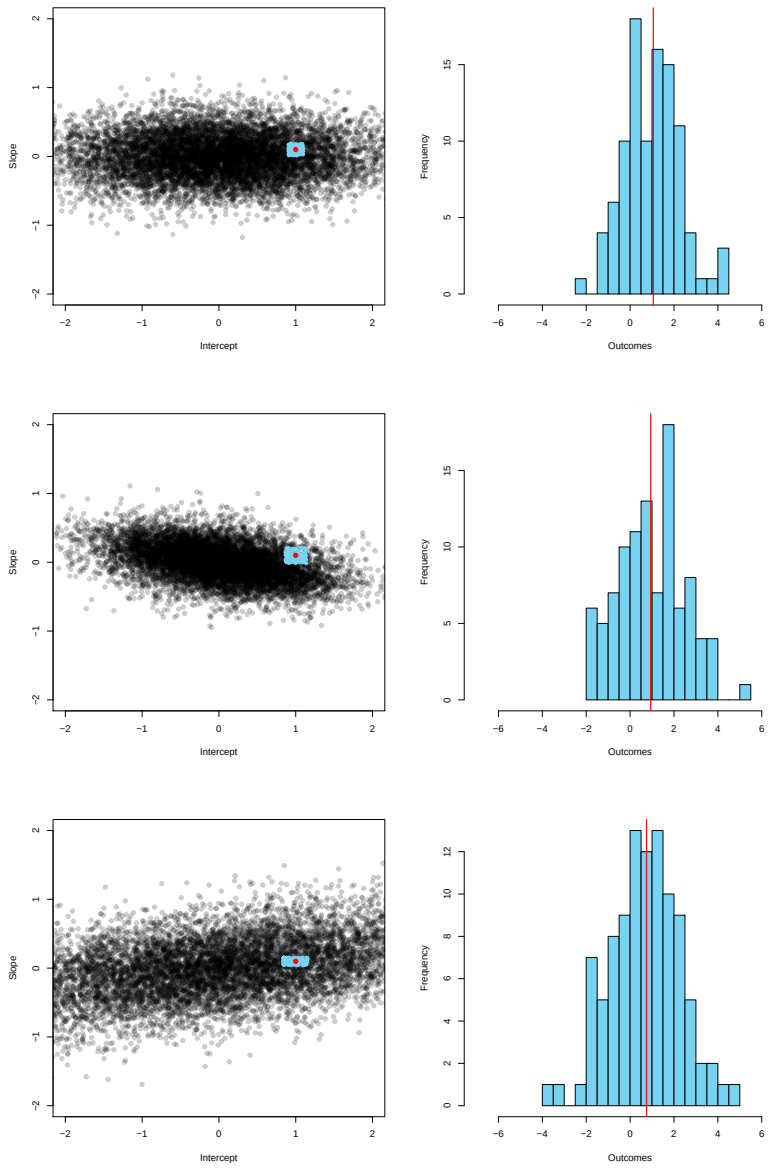

Figure 4.5: **Empirical Bayes Illustration** These plots display the neighborhood of 100 out of 10,000 patients around a patient with an intercept of 1 and a slope of .1 and the associated outcome. The top row is using true intercepts and slopes, the middle row using Empirical Bayes estimates, and the bottom row using a line for each patient. Red lines display the median of a given neighborhood.

	Min	Q1	Med	Q3	Max
Intercept	-8.89	-5.44	-0.28	4.90	14.71
Slope	-3.62	-0.49	0.01	0.56	3.63

Table 4.1: **Random Effects in BOLD Data** This table summarizes the distribution of random effects from our mixed model.

random effects – that for intercept and that for slope, described in Table 4.1 – that can be used as covariates in the construction of our neighborhood. In other words, users may select patients who have a similar trajectory or a similar mean to their patient. We set visit times such that time 3 is 0 in the model and the intercept is then an estimate of how far above or below the average Roland value at time 3 the patient lies.

