HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

J Am Stat Assoc. 2020 ; 115(529): 90–106. doi:10.1080/01621459.2019.1609969.

Quantile Function on Scalar Regression Analysis for Distributional Data

Hojin Yang[†], **Veerabhadran Baladandayuthapani**[†], **Arvind U.K. Rao**[‡], **Jeffrey S. Morris**[†]

[†]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

[‡]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract

Radiomics involves the study of tumor images to identify quantitative markers explaining cancer heterogeneity. The predominant approach is to extract hundreds to thousands of image features, including histogram features comprised of summaries of the marginal distribution of pixel intensities, which leads to multiple testing problems and can miss out on insights not contained in the selected features. In this paper, we present methods to model the entire marginal distribution of pixel intensities via the quantile function as functional data, regressed on a set of demographic, clinical, and genetic predictors to investigate their effects of imaging-based cancer heterogeneity. We call this approach *quantile functional regression*, regressing subject-specific marginal distributions across repeated measurements on a set of covariates, allowing us to assess which covariates are associated with the distribution in a global sense, as well as to identify distributional features characterizing these differences, including mean, variance, skewness, heavy-tailedness, and various upper and lower quantiles. To account for smoothness in the quantile functions, account for intrafunctional correlation, and gain statistical power, we introduce custom basis functions we call *quantlets* that are sparse, regularized, near-lossless, and empirically defined, adapting to the features of a given data set and containing a Gaussian subspace so non-Gaussianness can be assessed. We fit this model using a Bayesian framework that uses nonlinear shrinkage of quantlet coefficients to regularize the functional regression coefficients and provides fully Bayesian inference after fitting a Markov chain Monte Carlo. We demonstrate the benefit of the basis space modeling through simulation studies, and apply the method to Magnetic resonance imaging (MRI) based radiomic dataset from Glioblastoma Multiforme to relate imaging-based quantile functions to various demographic, clinical, and genetic predictors, finding specific differences in tumor pixel intensity distribution between males and females and between tumors with and without DDIT3 mutations.

Keywords

Basis Functions; Bayesian Modeling; Functional Regression; Imaging Genetics; Markov chain Monte Carlo; Probability Density Function

hyang6@mdanderson.org.

1. INTRODUCTION

Glioblastoma multiforme (GBM), also known as glioblastoma and grade IV astrocytoma, is the most common and most aggressive cancer that begins within the brain. Studying GBM is difficult in that the cause of most cases is unclear, there is no known way to prevent the disease, and most people diagnosed with GBM survive only 12 to 15 months, with less than 3% to 5% surviving longer than five years (Tutt 2011). Most GBM diagnoses are made by medical imaging such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET). MRI is frequently chosen because it offers a wide range of high-resolution image contrast that can serve as indicators for clinical decision making or for tumor progression in GBM studies. A GBM tumor, which usually originates from a single cell, demonstrates heterogeneous physiological and morphological features as it proliferates (Marusyk, Almendro and Polyak 2012). Those heterogeneous features make it difficult to predict treatment impacts and outcomes for patients with GBM. Investigating tumor heterogeneity is critical in cancer research since inter/intra-tumor differences have stymied the systematic development of targeted therapies for cancer patients (Felipe De Sousa, Vermeulen, Fessler and Medema 2013).

Our motivating dataset comes from The Cancer Imaging Archive (TCIA, cancerimagingarchive.net) – a comprehensive archive of biomedical images of various cancer types along with associated clinical and genomic data (described in detail in Section 4). As an illustration, the rightmost four plots of Figure 1 display MRI images for 4 patients with GBM, two males and two females, and with and without mutations in the DDIT3 gene, an important gene associated with GBM development, with tumor boundaries indicated by the black lines. The upper left plot contains smoothed density estimates of the pixel intensities while the bottom left plot contains the empirical quantile functions for these tumors. Features of these images may comprise clinically useful biomarkers since these pixel intensities denote the amount of contrast enhancement (or vascularization) on T1-weighted sequence; or extent of infiltration into neighboring tissue (in T2-weighted or fluid-attenuated inversion recovery (FLAIR) MR sequence). It is of scientific interest to study the pixel intensity distributions for a set of 64 GBM tumors of which these four are a subset and investigate their associations with various covariates including age, sex, tumor subtype, DDIT3 mutation status, EGFR mutation status, and survival status (> 12 months, ≤ 12 months), to assess how various aspects of GBM tumor heterogeneity are reflected in the tumor images.

Radiomics is a field of study to identify quantitative biomarkers from biomedical imaging data. The typical approach is to extract various features of the images and then relate them to various clinical and genetic outcomes. While some of these features characterize various spatial relationships among the pixel intensities, an important subset called *histogram features* (Just 2014) extracts information from the marginal distribution of pixel intensities within the tumor, such as the mean and variance. While the feature extraction strategy that is typical in radiomics is reasonable and often can yield meaningful results, it has numerous drawbacks. The exploratory regression analysis of numerous different summaries raises multiple testing problems, and if the key distributional differences are not contained in the pre-defined summaries, then this approach can miss out on important insights.

In this paper, we choose to represent and model the distribution through the quantile function, which has numerous advantages as described in Section 2.6, including a fixed, common domain [0, 1], their ease of estimation by order statistics without any need for smoothing parameter specification, and the ability to readily compute distributional moments. Thus, our approach is to represent each subject's data via their empirical quantile function $Q_i(p), p \in \mathcal{P} = [0, 1]$, computed from the order statistics, and then treat these as functional responses regressed on a set of scalar covariates $x_{ia}; j = 1, \ldots, A$ through $Q_i(p) = \beta_0(p) + \sum_{a=1}^{J} x_{ia}\beta_a(p) + E_i(p)$. This models the *distribution of subject-level distributions* as a function of subject-level covariates. We call the fitting of this model *quantile functional regression*, which is fundamentally different and distinguished from other models for quantile regression in existing literature in Section 2.1. Regression analysis using the quantile function as the response is based upon the Wasserstein metric between distributions (Dobrushin 1970), which can be shown to be equivalent to an L2 distance between the corresponding quantile functions.

One simple approach to fitting this model would be to interpolate each subject's data onto a common grid of $\mathcal{P}$ and then perform independent regressions for each interior point p. This would lead to estimators that are unbiased but ine cient, as they would not borrow strength across nearby p, which should be similar to each other. We refer to this strategy as *naive quantile functional regression*. As is typically done in other functional regression settings (see review article by (Morris 2015)), alternatively one could borrow strength across p using basis representations, with common choices including splines, principal components, and wavelets. In this paper, we will introduce a new strategy for construction of a custom basis set we call *quantlets* that is sparse, regularized, near-lossless, and empirically defined, adapting to the features of the given data set and containing the Gaussian distribution as a prespecified subspace so non-Gaussianness can be assessed. Representing the quantile functions with a quantlet basis expansion, we propose a Bayesian modeling approach for fitting the quantile functional regression model that utilizes shrinkage priors on the quantlet coefficients to induce regularization of the regression coefficients $\beta_a(p)$, and leading to a series of global and local inferential procedures that can first determine whether $\beta_a(p) \equiv 0$ and then assess which p and/or distributional summaries (e.g. mean/variance/skewness/Gaussianness) characterize any such difference. While based on quantile functions, our model will also be able to provide predicted distribution functions and densities for any set of covariates to use as summaries for users more accustomed to interpreting densities than quantile functions.

While developed in the context of our GBM motivating case study, the methods we develop are general and can be applied to a wide range of contexts in which multiple observations are obtained per subject and one wishes to associate subject-specific distributions to explanatory variables. This paper is organized as follows. In Section 2, we introduce the general quantile function regression model, introduce *quantlets*, describe how to construct a set of quantlet basis functions for a given data set, and describe our Bayesian approach to fitting the model. In Section 3, we describe simulation studies conducted to evaluate the finite-sample performance of our method and demonstrate the benefit of incorporating quantlet bases in the modeling. In Section 4, we apply our method to data in our GBM case

study and perform various investigations to obtain insightful scientific results. We provide concluding remarks in Section 5.

2. MODELS AND METHODS

2.1 Quantile Functions and Empirical Quantile Functions

Let Y be a real valued random variable which in our context, represents the pixel intensity from a tumor image in our GBM application, and $F_Y(y)$ be its cumulative distribution function (right-continuous) such that $F_Y(y) = \mathrm{P}(Y \leq y)$, and $p = F_Y(y)$ be the percentage of the population less than or equal to y. The quantile function of Y, defined for $p \in [0, 1]$, is defined as

$$Q(p) = Q_Y(p) = F_Y^{-1}(p) = \inf(y: F_Y(y) \geq p).$$

Given a sample of m repeated observations for a given subject, intensities for multiple pixels for the subject's tumor in our GBM application, let $Y_{(1)} \leq \cdots \leq Y_{(m)}$ be the corresponding order statistics. For $p \in [1/(m+1), m/(m+1)]$, a subject-specific empirical quantile function of Y can be computed, e.g. using linear interpolation across order statistics,

$$\widehat{Q}(p) = (1 - w)Y_{([(m+1)p])} + wY_{([m+1)p]+1)},$$

where $[x]$ is an integer less than or equal to x and w is a weight such that $(m + 1)p = [(m + 1)p] + w$. This empirical quantile function is an estimate of the true quantile function.

As shown in (Parzen 2004), for a fixed p, the empirical estimator is consistent and is asymptotically equivalent to a Brownian bridge when the density function $f_Y(y)$ exists and is positive. This can serve as a summary of the subject-specific distribution that does not require specification of any smoothing parameter, that in this paper we regress on outcomes to assess how they vary across covariates. In this paper, we are interested in studying outcomes Y that are absolutely continuous, meaning that the corresponding quantile functions are continuous and smooth, without jumps that would occur for discretely valued random variables. For brevity, we omit the estimator notation for the empirical quantile functions and just refer to them as $Q(p)$.

2.2 Quantile Functional Regression Model

Suppose that for a series of subjects $i = 1, \ldots, n$ we observe a sample of m_i observations from which we construct a subject-specific empirical quantile function $Q_i(p_j)$ for $p_j = j/(m_i + 1)$; $j = 1, \ldots, m_i$, along with a set of A covariates $\boldsymbol{X}_i = (x_{i1}, \ldots, x_{iA})^T$, which are the demographic, clinical, and genetic factors described in the introduction for our GBM application. Note that by construction all subject-specific empirical quantile functions $Q_i(p_j)$ are non-decreasing in p. See Section 4 of the supplement for further discussion of monotonicity issues in this framework.

The quantile functional regression model is given by

$$Q_i(p) = \sum_{a=1}^{A} x_{ia}\beta_a(p) + E_i(p) = \boldsymbol{X}_i^T \boldsymbol{B}(p) + E_i(p), \quad (1)$$

where $\boldsymbol{B}(p) = (\beta_1(p), \ldots, \beta_A(p))^T$ is a column vector of length A containing unknown fixed functional coefficients for the quantile p and $E_i(p)$ is a residual error process, assumed to follow a mean zero Gaussian process with the covariance surface, $\Sigma(p_1, p_2) = \text{cov}\{E_i(p_1), E_i(p_2)\}$ and to be independent of $\boldsymbol{X}_i$. The structure of $(\cdot, \cdot)$ captures the variability across subject-specific quantiles, and the diagonals capture the intrasubject covariance across p. In practice, we will focus our modeling on $p \in \mathcal{P} = [\delta, 1 - \delta]$, with $\delta = \max_{i \leq n}\{1/(m_i + 1)\}$ being the most extreme quantile estimable from the subject with the fewest observed data points. In this paper, we are primarily interested in settings with at least moderately large numbers of observations per subject, i.e. m_i not too small, and in later studies will extend our work to sparse data settings with few observations per subject.