We choose to examine 250 neighbors for a patient who is 75 years old and consider what happens with a medium and high average disability and a medium and high disability trajectory. We scale globally and our scaling values are 0.12 for age, 0.88 for intercept, and 2.73 for slope. Note that not only does the strength of the association between each variable and the outcome differ but so also does the range, and both of these contribute to the difference in scaling parameters. Unsurprisingly, the bulk of the outcomes shifts to higher disability values when our patient has a higher intercept – patients with worse histories have worse outcomes (Figure 4.6). We see similar results when considering patients with worse trajectories; a worse trajectory leads to a worse outcome on average (Figure 4.7). Also note that our neighborhood increases in size as we approach the edge of the data. In all scenarios we obtain our familiar bimodal outcome distribution. In addition to using one measure of history alone, we can also use both average disability and disability trajectory; we consider a patient with exactly average (i.e., 0) values on both and again we take 250 neighbors, as seen in Figure 4.8, and note that the two measures of history are rather unsurprisingly correlated.

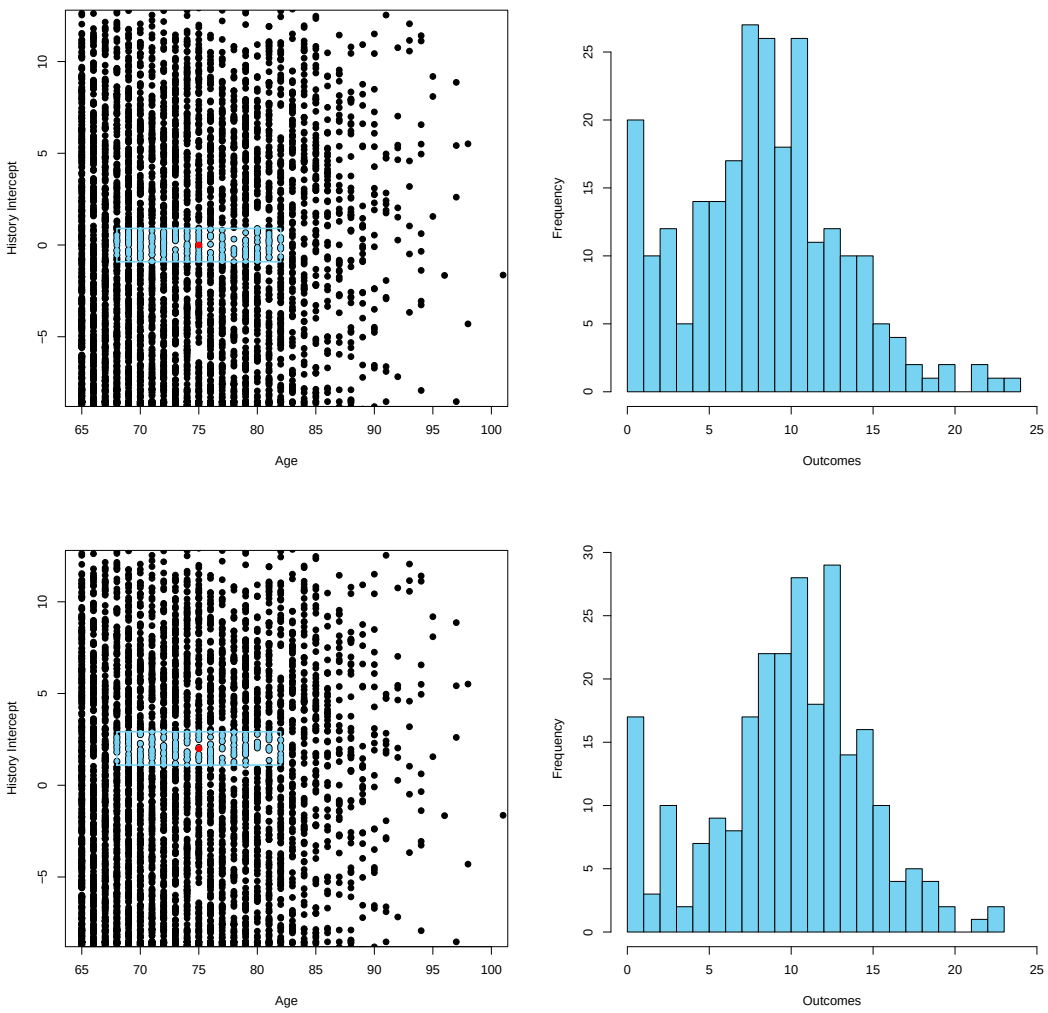

Figure 4.6: **Use of Average Patient History in BOLD Data** Here we examine a patient who is 75 years old. In the top set of plots, we consider a patient whose EB intercept is 0 (i.e., exactly average). The neighborhood for this patient extends from 68 to 82 years of age and from -0.91 to 0.91 on the EB intercept scale. In the second, we consider a patient with a higher than average intercept of 2. The neighborhood for this patient extends from 68 to 82 years of age and from 1.10 to 2.92 on the EB intercept scale. Each neighborhood has 250 patients and is globally scaled.