To place our model in the proper context within the current literature on quantile and functional regression, Table 1 lists various types of regression in terms of response and objective function. In contrast to classical regression, which specifies the mean of the response conditional on a set of covariates, *quantile regression* (He and Liang 2000; Koenker 2005; Yang and He 2015) works by estimating a pre-specified p-quantile of the response distribution conditional on the covariates, either with independent (Koenker 2004; Hao and Naiman 2007; Davino, Furno and Vistocco 2013) or spatially/temporally correlated errors (Koenker 2004; Reich, Fuentes and Dunson 2012; Reich 2012). Most existing methods fit independent quantile regressions for each desired p, which can lead to crossing quantile planes, although recent methods (e.g. Yang and Tokdar (2017)) jointly model all quantiles, borrowing strength across p using Gaussian process priors. Parallel to these e orts are methods to perform *Bayesian density regression* (Dunson, Pillai and Park 2007), in which the density of the response variable is modeled as a function of covariates via dependent Dirichlet processes (Muller, Erkanli and West 1996; MacEachern 1999; Griffin and Steel 2006; Dunson 2006). These quantile regression models are inherently different from the setting of this paper, as they are modeling the quantile of the *population* given covariates, while our framework is modeling the quantile function of each *subject* as a function of subject-specific covariates. Another difference is that, in general, these methods do not model intrasubject correlation in settings for which there is more than one observation per subject.

Other regression methods have been designed for functional responses. There is a subset of the functional regression literature (see Morris (2015) for an overview) that involve regression of a functional response on a set of covariates, with classical functional regression focusing on the mean function conditional on covariates (Faraway 1997; Wu and Chiang 2000; Guo 2002; Ramsay and Silverman 2006; Morris and Carroll 2006; Reiss, Huang and Mennes 2010; Goldsmith, Wand and Crainiceanu 2011; Goldsmith, Bobb, Crainiceanu, Caffo and Reich 2012; Scheipl, Staicu and Greven 2015; Meyer, Coull, Versace, Cinciripini and Morris 2015), and *functional quantile regression* that computes the quantile of functional response conditional on covariates, using the check function as the objective function (Brockhaus, Scheipl, Hothorn and Greven 2015; Brockhaus and Rügamer 2015) or

the asymmetric Laplace likelihood as a Bayesian analog (Liu, Li and Morris 2018). Again these methods are not modeling subject-specific, but rather population-level quantiles. Other recent works on functional quantile regression have focused on the quantile of the scalar response distribution regressed on a set of functional covariates (Ferraty, Rabhi and Vieu 2005; Cardot, Crambes and Sarda 2005; Chen and Müller 2012; Kato 2012; Kato, Galvao and Montes-Rojas 2012; Li, Wang, Maity and Staicu 2016), which is also an inherently different problem from the one addressed here.

All of these methods differ, fundamentally, from the quantile functional regression framework described in this paper. For these methods, the quantile regression is computing the p^{th} quantile of the population given covariates X, while in our case, we are interested in modeling the p^{th} quantile of an individual subject's distribution given X. In our case, we are modeling the empirical quantile function for each subject as the response, and using a classical (mean) regression of these subject-specific quantile functions onto a set of scalar covariates, i.e. estimating the expected quantile function for a subject given a set of covariates. Note that this regression problem is based upon the Wasserstein metric between distributions (Dobrushin 1970), which can be shown to be equivalent to an L2 distance between the corresponding quantile functions. It would also be possible to compute the q^{th} quantile of the distribution of specific empirical quantile functions for each p conditional on covariates, which could be dubbed *quantile functional quantile regression*, but this model is not addressed in the current paper.

2.3 Quantlet Basis Functions

If all empirical quantile functions are sampled on (or interpolated onto) the same grid (i.e. $m_i \equiv m \forall i = 1, \ldots, n$), then a simple way to fit model (1) would be to fit separate linear regressions for each p. However, this naive approach would treat observations across p as independent. This leads to a regression model that fails to borrow strength across p, and thus is expected to be inefficient for estimation of the functional coefficients $\beta_a(p)$, and ignores correlation across p in the residual error functions $E_i(p)$, which would adversely affect any subsequent inference. We call this approach *naive quantile functional regression* in our comparisons below.

Basis function representations can be used to induce smoothness across p in $\beta_a(p)$ and capture intra-subject correlation in the residual error functions $E_i(p)$. In existing functional regression literature, common choices for basis functions include splines, Fourier, wavelets, and principal components, and smoothness is induced across p by regularization of the basis coefficients via L1 or L2 penalization (Morris 2015). Here, we introduce a strategy to construct a custom basis set called *quantlets* for use in the quantile functional regression model that have many desirable properties, including regularity, sparsity, near-losslessness, interpretability, and empirical determination allowing them to capture the salient features of the empirical quantile functions for a given data set.

We empirically construct the quantlets for a given data set as a common near-lossless basis that can nearly perfectly represent each subject's empirical quantile function, and then we use these basis functions as building blocks in our quantile functional regression model as

described later. Given a sample of subject-specific empirical quantile functions, we construct a quantlet basis set by the following steps:

1. Construct an overcomplete dictionary that contains bases spanning the space of Gaussian quantile functions plus a large number of Beta cumulative density functions. For each subject, use regularization to choose a sparse set among these dictionary elements.
2. Take the union of all selected dictionary elements across subjects, and find a subset that simultaneously preserves the information in each empirical quantile function to a specified level, as measured by the cross-validated concordance correlation coefficient.
3. Orthogonalize this subset using Gram-Schmidt, apply wavelet denoising to regularize the orthogonal basis functions, and then re-standardize.

We refer to the set of basis functions resulting from this procedure as *quantlets*. We describe these steps in detail and then discuss their properties. See Figure 2 for an overview of the entire procedure, for which each step is given as follows.

Form overcomplete dictionary: Suppose that $L^2(\Pi(\mathscr{P}))$ is a Banach space such that $\{Q: p \in \mathscr{P} \to \mathbb{R}$ measurable s.t. $\|Q\|_2 = (\int Q(p)|^2 d\Pi(p))^{1/2} < \infty\}$, where Π is a uniform density with respect to the Lebesgue measure. We define the first two basis functions to be a constant basis $\xi_1(p) = 1$ for $p \in [0, 1]$ and standard normal quantile function $\xi_2(p) = \Phi^{-1}(p)$. These orthonormal bases span the space of all Gaussian quantile functions, with the first coefficient corresponding to the mean and the second coefficient the variance of the distribution. We form an overcomplete dictionary that includes these along with a large number of dictionary elements constructed from Beta cumulative density functions (CDF). The shape of the Beta CDF is able to follow a "steep-flat-steep" shapes that we have observed characterize the features of empirical quantile functions in a wide array of applications, so has the potential for e cient representation.

The individual dictionary elements $\xi_k(p)$ are given by

$$\xi_k(p) = P_N{}^{\perp}\left(\frac{F_{\theta_k}(p) - \mu_{\theta_k}}{\sigma_{\theta_k}}\right) = P_N{}^{\perp}\left(\int_0^1 \left(I(u \le p) - \mu_{\theta_k}\right)/\sigma_{\theta_k} dF_{\theta_k}(u)\right), \quad (2)$$

where $F_{\theta_k}(p)$ is the CDF of a Beta(θ_k) distribution for some positive parameters $\theta_k = \{a_k, b_k\}$, $\mu_{\theta_k} = \int_0^1 F_{\theta_k}(u)du$ and $\sigma^2_{\theta_k} = \int_0^1 \left(F_{\theta_k}(u) - \mu_{\theta_k}\right)^2 du$ are the centered and scaled values of these distributions for standardization, respectively, and $P_N{}^{\perp}$ indicates the projection operator onto the orthogonal complement to the Gaussian basis elements $\xi_1(p)$ and $\xi_2(p)$, with $P_N{}^{\perp}\{f(p)\} = f(p) - \xi_1(p)\int_0^1 f(p)\xi_1(p)dp - \xi_2(p)\int_0^1 f(p)\xi_2(p)dp$. Put together, the set $\mathscr{D}^O = \{\xi_1, \xi_2\} \cup \{\xi_k : \theta_k \in \Theta\}$ comprises an *overcomplete dictionary* family on $\Theta \subset \mathbb{R}^2_+$. In practice, to fix the number of dictionary elements, we choose a grid on the parameter space to obtain $\Theta = \{\theta_k = (a_k, b_k)\}_{k=3}^{K^O}$ by uniformly sampling on $\Theta \subset (0, J)^2$ for some sufficiently

large J, and choosing K^O to be a large integer (e.g. we use $K^O = 12,000$ in this paper). Details of how to select Θ can be found in the Supplementary materials. If desired, this dictionary can be arbitrarily expanded to include any other basis functions on [0, 1] that one might think could capture salient features of the given data set.

The use of a large dictionary of Beta CDF in this step is supported by the following theorem, that demonstrates that any quantile function whose first derivative is absolutely continuous can be represented by a conical combination of Beta CDFs.

Theorem 2.1. *Let $Q(p)$ be a quantile defined on $p \in [0, 1]$, $F_{k,n}(p)$ be a beta cumulative distribution function defined as $F_{k,n}(p) = \int_0^p \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} x^k (1-x)^{n-k} dx$,* and *$q(p)$ be the first derivative of $Q(p)$. Define*

$$Q_n(p) = \sum_{k=0}^{n} c_{k,n} \int_0^p \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} x^k (1-x)^{n-k} dx = \sum_{k=0}^{n} c_{k,n} F_{k,n}(p),$$

where $c_{k,n} = a_k/(n+1)$, $a_k = q(k/n)$ and $0 \le x \le 1$. Suppose that $q(x): [0,1] \to \mathbb{R}$ be continuous function for the sufficiently small $\delta > 0$, that there exists a constant C such that $\|q\|_\infty = \sup_{x \in [0,1]} |q(x)| \le C$, and that $c_{k,n} \to c_k$ for each k, where c_k is some constant. Then for any $p \in [0, 1]$

$$\lim_{n \to \infty} Q_n(p) = Q(p).$$

This theorem provides justification for using a dictionary containing many Beta CDF to represent the empirical quantile functions, and supports the notion that given a large enough dictionary, the linear combination of beta CDFs should be sufficient for representing each individual's empirical quantile function.

Sparse selection of dictionary elements: For each i, we use regularization via penalized likelihood to obtain a sparse set of dictionary elements to represent each subject's empirical quantile function. While other choices of penalty could be used, here we use the Lasso (Tibshirani 1996), minimizing

$$\left\| Q_i(p) - \sum_{k \in \mathscr{D}^O} \xi_k(p) Q_{ik}^O \right\|_2^2 + \lambda_i \sum_{k \in \mathscr{D}^O} \left\| Q_{ik}^O \right\|_1, \tag{3}$$

for a fixed positive constant λ_i, where the choice of each λ_i is determined by cross validation and Q_{ik}^O are basis coefficients for the elements of $\mathscr{D}^O$. The standardization of the basis functions ensures they are on a common scale which is important for the regularization method. By using the regularization methods, we obtain different sets of selected *dictionary* elements for each subject, denoted by $\mathscr{D}_i = \left\{ \xi_k \in \mathscr{D}^O : Q_{ik}^O \neq 0 \right\}$. Taking the union across subjects, we obtain a unified set of *dictionary* elements denoted by $\mathscr{D}^U = \cup_{i=1}^n \mathscr{D}_i$, which we construct to always include the Gaussian basis functions ξ_1 and ξ_2.

Finding near-lossless common basis: The above sparse selection is done for each subject i, however, we would like to use a common basis across all subjects to fit the quantile functional regression model. The unified set of dictionary elements $\mathscr{D}^U$ is likely to be very redundant, with some of the dictionary elements selected for many subjects' empirical quantile functions and many others selected for only a few subjects, and not all necessary. We would like to find a common basis set $\mathscr{D}^{\mathscr{C}}$ that is as sparse as possible while retaining virtually all of the information in the original empirical quantile functions. We call such a basis *near-lossless*, which we define more precisely below.

As a measure of losslessness, we use the leave-one-out concordance correlation coefficient (LOOCCC), $\rho_{(i)}$. This quantifies the ability of a basis set $\mathscr{D}^U_{(i)}$ that has been empirically constructed using all samples except the ith one to represent the observed quantile function

$$\rho_{(i)} = \frac{\mathrm{Cov}\left(Q_i(\cdot), \sum_{k \in \mathscr{D}^U_{(i)}} \xi_k(\cdot) Q^U_{ik}\right)}{\mathrm{Var}(Q_i(\cdot)) + \mathrm{Var}\left(\sum_{k \in \mathscr{D}^U_{(i)}} \xi_k(\cdot) Q^U_{ik}\right) + \left[\mathrm{E}(Q_i(\cdot)) - \mathrm{E}\left(\sum_{k \in \mathscr{D}^U_{(i)}} \xi_k(\cdot) Q^U_{ik}\right)\right]^2}, \quad (4)$$

where Cov, Var and E are taken with respect to Π and Q^U_{ik} are basis coefficients corresponding to the elements ξ_k contained in the set $\mathscr{D}^U_{(i)}$.

This measure $\rho_{(i)} \in [0, 1]$, with $\rho_{(i)} = 1$ indicating the basis set $\mathscr{D}^U_{(i)}$ is sufficiently rich such that there is no loss of information about $Q_i(p)$ in its corresponding projection. One advantage of this measure over other choices such as mean squared error is that it is scale-free, in the sense that it is invariant to the scale of the quantile functions Q_i and the basis functions ξ_k. Aggregating across subjects, we can compute $\rho^0 = \min_i\{\rho_{(i)}\}$ or $\bar{\rho} = \mathrm{mean}_i\{\rho_{(i)}\}$ to summarize the ability of the chosen basis to reconstruct the observed data set in its entirety, with $\bar{\rho}$ the average across all subjects and ρ^0 the worst case. If $\rho^0 = 1$, we say this basis is *lossless*, and if $\rho^0 > 1 - \epsilon$ for some small ϵ then we say this basis is *near-lossless*.

To find a sparse yet near-lossless basis set, we define a sequence of reduced basis sets $\left\{\mathscr{D}^U_{(i)c}, \mathscr{C} = 1, \ldots, n-1\right\}$ that contain the Gaussian basis functions ξ_1 and ξ_2 plus all dictionary elements $\xi_k(p)$ that are selected for at least $\mathscr{C}$ of the $n-1$ empirical quantile functions, excluding the ith one, i.e. $\mathscr{D}^U_{(i)c} = \left\{\xi_k, k: \sum_{i' \neq i = 1}^n I\left(Q^O_{i'k} \neq 0\right) \geq \mathscr{C}\right\}$. We can construct plots of ρ^0 or $\bar{\rho}$ vs. $\mathscr{C}$ to choose a value of $\mathscr{C}$ that leads to a sparse basis that can recapitulate the observed data at the desired level of accuracy (as shown below). Given this choice, we next compute the corresponding reduced basis set using all of the data $\mathscr{D}^c = \{\xi_k, k: \sum_{i=1}^n I\left(Q^O_{ik} \neq 0\right) \geq \mathscr{C}\}$ containing $K = K_{\mathscr{C}}$ basis coefficients. The left panel of Figure 3 contains this plot for our GBM data set. From this, we select $\mathscr{C} = 10$ which leads to $K_{\mathscr{C}} = 27$ basis functions as this number of basis preserves a concordance of at least $\rho^0 = 0.990$ for each subject ($\epsilon = 0.01$) and an average concordance of $\bar{\rho} = 0.998$.

Orthogonalization and Denoising: Next, we use Gram-Schmidt to orthogonalize the basis set $\mathscr{D}^{\mathscr{C}}$ to generate an orthogonalized basis set $\mathscr{D}^{\perp} = \left\{\psi_k^{\perp}(p), k = 1, \ldots, K\right\}$, where $\left\{\psi_1^{\perp}(\cdot), \psi_2^{\perp}(\cdot)\right\} = \{\xi_1(\cdot), \xi_2(\cdot)\}$, comprise the Gaussian basis and $\left\{\psi_k^{\perp}(\cdot), k = 3, \ldots, K\right\}$ are orthogonalized basis functions computed from and spanning the same space as the remaining bases in $\mathscr{D}^{\mathscr{C}}$, indexed in descending order of their total percent variability (total energy) explained for the given data set. Specifically, suppose that $Q_{ik}^{\perp}$, $k = 1, \ldots, K$ and $i = 1, \ldots, n$ are the empirical coefficients corresponding to the elements of $\mathscr{D}^{\perp}$, ordered as in $\mathscr{D}^{\mathscr{C}}$. We compute the percent total energy for basis k as $\mathscr{E}_k = \sum_{i=1}^{n} Q_{ik}^{\perp\,2} / \sum_{i=1}^{n} \sum_{k=1}^{k} Q_{ik}^{\perp\,2}$, and then relabel ψ_k, $k = 3, \ldots, K$ to be in descending order of $\mathscr{E}_k$.

In practice, we have observed that the first number of orthogonal basis functions are relatively smooth, but the later basis functions can be quite noisy, sometimes with high-frequency oscillations. As we do not believe these oscillations capture meaningful features of the empirical quantile functions, we regularize the orthogonal basis functions using wavelet denoising to adaptively remove these oscillations. For a choice of mother wavelet function $\varphi_{j,l}(p) = 2^{j/2}\varphi(2^j p - l)$ with integers j, l, we construct the wavelet shrunken and denoised basis function $\psi_k^{\dagger}(p)$ (Donoho, Johnstone, Kerkyacharian and Picard 1995), given by

$$\psi_k^{\dagger}(p) = \sum_{j=0}^{J} \sum_{l=1}^{L_j} d_{k,j,l}^{\dagger} \varphi_{j,l}(p), \tag{5}$$

where L is a grid of size $L = 2^{10} = 1024$ for our GBM data, $d_{k,j,l} = \int \psi_k^{\perp}(p)\varphi_{j,l}(p)dp = \left\langle \psi_k^{\perp}, \varphi_{j,l} \right\rangle$, $d_{k,j,l}^{\dagger} = d_{k,j,l}$ if $\left|d_{k,j,l}\right| > \sigma\sqrt{2\log L}$ and $d_{k,j,l}^{\dagger} = 0$ If $\left|d_{k,j,l}\right| \leq \sigma\sqrt{2\log L}$. We use an empirical estimator for σ that is the median absolute deviation of the wavelet coefficients at the highest frequency level J. Details for denoising are described in Section 1.1 of the supplement.

After applying the denoising method to all of the orthogonal basis functions in the set $\mathscr{D}^{\perp}$ to get $\mathscr{D}^{\dagger} = \left\{\psi_k^{\dagger}(p), k = 1, \ldots, K\right\}$, we re-standardize these basis functions by $\psi_k(p) = \left(\psi_k^{\dagger}(p) - \mu_k^{\dagger}\right)/\sigma_k^{\dagger}$ for $k = 3, \ldots, K$ with $\mu_k^{\dagger} = \int_0^1 \psi_k^{\dagger}(p)dp$ and $\sigma_k^{\dagger} = \sqrt{\int_0^1 \left\{\psi_k^{\dagger}(p) - \mu_k^{\dagger}\right\}^2 dp}$ such that $\int_0^1 \psi_k(p)dp = 0$ and $\int_0^1 \psi_k(p)\psi_k(p)dp = 1$ for $k = 3, \ldots, K$.

We refer to the resulting basis set $\mathscr{D} = \{\psi_k(p), k = 1, \ldots, K\}$ as the *quantlets*, which we use as the basis functions in our quantile functional regression modeling. Figure 4 contains the first 16 quantlet basis functions from the GBM data set.

Properties of quantlets: These quantlets have numerous properties that makes them useful for modeling in our quantile functional regression framework.

- **Empirically defined:** The empirical quantile functions for different applications can have very different features and characteristics. Given their derivation from the observed data, the quantlets are customized to capture the features underlying the given data set, giving them advantages over pre-specified bases like splines, wavelets, or Fourier series.
- **Near-losslessness:** By construction, the set of quantlets are at least *near-lossless* in the sense that the basis is sufficiently rich to almost completely recapitulate the empirical quantile functions $Q_i(p)$. As a result, we can project the empirical quantiles into the space spanned by the quantlets with negligible error, and thus it is reasonable to consider modeling the quantlet coefficients for the empirical quantile functions as observed data.
- **Regularity:** The denoising step tends to remove any wiggles or high frequency noise from the orthogonal basis functions $\psi_k^{\perp}(p)$, leading to visually pleasing yet adaptive basis functions that are relatively smooth and regular. We have found these tend to be more regular looking than other empirically determined basis functions like principal components (compare Figure 4 to Supplementary Fig 5).
- **Sparsity:** The procedure we have defined to construct the quantlets tends to also produce a basis set that is relatively low dimensional and thus a sparse representation. We have found these basis functions to have similar sparsity to principal component bases, measured by computing the average LOOCCC $\bar{\rho}$ for quantlets and analogously for principal components (i.e. computing the principal components leaving out the ith sample, and then estimating $\rho_{(i)}$ measuring the losslessness of the resulting basis set) – see Figure 3B and Figure 5C. Using a low dimensional basis enhances the computational speed of our procedure and reduces the uncertainty in the quantile functional regression coefficients $\beta_d(p)$, as can be seen in our sensitivity analyses (Supplementary Table 5).
- **Interpretability:** Unlike principal components, the quantlets have some level of interpretability in that the first two basis functions define the space of all Gaussian quantile functions (see Figure 4). For Gaussian data, only the first two basis functions will be needed, while comparing with dimensions $k = 3, \ldots, K$ provides a measure of the degree of *non-Gaussianity* in the distribution. The remaining quantlets for $k \geq 3$ are not necessarily interpretable since they are empirically determined, but by our observation for many data sets the next two quantlets capture some sense of skewness and some sense of heavy-tailedness like kurtosis.

2.4 Quantlet-based Modeling in Quantile Functional Regression

Given the ith empirical quantile function $Q_i(p_j)$ evaluated at $p_j = j/(m_i + 1)$, $j = 1, \ldots, m_i$, constructed from the order statistics $Y_{i(j)}$, $j = 1, \ldots, m_i$, and a quantlet basis set $\mathscr{D} = \{\psi_k(p), k = 1, \ldots, K\}$ derived as described in Section 2.3, we write a quantlet basis expansion $Q_i(p_j) = \sum_{k=1}^{k} Q_{ik}^{*} \psi_k(p_j)$ with Q_{ik}^{*} being the kth empirical quantlet basis function for subject i. For this paper, we will assume that $K < \min_i(m_i)$, with the understanding that K

$\ll \min_i(m_i)$ for an extremely large number of applications, including our GBM data. Extensions of this framework to sparse data settings for which $m_i < K$ for some i are tractable and of interest, but given the length and complexity of this paper and the additional challenges raised by this sparse case, we will leave it to future work.

With $\boldsymbol{Q}_i = \left[Q_i(p_1), \ldots, Q_i\left(p_{m_i}\right)\right]$ a row vector containing the ith empirical quantile function and $\boldsymbol{\Psi}_i$ a $K \times m_i$ matrix with element $\boldsymbol{\Psi}_i(k, j) = \psi_k(p_j)$, we can compute the $1 \times K$ vector of empirical quantlet coefficients $\boldsymbol{Q}_i^* = \left[Q_{i1}^*, \ldots, Q_{iK}^*\right]$ by $\boldsymbol{Q}_i^* = \boldsymbol{Q}_i \boldsymbol{\Psi}_i^-$, where $\boldsymbol{\Psi}_i^- = \boldsymbol{\Psi}_i^T\left(\boldsymbol{\Psi}_i \boldsymbol{\Psi}_i^T\right)^{-1}$ is the generalized inverse of $\boldsymbol{\Psi}_i$. Based on the *near-lossless* property of the quantlets by design, $\boldsymbol{Q}_i^*$ contains virtually all of the information in the raw data $\boldsymbol{Q}_i$, and thus we model these as our data. Concatenating $\boldsymbol{Q}_i^*$ across the n subjects, we are left with a $n \times K$ matrix $\boldsymbol{Q}^*$, and consider obtaining estimates and inference on the quantiles and parameters of model (1) on any desired grid of p of size J, by $\boldsymbol{Q}(\mathscr{P}) = \boldsymbol{Q}^* \boldsymbol{\Psi}$ and $\boldsymbol{B}(\mathscr{P}) = \boldsymbol{B}^* \boldsymbol{\Psi}$ with $\boldsymbol{\Psi}$ a $K \times J$ matrix with elements $\psi_k(p_j)$, where $\boldsymbol{B}^*$ an $A \times K$ matrix of corresponding quantlet-space regression coefficients.