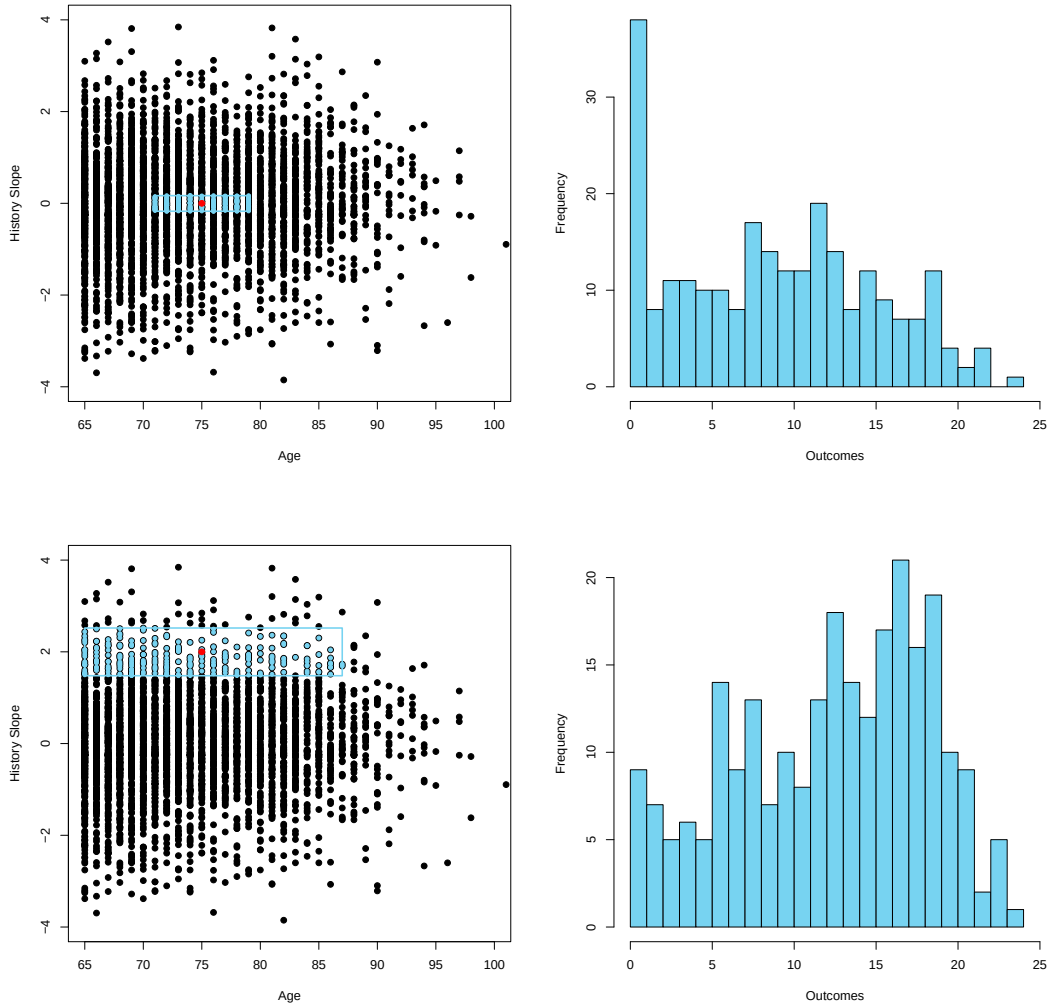

Figure 4.7: **Use of Patient Trajectory in BOLD Data** Here we examine a patient who is 75 years old. In the top set of plots, we consider a patient whose EB slope is 0 (i.e., exactly average). The neighborhood for this patient extends from 71 to 79 years of age and from -0.17 to 0.17 on the EB slope scale. In the second, we consider a patient with a higher than average slope of 2. The neighborhood for this patient extends from 65 to 87 years of age and from -1.48 to 2.52 on the EB slope scale. Each neighborhood has 250 patients and is globally scaled.

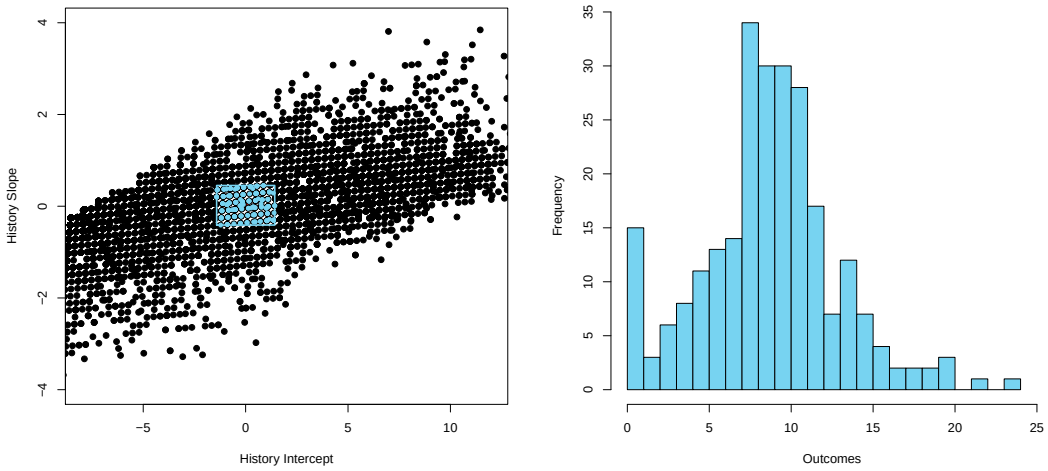

Figure 4.8: **Use of Patient Average History and Trajectory in BOLD Data** Here we examine a patient who has an average history on both slope and intercept (i.e., a value of 0 for both). Each neighborhood has 250 patients and is globally scaled. The neighborhood for this patient extends from an intercept of -1.42 to 1.42 and a slope of -.41 to .45.

4.6 Discussion

We use Empirical Bayes methods to reduce the dimension of patient histories in order to take advantage of the information contained within them. This has the additional advantage of shrinking patient values toward the overall mean, which prevents the values of patients for whom we have less information from becoming unreasonably large in magnitude. The summaries we obtain are easily interpretable as patient-specific intercepts, slopes, etc., and give us a convenient measure of how patients' previous experiences compare to one another. However, determining which parameters to estimate can require substantial subject-matter expertise. These history scores can be used in conjunction with other scores, such as the propensity score.