The Wasserstein distance between cumulative distribution functions (Bickel and Freedman 1981) is defined as $L(F, G) = \inf_{U,V} \|U - V\|_m$ for F and G two distribution, where all pairs of random variables (U, V) are followed from F and G, respectively. Following Bellemare, Dabney and Munos (2017), the infimum is attained by the inverse transformation of a random variable $\mathscr{P}$ uniformly distributed on [0, 1]. i.e., $L(F, G) = \int_0^1 \left|F^{-1}(p) - G^{-1}(p)\right|^m dp$. The quantile functional regression model of (1) is a framework for the Wasserstein distance with $m = 2$, minimizing the empirical risk $\sum_{i=1}^n \int_0^1 \left|Q_i(p) - \sum_{a=1}^A x_{ia}\beta_a(p)\right|^2 dp$. Based on the matrix notation, we rewrite the empirical risk for the Wasserstein loss function as

$$\operatorname{tr}\left[(\boldsymbol{Q}^*\boldsymbol{\Psi} - \boldsymbol{X}\boldsymbol{B}^*\boldsymbol{\Psi})(\boldsymbol{Q}^*\boldsymbol{\Psi} - \boldsymbol{X}\boldsymbol{B}^*\boldsymbol{\Psi})^T\right], \tag{6}$$

where $\boldsymbol{X}$ is an $n \times A$ matrix with $\boldsymbol{X}(i, a) = x_{ia}$. It follows from the corresponding normal equation $\boldsymbol{X}^T\boldsymbol{Q}^*\boldsymbol{\Psi}\boldsymbol{\Psi}^T = \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{B}^*\boldsymbol{\Psi}\boldsymbol{\Psi}^T$ that the minimizers $\widehat{\boldsymbol{B}^*} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Q}^*$ is seen to be a point estimator in a quantile functional regression framework like ours. This partially motivates our approach of performing the regressions on the quantile scale.

We can also consider regressing on the covariates in the *quantlet space model*

$$\boldsymbol{Q}^* = \boldsymbol{X}\boldsymbol{B}^* + \boldsymbol{E}^*, \tag{7}$$

where $\boldsymbol{E}^*$ an $n \times K$ matrix of quantlet space residuals. From (7), we can relate this quantlet-space model back to the original quantile functional regression model (1) through the quantlet basis expansions $\beta_a(p) = \sum_{k=1}^K B_{ak}^* \psi_k(p)$ and $E_i(p) = \sum_{k=1}^K E_{ik}^* \psi_k(p)$. The rows of $\boldsymbol{E}^*$ are assumed to be independent and identically distributed mean-zero Gaussians, with $\mathbf{E}_i^* \sim N(0, \Sigma^*)$, where $A_{i\cdot}$ or $A_{\cdot i}$ denotes the ith row or column of the matrix A. Here, we assume $\Sigma^* = \operatorname{diag}_k\left\{\sigma_k^2\right\}$, which enables us to fit in parallel the models for each column,

$\boldsymbol{Q}^*_{\cdot k} = \boldsymbol{X}\boldsymbol{B}^*_{\cdot k} + \boldsymbol{E}^*_{\cdot k}$, $k = 1, \ldots, K$, and yet accommodate correlation across p since modeling in the quantlet space induces correlation in the original data space, with the covariance operator for $E_i(p)$ given by $\Sigma(p, p') = \text{cov}\{E_i(p), E_i(p')\} = \boldsymbol{\Psi}(p)\Sigma^*\boldsymbol{\Psi}(p')$, where $\boldsymbol{\Psi}(p) = (\psi_1(p), \ldots, \psi_K(p))^T$. The empirical nature of the derived quantlets makes this structure well-equipped to capture the key correlations across p in the observed data, as shown for our real data set (See Supplementary Figure 9). If desired, one could model Σ^* as an unconstrained $K \times K$ matrix, which would provide additional flexibility in the precise form of Σ but at a potentially much greater computational cost.

2.5 Bayesian Modeling Details

This model could be fit using vague conjugate priors for the regression coefficients, $B^*_{ak} \sim N(0, \tau^2)$ for some extremely large τ^2. This could be called a *quantlet-no sparse regularization* approach. It would result in virtually no smoothing of $\beta_a(p)$ relative to the naive (one-p-at-a-time) quantile functional regression model, but it would still account for correlation across p in the residual errors, so may have inferential advantages over the naive approach. We can further improve performance by inducing regularity and smoothness in the quantile functional regression parameters $\beta_a(p)$, which we accomplish through regularization or shrinkage priors, as is customary for Bayesian functional regression models.

In order to fit model in the quantlet space model using a Bayesian approach, we also need to specify priors on the variance components $\{\sigma^2_k, k = 1, \ldots, K\}$. We place a vague proper inverse gamma prior on each diagonal element σ^2_k given by $\sigma^2_k \sim$ inverse-gamma($\nu_0/2$, $\nu_0/2$), where ν_0 is some relatively small positive constants. Other relatively vague priors could also be used. If one wanted to allow Σ^* to be unconstrained, an Inverse Wishart prior could be assumed for the $K \times K$ matrix. The likelihood function is given $\boldsymbol{Q}^*_{\cdot k} \sim N\left(\boldsymbol{X}\boldsymbol{B}^*_{\cdot k}, \sigma^2_k \boldsymbol{I}\right)$ in the projected space for each $k = 1, \ldots, K$.

We fit the quantlet space model in (7) using Markov chain Monte Carlo (MCMC). Let $\mathbf{Q}^*_{\cdot k}$ and $\boldsymbol{B}^*_{\cdot k}$ be the kth column vector of $\boldsymbol{Q}^*$ and $\boldsymbol{B}^*$, respectively. For each quantlet basis $k = 1, \ldots, K$, we sample the ath covariate effect from $f\left(B^*_{ak} | \mathbf{Q}^*, \boldsymbol{B}^*_{(-a)k}, \sigma^2_k\right)$, where $\boldsymbol{B}^*_{(-a)k}$ is a vector of length $A - 1$ containing all covariate effects except the ath of B^* in model (7) for the kth quantlet coefficient. We repeat this procedure for all covariates, $a = 1, \ldots, A$ and quantlet basis function $k = 1, \ldots, K$. This distribution is a mixture of a point mass at zero and a normal distribution, with normal mixture proportion α_{ak} and the mean and variances of the normal distribution μ_{ak} and v_{ak} given by

$$B^*_{ak} \equiv B^*_{ah_k, l} \sim \alpha_{ah_k, l} N\left(\mu_{ah_k, l}, v_{ah_k, l}\right) + \left(1 - \alpha_{ah_k, l}\right) I_0$$

where $\alpha_{ah_k, l}$, $\mu_{ah_k, l}$ and $v_{ah_k, l}$ are given by

$$\alpha_{ah_k, l} = \text{P}\left(\gamma_{ah_k, l} = 1 \middle| \mathbf{Q}^*_{k,k}, \boldsymbol{B}^*_{(-a)k}, \sigma^2_k\right) = \hat{O}_{ah_k, l} / \left(\hat{O}_{ah_k, l} + 1\right),$$

$$\mu_{ah_k,l} = \hat{B}^*_{ah_k,l}\left(1 + V_{ah_k,l}/\tau_{ah_k,l}\right)^{-1}, \quad \upsilon_{ah_k,l} = V_{ah_k,l}\left(1 + V_{ah_k,l}/\tau_{ah_k,l}\right)^{-1},$$

$$\hat{O}_{ah_k,l} = \frac{\hat{\pi}_{ah}}{1-\hat{\pi}_{ah}}\left(1 + V_{ah_k,l}/\tau_{ah_k,l}\right)^{-1/2}\exp\left\{\frac{1}{2}\zeta^2_{ah_k,j}\frac{V_{ah_k,l}/\tau_{ah_k,l}}{1 + V_{ah_k,l}/\tau_{ah_k,l}}\right\},$$

$$\zeta_{ah_k,l} = \hat{\beta}^*_{ah_k}/V^{1/2}_{ah_k,l}, \quad V_{ah_k,l} = \left(\sum_{i=1}^{n} x^2_{ia}/\sigma^2_k\right)^{-1},$$

and $\hat{B}^*_{ah_k,l}$ is frequentist estimator mentioned in Subsection 2.5. For each quantlet basis $k = 1, \ldots, K$, we sample σ^2_k from its complete conditional

$$\mathrm{P}\left(\sigma^2_k \middle| \boldsymbol{B}^*_{\cdot k}, \mathbf{Q}^*_{\cdot k}, \boldsymbol{X}\right) \sim \text{Inverse Gamma}\left\{(\nu_0 + n)/2, \left(\nu_0 + \mathrm{SSE}\left(\boldsymbol{B}^*_k\right)\right)/2\right\},$$

where $\mathrm{SSE}\left(\boldsymbol{B}^*_{\cdot k}\right) = \mathbf{Q}^{*T}_{\cdot k}\left(\mathbf{I} - \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\right)\mathbf{Q}^*_{\cdot k}$. (See Subsection 1.2 of the supplement for details of MCMC).

2.6 Posterior Inference

After obtaining posterior samples for all quantities in the *quantlet space* model (7), these posterior samples are transformed back to the *data space* using $\beta^{(m)}_a(p) = \sum_{k=1}^{K} B^{*(m)}_{ak}\psi_k(p)$, $m = 1, \ldots, M$ where M is the number of MCMC samples after burn in and thinning. From these posterior samples, various Bayesian inferential quantities can be computed, including point wise and joint credible bands, global Bayesian p-values, and multiplicity-adjusted probability scores, as detailed below. These can be computed for $\beta_a(p)$ itself or any transformation, functional, or contrast involving these parameters.

Point and joint credible bands: Pointwise credible intervals for $\beta_a(p)$ can be constructed for each p by simply taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior samples. Use of these local bands for inference does not control for multiple testing, however. Joint credible bands have global properties, with the $100(1 - \alpha)\%$ joint credible bands for $\beta_a(p)$ satisfying $\mathrm{P}(L(p) \le \beta_a(p) \le U(p) \quad \forall p \in \mathscr{P}) \ge 1 - \alpha$. Using a strategy as described in (Ruppert, Wand and Carroll 2003), we can construct joint bands by

$$\boldsymbol{J}_{a,\alpha}(p) = \hat{\beta}_a(p) \pm q_{(1-\alpha)}\left[\widehat{\mathrm{St{\cdot}Dev}}\left\{\hat{\beta}_a(p)\right\}\right], \tag{8}$$

where $\hat{\beta}_a(p)$ and $\widehat{\mathrm{St.Dev}}\left\{\hat{\beta}_a(p)\right\}$ are the mean and standard deviation for each fixed p taken over all MCMC samples. Here the variable $q_{(1-\alpha)}$ is the $(1-\alpha)$ quantile taken over all MCMC samples of the quantity

$$Z_a^{(m)} = \max_{p \in \mathcal{P}} \left| \frac{\beta_a^{(m)}(p) - \hat{\beta}_a(p)}{\widehat{\text{St·Dev}}\left\{\hat{\beta}_a(p)\right\}} \right|.$$

SimBaS and GBPV: Following Meyer et al. (2015) we can construct $J_{a,\alpha}(p)$ for multiple levels of α and determine for each p the minimum α such that 0 is excluded from the joint credible band, which we call *Simultaneous Band Scores (SimBaS)*, $\mathrm{P}_{a,SimBaS}(p) = \min\{\alpha : 0 \notin J_\alpha(p)\}$, which can be directly estimated by

$$\mathrm{P}_{a,\ SimBaS}(p) = M^{-1} \sum_{m=1}^{M} I\left\{ \left| \frac{\hat{\beta}_a(p)}{\widehat{\text{St·Dev}}\left\{\hat{\beta}_a(p)\right\}} \right| \le Z_a^{(m)} \right\}.$$

These can be used as local probability scores that have global properties, effectively adjusting for multiple testing. For example, we can flag all $\{p : \mathrm{P}_{a,SimBaS}(p) < \alpha\}$ as significant. From these we can compute $P_{a,Bayes} = \min_p\{P_{a,SimBaS}(p)\}$, which we call *global Bayesian p-values* (GBPV) such that we reject the global hypothesis that $\beta_a(p) \equiv 0$ whenever $P_{a,Bayes} < \alpha$.

Probability scores for distributional moments: As mentioned in Section 2.1, distributional moments can be constructed as straightforward functions of the quantile function, and thus from posterior samples of quantile functional regression parameters one can construct posterior samples of these moments for various levels of covariates $\boldsymbol{X}$. Denoting $\beta^{(m)}(p) = \left(\beta_1^{(m)}(p), \ldots, \beta_A^{(m)}(p)\right)^T$ for each MCMC sample $m = 1, \ldots, M$, posterior samples of distributional moments conditional on $\boldsymbol{X}$ are given by

$$\begin{aligned}
\mu_{\boldsymbol{X}}^{(m)} &= \int_0^1 \boldsymbol{X}^T \beta^{(m)}(p) dp \\
\sigma_{\boldsymbol{X}}^{2(m)} &= \int_0^1 \left(\boldsymbol{X}^T \beta^{(m)}(p) - \mu_{\boldsymbol{X}}^{(m)}\right)^2 dp, \\
\xi_{\boldsymbol{X}}^{(m)} &= \int_0^1 \left(\boldsymbol{X}^T \beta^{(m)}(p) - \mu_{\boldsymbol{X}}^{(m)}\right)^3 / \sigma_{\boldsymbol{X}}^{3(m)} dp, \text{ and} \\
\varphi_{\boldsymbol{X}}^{(m)} &= \int_0^1 \left(\boldsymbol{X}^T \beta^{(m)}(p) - \mu_{\boldsymbol{X}}^{(m)}\right)^4 / \sigma_{\boldsymbol{X}}^{4(m)} dp\,.
\end{aligned} \tag{9}$$

The conditional expectations of other basic statistics are similarly derived. We can construct posterior probability scores to assess differences of moments between groups or specific levels of continuous covariates as follows. For each posterior sample, we compute the appropriate moment from the formulas in (9) for two covariate levels, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and compute the difference, e.g. for the mean $\Delta_m = \mu_{1m} - \mu_{2m}$. Then, we define the posterior probability score for the comparison as:

$$P_{\mu_1 - \mu_2} = 2\min\{M^{-1} \sum_{m=1}^{M} I(\Delta_m > 0)\}, M^{-1} \sum_{m=1}^{M} I(\Delta_m < 0)\}$$

In assessing a dichotomous covariate x_a, we compare $x_a = 0$ and $x_a = 1$ while holding all other covariates at the mean, while when assessing a continuous covariate we compute differences for two extreme values of x_a, with the corresponding probability scores for the respective moments denoted $P_{a,\mu}$, $P_{a,\sigma}$, $P_{a,\xi}$, or $P_{a,\varphi}$.

Summarizing Gaussianity: As mentioned above, the first two quantlets form a complete basis for the space of Gaussian quantile functions, so by comparing the first two coefficients to the remainder one can obtain a rough measure of "Gaussianity" of the predicted distribution for a given set of covariates **X**. One measure that can be computed is $\sum_{k=1}^{2}\left(\mathbf{X}\hat{\beta}_{ak}\right)^2 / \sum_{k=1}^{K}\left(\mathbf{X}\hat{\beta}_{ak}\right)^2$, which will be on [0, 1], with a value of 1 precisely when the predicted quantile function is completely determined by the first two (Gaussian) bases and smaller scores indicating greater degrees of non-Guassianity.

Predicted PDF and CDF: To some researchers, distribution functions or probability density functions are more intuitive than quantile functions, and given their one-to-one relationship, it is possible to construct CDF or PDFs from the posterior samples as follows. CDFs can be constructed by simply plotting p vs. $\mathrm{E}\left\{\hat{Q}(p)|\mathbf{X},\mathbf{Y}\right\}$, and given posterior samples of the predicted quantile functions on an equally spaced grid $0 < p_1, \ldots, p_J < 1$, one can estimate predicted pdf for a set of covariates as described in Section 1.3 of the supplement.

Following is our recommended sequence of Bayesian inferential procedures.

1. Compute the global Bayesian p-value $P_{a,Bayes}$ for each predictor or contrast.
2. For any covariates for which $P_{a,Bayes} < \alpha$, characterize the differences:

   2a. Flag which probability grid points p are different using $P_{SimBas}(p) < \alpha$.

   2b. Compute moments; assess which moments differ according to the covariates.

   2c. Assess whether the degree of Gaussianity appears to differ across covariates.
3. If desired, compute the predicted densities or CDFs for any set of covariates.

3. SIMULATION STUDY

We conducted a simulation study to evaluate the performance of the quantile functional modeling framework and the use of quantlet basis functions. We generated random samples for four groups of subjects whose mean quantile function was assumed to be from a skew normal distribution

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\eta}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\eta}{\omega}\right)\right) \tag{10}$$

with the respective values of (η, ω, α) being (1, 5, 0), (3, 5, 0), (1, 6.5, 0), and (9.11, 7.89, −4), which correspond to a $N(1, 5)$, $N(3, 5)$, $N(1, 6.5)$, and a skewed normal with mean 1, variance 5, and skewness −0.78 denoted by $SN(1, 5, -0.78)$. Panels A and E of Figure 5

below show the densities and quantile functions, respectively, corresponding to these distributions.

For each group $j = 1, \ldots, 4$, we generated the random process $Q_{ij}(p)$ for $i = 1, \ldots n$ subjects, taking 1024 samples from the corresponding skewed normal distribution, with $p \in \mathcal{P} = [1/1025, \ldots, 1024/1025]$, and some correlated noise $\epsilon_{ij}(p)$ added to allow some random biological variability in the individual subjects' distributions. That is, $Y_{ij}(p) = \beta_j(p) + \epsilon_{ij}(p)$, where $\epsilon_{ij}(p)$ follows an Ornstein-Uhlenbeck process such that $\mathrm{Cov}(\epsilon_{ij}(p), \epsilon_{ij}(p')) = 0.9^{|p-p'|}$.

After constructing the empirical quantile function $Q_{ij}(p)$ by reordering $Y_{ij}(p)$ in p, the quantile functional regression model we fit to these data was

$$Q_{ij}(p) = \sum_{a=1}^{4} X_{ija}\beta_a(p) + \epsilon_{ij}(p), \tag{11}$$

with covariates defined such that $X_{ij1} = 1$ is for the intercept and $X_{ija} = \delta_{j=a}$ for $a = 2, 3, 4$ group indicators for groups 2–4. Note that with this parameterization, the means of the four groups are, respectively, $\beta_1(p)$, $\beta_1(p) + \beta_2(p)$, $\beta_1(p) + \beta_3(p)$, and $\beta_1(p) + \beta_4(p)$, and by construction $\beta_2(p)$ represents a location offset, $\beta_3(p)$ a scale offset, and $\beta_4(p)$ a skewness offset. Panel E of Figure 5 displays the true mean quantiles for each group and panel F the true values for these quantile functional regression coefficients.

We constructed a quantlet basis set for this data set as described above, with some results summarized in panels B, C, and D of Figure 5. The union set $\mathcal{D}^U = \cup_{i=1}^{n} \mathcal{D}_i$ included 2, 868 basis functions, and we chose a *common* set, $\mathcal{D}^{\mathcal{C}}$, that retained 10 basis functions, which resulted in a near-lossless basis set with $\rho^0 = 0.997$ (see $K_{\mathcal{C}} = 10$ in panel B). After orthogonalization, denoising, and re-standardization, the set of quantlets had sparsity properties similar to principal components (see panel C), and the fitted *quantlet* projection almost perfectly coincided with the observed data for all of the empirical quantile functions (panel D). Supplemental Figure 4 contains a plot of these 10 quantlet basis functions.

We applied several different approaches to these data: (A) naive quantile regression method (separate classical quantile regressions for each p by using *rq* function in *quantreg* R package (Koenker 2005)), (B) naive quantile functional regression approach (separate functional regressions for each subject-specific quantile p), (C) principal components method (quantile functional regression using PCs as basis functions), (D) *quantlet* without sparse regularization, (E) *quantlet* with sparse regularization, and (F) Gaussian model (quantlet approach but keeping only the first two coefficients). The naive quantile regression method (A) ignores all intrasubject correlation in the data and estimates the *population* quantile conditional on covariates, not the *subject-specific* quantile conditional on covariates desired in this quantile functional regression setting, but it is included here since it is an approach some researchers might try in this setting. In each case, the MCMC was run for 2, 000 iterations, keeping every one after a burn-in of 200. The results are shown in Supplementary Figure 8. We compared the methods in terms of the area within the joint

credible region and the corresponding integrated coverage rate, defined respectively as $\mathscr{A}(a) = \int_0^1 |J_{a,\alpha}^{upper}(p) - J_{a,\alpha}^{lower}(p)|^2 dp$ and $\mathscr{C}(a) = \int_0^1 I\left(J^{lower}(p) \leq \beta_a(p) \leq J_{a,\alpha}^{upper}(p)\right) dp$, where $J_{a,\alpha}^{upper}(p)$ and $J_{a,\alpha}^{lower}(p)$ are the upper and lower joint credible bands, respectively.

To investigate the degree of monotonicity afforded by the model, we constructed predicted quantile functions for a broad range of covariate values, and computed the degree of ϵ-monotonicity, defined to be $P_\epsilon^{\mathscr{M}}(X) = \int_0^1 I\left[\widehat{Q}(p|X) - \max_{p' < p}\left\{\widehat{Q}(p|X)\right\} > \epsilon\right] dp$ for some ϵ considered negligibly small in the context of the scale of Y in the current data set. We report the empirical rates of the ϵ-monotonicity as $1 - n^{-1}\sum_{i=1}^n P_\epsilon^{\mathscr{M}}(X_i)$. This empirical summary measure can be used to assess if a given model produces predictors with significant non-monotonicities across p or not.

Table 2 reports $\mathscr{A}(a)$ and $\mathscr{C}(a)$ for all quantile functional coefficients. Methods A-E all had good coverage properties, but use of the basis functions in modeling (C, D, E) clearly led to tighter joint credible bands than the naive quantile regression and naive quantile function regression methods that did not borrow strength across p, as expected, and the use of sparse regularization (E) led to tighter bands than the quantlet method with no shrinkage (D). Supplementary Figure 8 demonstrates the wiggliness and extremely wide joint credible bands of the naive methods. Note also that for the coefficient with significant skewness $\beta_4(p)$, the Gaussian model (F) had extremely poor coverage, while for the coefficients corresponding to the Gaussian groups, the quantlet model (E) had performance no worse than the Gaussian method. This is encouraging, suggesting that when the quantile functions are Gaussian there is not much loss of efficiency from using a richer quantlet basis set.