Chapter 5

DISCUSSION

We have succeeded in our goal to create an interpretable localized estimate of the distribution of patient outcomes with fixed precision. We created this estimate for subjects similar to rather than exactly the same as a given patient, and we collected those similar to the patient in an axis parallel neighborhood, the importance of whose axes were determined by adaptive neighborhood scaling. We note that, although our examples here are for an outcome score, our method can easily be adapted to survival data using local Kaplan-Meier methods. Our method would be particularly helpful for those hoping to examine subgroups in clinical trials, for medical practitioners wishing to examine the full range of likely outcomes for their patients, and for researchers developing inclusion/exclusion criteria for their studies based on the outcomes they require. It could also be used to assess the calibration of a higher-dimensional model.

The major limitation of our work is that it is suitable for low-dimensional situations only. Because dimension reduction is critical to our work, it is important to explore different ways to reduce multiple candidate predictors into summary scores or selected predictors and to communicate those summary scores in a meaningful manner. In general, we can consider three key types of scores: prognosis, treatment benefit, and propensity. We have already worked with the propensity score, although we additionally suggest consideration of a local propensity score, created in a similar manner to local scaling. A treatment benefit score is simply that, an estimate for each patient of whether or not the treatment will help him and to what degree. A prognostic score would consider the outcome for untreated patients, so that one could consider what actually happens to those we expect to have similar outcomes to a given patient.

We would welcome opportunities to work more on our tool, in order to operationalize it for specific disease areas and improve its functionality based on the principles of human-centered de-

sign. In particular, we hope to help the user in his choice of neighborhood size by showing him the relationship between the number of points in the neighborhood and the distance the neighborhood extends about the point of interest. This would more easily allow him to determine whether he is at a point of diminishing returns than is possible with our current tool. We would also like to determine whether there is a role for communicating and displaying the uncertainty associated with the empirical distribution of the outcome.

We ultimately see enormous potential for the development of personalized predictions due to the advent of EHRs and modern interest in connecting patient outcome data back to the patient himself. This research provides an initial exploration into novel methods that can be generalizable, reliable, and effective in a variety of clinical and experimental settings.

BIBLIOGRAPHY

[1] A. Aue, T.C.M. Lee, H. Wang, et al. Local bandwidth selection via second derivative segmentation. *Electronic Journal of Statistics*, 6:478–500, 2012.

[2] P.C. Austin. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Medical Decision Making*, 2009.

[3] P.C. Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

[4] M.A. Brookhart, S. Schneeweiss, K.J. Rothman, R.J. Glynn, J. Avorn, and T. Stürmer. Variable selection for propensity score models. *American journal of epidemiology*, 163(12):1149–1156, 2006.

[5] G. Casella and R.L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[6] J. Chang. Stochastic processes. As yet unpublished book., 1999.

[7] E.R. Dickson, P.M. Grambsch, T.R. Fleming, L.D. Fisher, and A. Langworthy. Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology*, 10(1):1–7, 1989.

[8] R.B. DAgostino, R.S. Vasan, M.J. Pencina, P.A. Wolf, M. Cobain, J.M. Massaro, and W.B. Kannel. General cardiovascular risk profile for use in primary care the framingham heart study. *Circulation*, 117(6):743–753, 2008.

[9] T.M. Egan, S. Murray, R.T. Bustami, T.H. Shearon, K.P. McCullough, L.B. Edwards, M.A. Coke, E.R. Garrity, S.C. Sweet, D.A. Heiney, and F.L. Grover. Development of the new lung allocation system in the united states. *American Journal of Transplantation*, 6(5p2):1212–1227, 2006.

[10] G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs. *Longitudinal data analysis*. CRC Press, 2009.

[11] S. Guo and M.W. Fraser. *Propensity score analysis: Statistical methods and applications*, volume 12. Sage Publications, 2009.

[12] B.B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

[13] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, and J.G. Conde. Research electronic data capture (redcap)a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics*, 42(2):377–381, 2009.

[14] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning*. Springer New York, 2009.

[15] P.J. Heagerty and B.A. Comstock. Exploration of lagged associations using longitudinal data. *Biometrics*, 69(1):197–205, 2013.