Supplementary Figure 10 depicts the simultaneous band scores $P_{SimBaS}(p)$ for the two contrast functions associated with the scale effect $\beta_3(p)$ and skewness effect $\beta_4(p)$, with regions of p for which $P_{SimBaS}(p) < 0.05$ are flagged as significantly different. As seen in Supplementary Figure 10, we expect to flag the tails in the scale effect and a broad region in the middle and in the extreme tails for the skewness effect. Note how the quantlet method with sparse regularization (E) flagged a larger set of regions than the other approaches, especially (B). In all cases, the global adjusted Bayesian p-values $P_{Bayes} = \min\{P_{map}(p)\}$ were less than 0.0005; hence, the null hypothesis $\beta_a(p) \equiv 0$ was rejected in all models.

We computed posterior probability scores to compare the mean, standard deviation, and skewness for each pair of distributions (Table 3), and Supplemental Table 2 contains the posterior means and credible intervals for each summary. We see that the basis function methods (C-E) all flagged the correct differences, while the naive quantile functional regression approach (B) had major type I error problems in the moment tests and the Gaussian method (F) unsurprisingly was unable to detect differences in skewness. As an additional comparison, we also applied the so-called *feature extraction* approach (G), which involved first computing the moments from the set of values for each subject and then performing statistical test comparing these across the groups. Encouragingly, we found these results were near identical to those found using our quantile functional regression with quantlets(E), suggesting that our unified functional modeling approach does not lose power

relative to feature extraction approaches when the distributional differences are indeed contained in the moments.

Constructing predicted quantile functions for a wide range of predictors and assessing ϵ-monotonicity, we found that the all predicted quantile functions from the quantlet-based methods were monotone, while the naive quantile functional regression method had ϵ-monotonicity of 25.8% and 96.8% for $\epsilon = 0.001$ and 0.01, respectively, demonstrating that quantlet basis functions encouraged the predicted quantile functions to be monotone in p.

4. QUANTILE FUNCTIONAL REGRESSION ANALYSIS OF GBM DATA

In our GBM case study, radiologic images consisting of pre-surgical T1-weighted post contrast MRI sequences from 64 patients were obtained from the Cancer Imaging Archive (cancerimagingarchive.net), along with measurements of certain covariates, including sex (21 females, 43 males), age (mean 56.5 years), DDIT3 gene mutation (6 yes, 58 no), EGFR gene mutation (24 yes, 40 no), GBM subtype (30 mesenchymal, 34 other), and survival status (25 less than 12 months, 39 greater than or equal to 12 months), where Tutt (2011) has pointed out that most people diagnosed with GBM survive only 12 to 15 months, so that we followed this and used 12 moments as the cutoff in our context. This cut-off is commonly referred to as an extreme discordant phenotype design (Nebert 2000) and is a well-established grouping to enhance signals relevant to survival (Tyekucheva, Marchionni, Karchin and Parmigiani 2011).

Following Saha, Banerjee, Kurtek, Narang, Lee, Rao, Martinez, Bharath, Rao and Baladandayuthapani (2016), registration and inhomogeneity correction were conducted using Medical Image Processing and Visualization (MIPAV) software. Inhomogeneity correction known as nonparametric, nonuniform intensity normalization (N3) correction was conducted to remove the shading artifacts in MRI scans. Then, tumors were segmented in 3-D by clinical experts using the Medical Image Interaction Toolkit. Images and their 3-D tumor masks were subsequently re-sliced for isotropic pixel resolution using the NIFTI toolbox in MATLAB. From these re-sliced images, the slice with largest tumor area in the T1-post contrast image was selected as the Regions of Interest (ROI) for analysis. We extracted the set of m_i pixel intensities within the ROI for each patient $i = 1, \ldots, n = 64$, where the number of pixels within the tumor ranged from 371 to 3421.

Model:.

We sorted the pixel intensities for each patient, yielding an empirical quantile function $Q_i(p_{ij})$ on a grid of observational points $p_{ij} = j/(m_i + 1)$, $j = 1, \ldots, m_i$. We related these to the clinical, demographic, and genetic covariates using the following quantile functional regression model:

$$\begin{aligned} Q_i(p) = {} & \beta_{\text{overall}}(p) + x_{\text{sex},i}\beta_{\text{sex}}(p) + x_{\text{age},i}\beta_{\text{age}}(p) + x_{\text{DDIT3},i}\beta_{\text{DDIT3}}(p) \\ & + x_{\text{EGFR},i}\beta_{\text{EGFR}}(p) + x_{\text{Mesenchymal},i}\beta_{\text{Mesenchymal}}(p) + x_{\text{survival},i}\beta_{\text{survival}\ 12}(p) \\ & + E_i(p). \end{aligned} \tag{12}$$

We constructed quantlets for these data using the procedure described in Section 2.3. After the first step, we were left with a union basis set $\mathscr{D}^U$ containing 546 basis functions. The first panel of Figure 3 plots the near-losslessness parameters ρ_0 and $\bar{\rho}$ against the number of basis coefficients $K_{\mathscr{C}}$ in the reduced set. Based on this, we selected the combined basis set $\mathscr{D}_{\mathscr{C}}$ for $\mathscr{C} = 10$, which contained $K_{\mathscr{C}} = 27$ basis functions and was near-lossless, with $\rho^0 = 0.990$ and $\bar{\rho} = 0.998$. We then orthogonalized, denoised, and re-standardized the resulting basis to yield the set of quantlets, the first 16 of which are plotted in Figure 4. As shown in panel 2 of Figure 3, these quantlets yielded a basis with similar sparsity property as principal components computed from the empirical quantile functions.

After computing the quantlet coefficients for each subject's empirical quantile function, we fit the quantlet-space version of model (12) as described above, obtaining 2, 000 posterior samples after a burn-in of 200, after which the results were projected back to the original quantile space to yield posterior samples of the functional regression parameters in model(12). MCMC convergence diagnostics were computed, and suggested that the chain mixed well (Supplementary Figure 17). From these, we constructed 95% point wise and joint credible bands for each $\beta_a(p)$ and computed the corresponding simultaneous band scores $P_{a,SimBaS}(p)$ and global Bayesian p-values $P_{a,Bayes}$ as described in Section 2.6.

Results:.

Figure 6 summarizes the estimation and inference for each of the covariates in the model. For each covariate there is one panel presenting the functional predictor $\beta_a(p)$ along with the point wise (grey) and joint (black) credible bands, and an indicator of which p are flagged such that $\beta_a(p) \neq 0$ (orange lines indicating $P_{a,SimBaS}(p) < 0.05$). The other panel contains density estimates for each covariate level (holding all others at the mean), computed as outlined in the supplementary materials, along with posterior probability scores summarizing whether the mean, variance or skewness appeared to di er across these groups. Supplementary Table 3 contains measures of the relative Gaussianness of the distributions for the various groups along with 95% credible intervals.

The global Bayesian p-values for testing $\beta_a(p) \equiv 0$ for each covariate are in the corresponding figure panel headers, and reveal that for sex (p=0.016) and DDIT3 (p=0.012), the functional covariates are flagged as significant, and for the mesenchymal subtype (p=0.087) and survival (p=0.067) endpoints, there was some indication of a possible trend. We see that for sex, there was evidence of a mean shift (p=0.004) with females tending to have higher pixel intensities than males, especially in the upper tails of the distribution, and the female distribution appearing to be slightly more Gaussian than the males. For DDIT3, we see evidence of a mean and variance shift, with tumors with DDIT3 mutation tending to have higher intensities and greater variability than those without, especially in the upper tail of the distribution. The mesenchymal subtype, while not flagged as statistically significant in the global test, shows some tendency for a mean shift with the mesenchymal subtype tending to have higher distributional values and perhaps slightly more non-Gaussian characteristics. Follow-up studies can assess the significance of this upward shift in distribution of pixel intensities for female patients and DDIT3 mutated tumors.

One cause of higher pixel intensities in MRI images of tumors is greater accumulation of fluid in body tissues, called edema, which can be an indicator of poor prognosis (Zinn, Majadan, Sathyan, Singh, Majumder, Jolesz and Colen 2011). Thus, it may be true that female patients and patients with DDIT3 mutations have tumors with greater edema, which is plausible given results in the literature showing DDIT3 mutation is associated with shorter survival time (Saha et al. 2016). Follow-up studies can assess the significance of this upward shift in distribution of pixel intensities (or the extent of tumor vascularisation) for female patients and DDIT3 mutated tumors. Since the gender has the specific effect on GBM (Colen, Wang, Singh, Gutman and Zinn 2014) and DDIT3 also plays a key role in resistance to therapy, due to its hypoxia-related activity (Ragel, Couldwell, Gillespie and Jensen 2007), and in GBM tumorogenesis (Ping, Deng, Wang, Zhang, Zhang, Xu, Zhao, Fan, Yu, Xiao et al. 2015), where DDIT3 is a p53 driven gene (Tivnan 2016) suggesting that this radiographic observation might be associated with p53 associated cell death (showing as lower T2, FLAIR or T1c signal), our findings have a strong connection with the results in the existing literature. In addition, we notice that the longer survival time tends to be shifted to the right, representing higher intensity and higher vascularisation. As pointed out by Gilbert (2016), one of the only effective therapeutic strategies for GBM is antiangiogeneic therapies, and it would make sense that patients with greater baseline vasculature would be more likely to respond to therapy, and thus experience improved survival times.

Sensitivity Analysis and Comparison:

Our results are presented for $K = 27$ basis functions, but to assess sensitivity to choice of K we also ran our model for a wide range of possible values of K, with Supplementary Table 5 showing global Bayesian p-values for the entire range of potential values for K (from 546 to 2), along with run time. The run time tracks linearly with K. Note that we get the same substantive results over the range of basis sizes, so results are quite robust to choice of number of quantlets. However, keeping more quantlets than necessary clearly adds to the uncertainty of parameter estimates, as indicated by the larger joint band widths. Also, keeping too few basis functions can lead to some missed results and also wider joint band widths. Moderate basis sets that are as parsimonious as possible while retaining the near-lossless property seem to give the tightest credible bands and thus the greatest power for global and local tests. We also performed a sensitivity analysis on the parameter ν_0 (inverse gamma prior) indicating the prior strength for the variance components and found that results for slightly larger or smaller values yielded nearly identical results. We lastly conducted a sensitivity analysis for lasso to see how selection of more or fewer dictionary elements via larger or smaller lasso parameters effects the ultimate number of quantlets. Choice of greater or fewer dictionary elements via larger or smaller lasso parameters still resulted in sparse sets of quantlet basis functions using the near-lossless criterion as can be seen in Figures 23 in Supplementary material. Also, from Figures 11, 21 and 22 in Supplementary material, we see that there are not dramatic changes on the final results.

To compare different methods with our quantlet with sparse regularization approach, we also applied to these data a quantlet approach with no sparse regularization and a naive quantile functional regression method modeling independently for each p (after interpolating onto a common grid). Posterior mean estimates, credible intervals, and other inferential summaries

are given in Supplementary Figure 11 and Table 6. Note that the quantlets method with sparse regularization tends to yield estimates that are smoother and with tighter joint credible bands than either the naive or the quantlets-no sparse regularization runs. As we can see in Figure 7 the differences between the quantlet and naive methods are substantial, and demonstrate the significant power gained by borrowing strength across p using the quantlet-based modeling approach. The completely naive quantile functional regression approach gave nonsensical results for this application (Supplementary Figure 19).

Supplementary Figure 18 contains the predicted quantiles functions over a grid of covariate combinations for this model. Although the quantile functional regression using quantlets does not explicitly impose monotonocity in the predicted quantile functions, we see that the predicted quantile functions are all monotone non-decreasing. See Section 4 of the supplement for further details and discussion of monotonicity issues.

Table 4 contains posterior probability scores assessing differences in moments for these three methods, plus a feature extraction approach in which moments were first calculated from each subject's samples and then statistically compared with a Bayesian regression fit. As in the simulations, we see that the naive quantile functional regression method appears to have type I error problems in the mean and variance. While our method (E) does not yield additional power when the distributional differences are captured by the moments (G), our approach does not give much power away in these settings and yet can detect distributional differences that are not contained in the moments, e.g. differences in specific extreme quantiles. Specifically, by the estimation and inference of our method, we can thoroughly understand the pixel intensity distribution for each of the covariates. For instance, for the male, (E) provide the insight that males have lower pixel intensities than females because the male effect, $B(p)$ along with the point wise and joint credible band has a decreasing tendency from the first panel of Figure 6.

5. DISCUSSION

In this paper, motivated by a clinical imaging application in cancer, we have introduced a strategy for regressing the distribution of repeated samples for a subject on a set of covariates through a model we call *quantile functional regression*. We distinguish this model from other types of quantile regression and functional regression methods in existing literature, in that it is regressing the *subject-specific* quantile, not the *population-level* quantile, on covariates, and accounts for intrasubject correlation. We describe how it serves as a middle ground between two commonly-used strategies of (1) performing a series of regressions on arbitrary summaries of the distribution such as mean or standard deviation and (2) independent regression models for each quantile p in a chosen set. Our approach models a subject's entire quantile function as a functional response, building in dependency across p in the mean and covariance using custom basis functions called *quantlets* that are empirically defined, near-lossless, regularized, sparse, and with some of the individual bases being interpretable. These basis functions have sparsity properties similar to principal components, but appear more regular and interpretable. They provide a flexible representation of the underlying quantile functions while containing a sufficient Gaussian basis as a subspace. The *quantlets* basis function that is successfully utilized to capture

distinct characteristics of the quantile function, consisting of the subspace spanned by the normal quantile and the subspace spanned by the mixture beta distributions. These quantlets are constructed based on a dictionary of Beta CDF, which can be shown to be sufficient for representing any quantile function with a first derivative that is uniformly continuous, and has numerous useful statistical properties including near-lossless representation, sparsity, regularity, and some interpretability.

We fit the quantile functional regression model using a Bayesian approach with sparse regularization priors on the quantlet space regression coefficients that smooths the regression coefficients and yields a broad array of Bayesian inferential summaries computable from the posterior samples of the MCMC procedure. For example, we can construct global tests of significance for each covariate using global Bayesian p-values, and then characterize these differences by flagging regions of p while adjusting for multiple testing, and obtaining probability scores for any moments or other summaries of the distributions.

In this paper, we have presented the quantile functional regression framework using a standard linear model with scalar covariates and independent Gaussian residual error functions, but as in other functional regression contexts the model can be extended to include other complex structures that extend the usability of the modeling framework. This includes functional covariates, nonparametric effects in the covariates x_{ia}, random effects and/or spatially/temporally correlated residual errors to accommodate correlation between subjects induced by the experimental design, and the ability to perform robust quantile functional regression to downweight outlying samples using heavier-tailed likelihoods. These types of flexible modeling components are available as part of the Bayesian functional mixed model (BayesFMM) framework that has been developed in recent years (Morris and Carroll 2006; Zhu, Brown and Morris 2011; Zhu, Brown and Morris 2012; Meyer et al. 2015; Zhang, Baladandayuthapani, Zhu, Baggerly, Majewski, Czerniak and Morris 2016; Zhu, Versace, Cinciripini and Morris 2018; Lee, Miranda, Baladandayuthapani, Rausch, Fazio, Downs and Morris 2018). By linking the software developed here to generate the quantlets and fit quantile functional regression models with the BayesFMM software, it will be possible to extend the quantile functional regression framework to these settings and thus analyze an even broader array of complex data sets generated by modern research tools.

Our approach has been designed with relatively high dimensional data in mind, i.e. data for which there are at least a moderately large number of observations per subject (at least 50 or 100). We are currently working on extensions of this method to handle lower dimensional data with fewer observations per subject, which requires a careful propagation of uncertainty in the estimators of the empirical quantile functions into the quantile functional regression. This propagation of uncertainty could also be done in larger sample cases like the one presented here, but given the substantial complexity and length already in this paper we leave this for future work. As mentioned in Section 5, we are currently working on extensions of this method to handle the empirical quantile estimator established by fewer/massive observations because of the different tumor size and the imperfection of the image segmentation per subject and leave this for future work in this paper

Also, in settings with enormous numbers of observations per subjects, e.g. millions to billions or more, the procedure described in this paper to construct the quantlets basis would be too computationally burdensome. Given that in those settings, it is unlikely that so many observations are needed to quantify the subject-specific quantile function, we have worked out algorithms to down-sample the empirical quantile functions in these cases in a way that engenders computational feasibility but is still near-lossless. This also will be reported in future work. Other data have measurements on many 1000s to 100,000s of subjects, which can be accommodated by computational adjustments of the procedure reported herein, but again we leave this for future work. In this paper, we focused on absolutely continuous random variables that have no jumps in the quantile functions. It is also possible to adapt our quantlet construction procedure to allow jumps at a discrete set of values, thus accommodating discrete valued random variables, but again this extension will be left for future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

REFERENCES

Bellemare MG, Dabney W, and Munos R (2017), "A distributional perspective on reinforcement learning," arXiv preprint arXiv:1707.06887, pp. 1–19.

Bickel PJ, and Freedman DA (1981), "Some asymptotic theory for the bootstrap," The Annals of Statistics, 9, 1196–1217.

Brockhaus S, and Rügamer D (2015), "FDboost: boosting functional regression models," R package version 0.0–8,.

Brockhaus S, Scheipl F, Hothorn T, and Greven S (2015), "The functional linear array model," Statistical Modelling, 15, 279–300.

Cardot H, Crambes C, and Sarda P (2005), "Quantile regression when the covariates are functions," Nonparametric Statistics, 17, 841–856.

Chen K, and Müller H-G (2012), "Conditional quantile analysis when covariates are functions, with application to growth data," Journal of the Royal Statistical Society: Series B, 74, 67–89.

Colen RR, Wang J, Singh SK, Gutman DA, and Zinn PO (2014), "Glioblastoma: imaging genomic mapping reveals sex-specific oncogenic associations of cell death," Radiology, 275, 215–227. [PubMed: 25490189]

Davino C, Furno M, and Vistocco D (2013), Quantile regression: theory and applications Wiley New York.

Dobrushin RL (1970), "Definition of random variables by conditional distributions," Teoriya Veroyatnostei i Primeneniya, 15, 469–497.

Donoho DL, Johnstone IM, Kerkyacharian G, and Picard D (1995), "Wavelet shrink age: asymptopia?," Journal of the Royal Statistical Society. Series B, 57, 301–369.

Dunson DB (2006), "Bayesian dynamic modeling of latent trait distributions," Biostatistics, 7, 551–568. [PubMed: 16488893]

Dunson DB, Pillai N, and Park J-H (2007), "Bayesian density regression," Journal of the Royal Statistical Society: Series B, 69, 163–183.

Faraway JJ (1997), "Regression analysis for a functional response," Technometrics, 39, 254–261.

Felipe De Sousa EM, Vermeulen L, Fessler E, and Medema JP (2013), "Cancer heterogeneity - a multifaceted view," EMBO reports, 14, 686–695. [PubMed: 23846313]

Ferraty F, Rabhi A, and Vieu P (2005), "Conditional quantiles for dependent functional data with application to the climatic" el niño" phenomenon," Sankhyā: The Indian Journal of Statistics, 67, 378–398.

Gilbert MR (2016), "Antiangiogenic therapy for glioblastoma: complex biology and complicated results," American Society of Clinical Oncology, 34, 1567–1570.

Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, and Reich D (2012), "Penalized functional regression," Journal of Computational and Graphical Statistics, 20, 830–851.

Goldsmith J, Wand MP, and Crainiceanu C (2011), "Functional regression via variational Bayes," Electronic journal of statistics, 5, 507–602.

Griffin JE, and Steel MJ (2006), "Order-based dependent Dirichlet processes," Journal of the American statistical Association, 101, 179–194.

Guo W (2002), "Functional mixed effects models," Biometrics, 58, 121–128. [PubMed: 11890306]

Hao L, and Naiman DQ (2007), Quantile regression Sage London.

He X, and Liang H (2000), "Quantile regression estimates for a class of linear and partially linear errors-in-variables models," Statistica Sinica, 10, 129–140.

Just N (2014), "Improving tumour heterogeneity MRI assessment with histograms," British journal of cancer, 111, 2205–2213. [PubMed: 25268373]

Kato K (2012), "Estimation in functional linear quantile regression," The Annals of Statistics, 6, 3108–3136.

Kato K, Galvao AF, and Montes-Rojas GV (2012), "Asymptotics for panel quantile regression models with individual effects," Journal of Econometrics, 170, 76–91.

Koenker R (2004), "Quantile regression for longitudinal data," Journal of Multivariate Analysis, 91, 74–89.

Koenker R (2005), Quantile regression Cambridge university press.

Lee W, Miranda M, Baladandayuthapani V, Rausch P, Fazio M, Downs C, and Morris JS (2018), "Bayesian semiparametric functional mixed models for longitudinal functional data with application to glaucoma data," Journal of the American Statistical Association to appear, doi:10.1080/01621459.2018.1476242.

Li M, Wang K, Maity A, and Staicu A-M (2016), "Inference in Functional Linear Quantile Regression," arXiv preprint arXiv:1602.08793,.

Liu Y, Li M, and Morris JS (2018), Function-on-scalar quantile regression with application to mass spectrometry proteomics data,, Technical report, MD Anderson.

MacEachern SN (1999), "Dependent nonparametric processes," ASA proceedings of the section on Bayesian statistical science, 1, 50–55.

Marusyk A, Almendro V, and Polyak K (2012), "Intra-tumour heterogeneity: a looking glass for cancer?," Nature Reviews Cancer, 12, 323–334. [PubMed: 22513401]

Meyer MJ, Coull BA, Versace F, Cinciripini P, and Morris JS (2015), "Bayesian function-on-function regression for multilevel functional data," Biometrics, 71, 563–574. [PubMed: 25787146]

Morris JS (2015), "Functional Regression," Annual Review of Statistics and Its Application, 2, 321–359.

Morris JS, and Carroll RJ (2006), "Wavelet-based functional mixed models," Journal of the Royal Statistical Society. Series B, 68, 179–199.

Muller P, Erkanli A, and West M (1996), "Bayesian curve fitting using multivariate normal mixtures," Biometrika, 83, 67–79.

Nebert DW (2000), "Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics," European journal of pharmacology, 410, 107–120. [PubMed: 11134663]

Parzen E (2004), "Quantile probability and statistical data modeling," Statistical Science, 19, 652–662.

Ping Y, Deng Y, Wang L, Zhang H, Zhang Y, Xu C, Zhao H, Fan H, Yu F, Xiao Y et al. (2015), "Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multidimensional genomic data," Nucleic acids research, 43, 1997–2007. [PubMed: 25653168]

Ragel BT, Couldwell WT, Gillespie DL, and Jensen RL (2007), "Identification of hypoxia-induced genes in a malignant glioma cell line (U-251) by cDNA microarray analysis," Neurosurgical review, 30, 181–187. [PubMed: 17486380]

Ramsay JO, and Silverman BW (2006), Functional data analysis New York: Springer.

Reich BJ (2012), "Spatiotemporal quantile regression for detecting distributional changes in environmental processes," Journal of the Royal Statistical Society: Series C, 61, 535–553.

Reich BJ, Fuentes M, and Dunson DB (2012), "Bayesian spatial quantile regression," Journal of the American Statistical Association, 106, 6–20.

Reiss PT, Huang L, and Mennes M (2010), "Fast function-on-scalar regression with penalized basis expansions," International Journal of Biostatistics, 6, 1–28.

Ruppert D, Wand MP, and Carroll RJ (2003), Semiparametric regression Cambridge university press.