[16] J.C. Hill, K.M. Dunn, M. Lewis, R. Mullis, C.J. Main, N.E. Foster, and E.M. Hay. A primary care back pain screening tool: identifying patient subgroups for initial treatment. *Arthritis Care & Research*, 59(5):632–641, 2008.

[17] L. Horváth and P. Kokoszka. Functional principal components. In *Inference for Functional Data with Applications*, pages 37–43. Springer, 2012.

[18] K.A. Ibrahim, N. Paneth, E. LaGamma, and P.L. Reed. Clinician opinion to design clinical trials that change standards-of-cares. *Pediatric Research*, 2009.

[19] K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[20] J.G. Jarvik, B.A. Comstock, B.W. Bresnahan, S. Nedeljkovic, D.R. Nerenz, Z. Bauer, A.L. Avins, K. James, J.A. Turner, P.J. Heagerty, et al. Study protocol: The back pain outcomes using longitudinal data (bold) registry. *BMC Musculoskeletal Disorders*, 13(1):64, 2012.

[21] J.G. Jarvik, L.S. Gold, B.A. Comstock, P.J. Heagerty, S.D. Rundell, J.A. Turner, A.L. Avins, Z. Bauer, B.W. Bresnahan, J.L. Friedly, et al. Association of early imaging for back pain with clinical outcomes in older adults. *JAMA*, 313(11):1143–1153, 2015.

[22] A.K. Jha. Meaningful use of electronic health records: the road ahead. *JAMA*, 304(15):1709–1710, 2010.

[23] M.M. Joffe and P.R. Rosenbaum. Invited commentary: propensity scores. *American Journal of Epidemiology*, 150(4):327–333, 1999.

[24] P.S. Kamath, R.H. Wiesner, M. Malinchoc, W. Kremers, T.M. Therneau, C.L. Kosberg, G. D'Amico, E.R. Dickson, and W.R. Kim. A model to predict survival in patients with end-stage liver disease. *Hepatology*, 33(2):464 – 470, 2001.

[25] R. Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

[26] R. Koenker and K.F. Hallock. Quantile regression. *The Journal of Economic Perspectives*, 15(4):pp. 143–156, 2001.

[27] F.P. Leacy and E.A. Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study. *Statistics in medicine*, 33(20):3488–3508, 2014.

[28] W.C. Levy, D. Mozaffarian, D.T. Linker, S.C. Sutradhar, S.D. Anker, A.B. Cropp, I. Anand, A. Maggioni, P. Burton, M.D. Sullivan, et al. The seattle heart failure model prediction of survival in heart failure. *Circulation*, 113(11):1424–1433, 2006.

[29] D.J. Lowsky, Y. Ding, D.K.K. Lee, C.E. McCulloch, L.F. Ross, J.R. Thistlethwaite, and S.A. Zenios. Ak-nearest neighbors survival probability prediction method. *Statistics in medicine*, 32(12):2062–2069, 2013.

[30] D.F. McCaffrey, G. Ridgeway, and A.R. Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[31] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[32] R.E. and J.T. Cacioppo. *The elaboration likelihood model of persuasion*. Springer, 1986.

[33] H. Robbins. An empirical bayes approach to statistics. In J. Neyman, editor, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163. Univ of California Press, 1956.

[34] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[35] RStudio and Inc. *shiny: Web Application Framework for R*, 2013. R package version 0.5.0.

[36] D.B. Rubin and N. Thomas. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, 95(450):573–585, 2000.

[37] M. Segal and K. Kedem. Geometric applications of posets. *Computational Geometry*, 11(3):143–156, 1998.

[38] G.R. Shorack and J.A. Wellner. *Empirical processes with applications to statistics*, volume 59. SIAM, 2009.

[39] S.S. Skiena. *The Algorithm Design Manual*. Springer, 2009.

[40] E.A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

[41] W. Stüte. Conditional empirical processes. *The Annals of Statistics*, 14(2):638–647, 1986.

[42] W. Stüte. On almost sure convergence of conditional empirical distribution functions. *The Annals of Probability*, 14(3):891–901, 1986.

[43] A.W. van der Vaart and J.A. Wellner. Empirical processes indexed by estimated functions. *Lecture Notes-Monograph Series*, pages 234–252, 2007.

[44] H. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2012.

[45] S. Vansteelandt and R.M. Daniel. On regression adjustment for the propensity score. *Statistics in medicine*, 33(23):4053–4072, 2014.

[46] G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer, 2009.

Appendix A

INFERENCE: A DETAILED VERSION

A.1 Construction of Empirical Processes

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

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

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

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

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

The uniform empirical process is then

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

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

$$\mathbb{U}_n \xrightarrow{d} \mathbb{U}. \tag{A.5}$$

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

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

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

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

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

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

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

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

Therefore,

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

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., $X_1, \ldots, X_n$ 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

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

and rewrite our empirical process as

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

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

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

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 $K_n(s,t) \to K(s,t)$ 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 $\lambda \geq 0$, $P(||\mathbb{X}|| \leq \lambda) \geq P(||\mathbb{U}|| \leq \lambda)$ [38].

A.2 Uncertainty in Neighborhood Location

For the uncertainty in neighborhood location, we turn to the δ-method [5]. 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 [5] to obtain

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

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

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

where σ_Δ^2 is the asymptotic variance of $\sqrt{n}(\hat{\Delta} - \Delta)$. Because Δ is simply a quantile of the distance, we may use the formula for the asymptotic variance of a sample quantile [38] to obtain

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

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

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

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

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

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

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 [43] to evaluate this component and we obtain, where V_β is the covariance matrix of the scaling parameters,

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

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,

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

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 $\Delta_n \to 0$ as $n \to \infty$, we turn to results from Stüte to handle the uncertainty in outcome (Expression 2.4) [41]; see Appendix A.4.1 for a more detailed

explanation of the mathematics. Assuming that $np_n^3 \to \infty$, we obtain

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

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

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

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 $\frac{m}{n} \to 0$ and

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

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

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

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

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

A.4.1 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 Stüte [41], 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 $\underline{\mathbf{X}}$. We let $(X_1, Y_1), (X_2, Y_2), \ldots$ 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 $np_n^3 \to \infty$ as $n \to \infty$. We define

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

and

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

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

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

we obtain

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

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,

$$\begin{aligned} m_n(x^*) &= p_n^{-1} \int yK\left(\frac{F(x^*) - F(x)}{p_n}\right) H_n(dx, dy) && \text{(A.32)} \\ &+ p_n^{-2} \int y[F_n(x^*) - F_n(x) - F(x^*) + F(x)]K'\left(\frac{F(x^*) - F(x)}{p_n}\right) H_n(dx, dy) && \text{(A.33)} \\ &+ p_n^{-3} \int y[F_n(x^*) - F_n(x) - F(x^*) + F(x)]^2 \frac{K''(\Lambda)}{2} H_n(dx, dy) && \text{(A.34)} \\ &\equiv I_1 + I_2 + I_3, && \text{(A.35)} \end{aligned}$$

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

$$\sqrt{np_n} I_3 \xrightarrow{p} 0 \tag{A.36}$$

as $n \to \infty$. 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 $\limsup_{n\to\infty} \int |y| H_n(dx, dy) < \infty$ 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 $n \to \infty$. 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,

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

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

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

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

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

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

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

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

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

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

with corresponding von Mises statistic

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

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

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

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

$$\xrightarrow{p} 0 \tag{A.47}$$

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

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

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

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

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

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

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

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

as $n \to \infty$ and hence

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

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

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

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

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,

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

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

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,

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

$$= 0 \tag{A.60}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\text{Var}(I_4) \to h(x^*) \int K^2(u) du$ 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 $n \to \infty$, and for all $\delta > 0$

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

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

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

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

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

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 $a \to \infty$. 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

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

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

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

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

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