Saha A, Banerjee S, Kurtek S, Narang S, Lee J, Rao G, Martinez J, Bharath K, Rao AU, and Baladandayuthapani V (2016), "DEMARCATE: Density-based Magnetic Resonance Image Clustering for Assessing Tumor Heterogeneity in Cancer,"NeuroImage: Clinical, 12, 132–143. [PubMed: 27408798]

Scheipl F, Staicu A-M, and Greven S (2015), "Functional additive mixed models," Journal of Computational and Graphical Statistics, 24, 477–501. [PubMed: 26347592]

Tibshirani R (1996), "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society. Series B, 58, 267–288.

Tivnan A (2016), Resistance to Targeted Therapies Against Adult Brain Cancers, International Publishing AG: Springer.

Tutt B (2011), "Glioblastoma Cure Remains Elusive Despite Treatment Advances," OncoLog, 56, 1–8.

Tyekucheva S, Marchionni L, Karchin R, and Parmigiani G (2011), "Integrating diverse genomic data using gene sets," Genome biology, 12, R105. [PubMed: 22018358]

Wu CO, and Chiang C-T (2000), "Kernel smoothing on varying coefficient models with longitudinal dependent variable," Statistica Sinica, 10, 433–456.

Yang Y, and He X (2015), "Quantile regression for spatially correlated data: an empirical likelihood approach," Statistica Sinica, 25, 261–274.

Yang Y, and Tokdar ST (2017), "Joint estimation of quantile planes over arbitrary predictor spaces," Journal of the American Statistical Association, pp. 1–14.

Zhang L, Baladandayuthapani V, Zhu H, Baggerly KA, Majewski T, Czerniak BA, and Morris JS (2016), "Functional CAR models for large spatially correlated functional datasets," Journal of the American Statistical Association, 111, 772–786. [PubMed: 28018013]

Zhu H, Brown PJ, and Morris JS (2011), "Robust, adaptive functional regression in functional mixed model framework," Journal of the American Statistical Association, 106, 1167–1179. [PubMed: 22308015]

Zhu H, Brown PJ, and Morris JS (2012), "Robust classification of functional and quantitative image data using functional mixed models," Biometrics, 68, 1260–1268. [PubMed: 22670567]

Zhu H, Versace F, Cinciripini P, and Morris JS (2018), "Robust functional mixed models for spatially correlated functional regression, with application to event-related potentials for nicotine-addicted individuals," Neuroimage, 181, 501–512. [PubMed: 30057352]

Zinn PO, Majadan B, Sathyan P, Singh SK, Majumder S, Jolesz FA, and Colen RR (2011), "Radiogenomic mapping of edema/cellular invasion MRI-phenotypes in glioblastoma multiforme," PloS one, 6, 1–11.

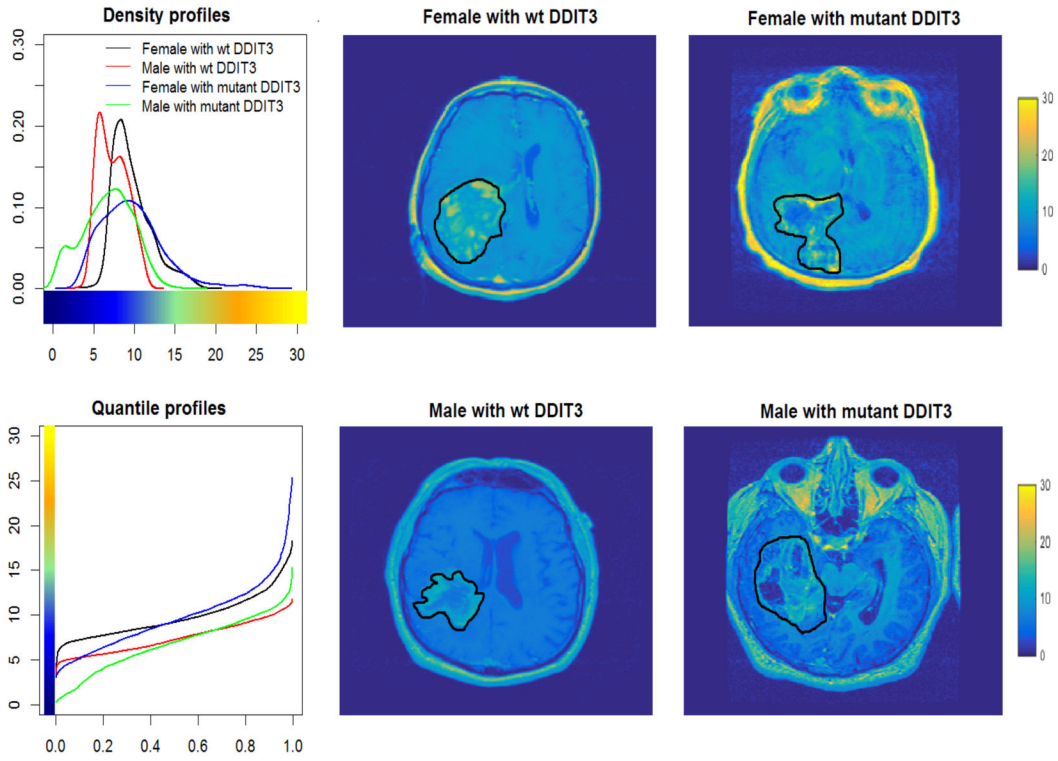

Figure 1:

Characterizing tumor heterogeneity from distributional summaries of MRI pixel intensities: the two graphs include kernel density estimates and the raw empirical quantile functions as representations of tumor heterogeneity (pixel intensities within the tumor); black line: female patient without DDIT3 mutation; red line: male patient without DDIT3 mutation; blue line: female patient with DDIT3 mutation; and green line: male patient with DDIT3 mutation. The images in other columns represent the T1-post contrast MRIs of the brains, with tumor boundaries indicated by black lines.

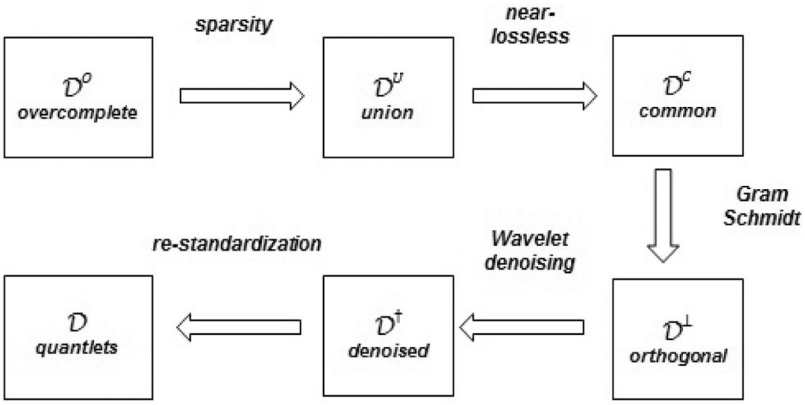

Figure 2:
Graphical illustration of the entire procedure for constructing the quantlets.

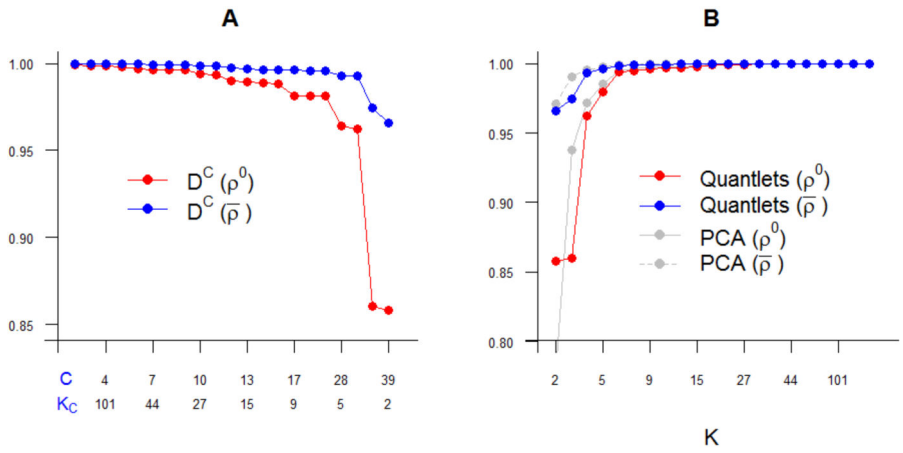

Figure 3:

Construction of Quantlet Bases. The concordance correlation for the GBM application: (A) minimum concordance (ρ^0, red) and average ($\bar{\rho}$, blue) across samples as function of $K_{\mathscr{C}}$, (B) ρ^0 and $\bar{\rho}$ for *quantlets* basis and principal components, varying with the number of basis coefficients.

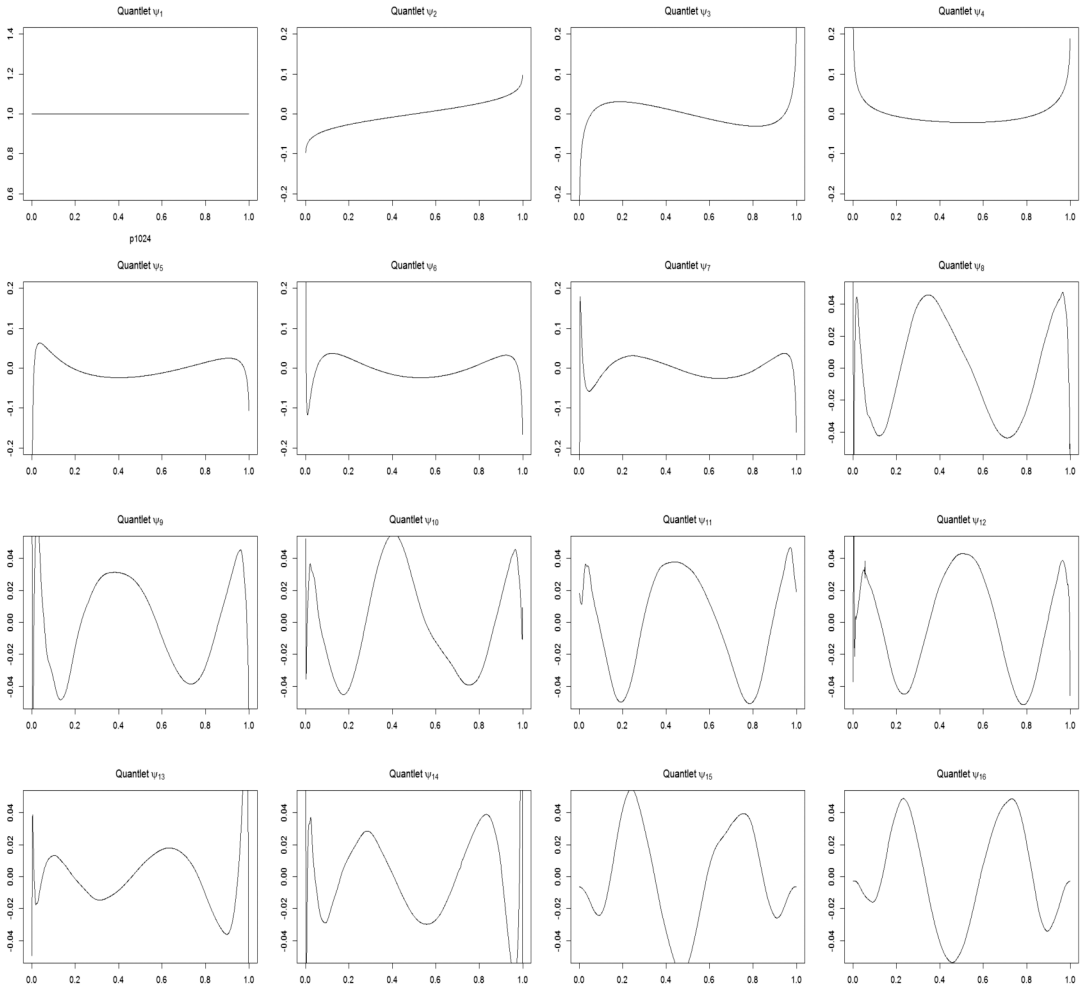

Figure 4:
First 16 quantlet basis functions for GBM data set.

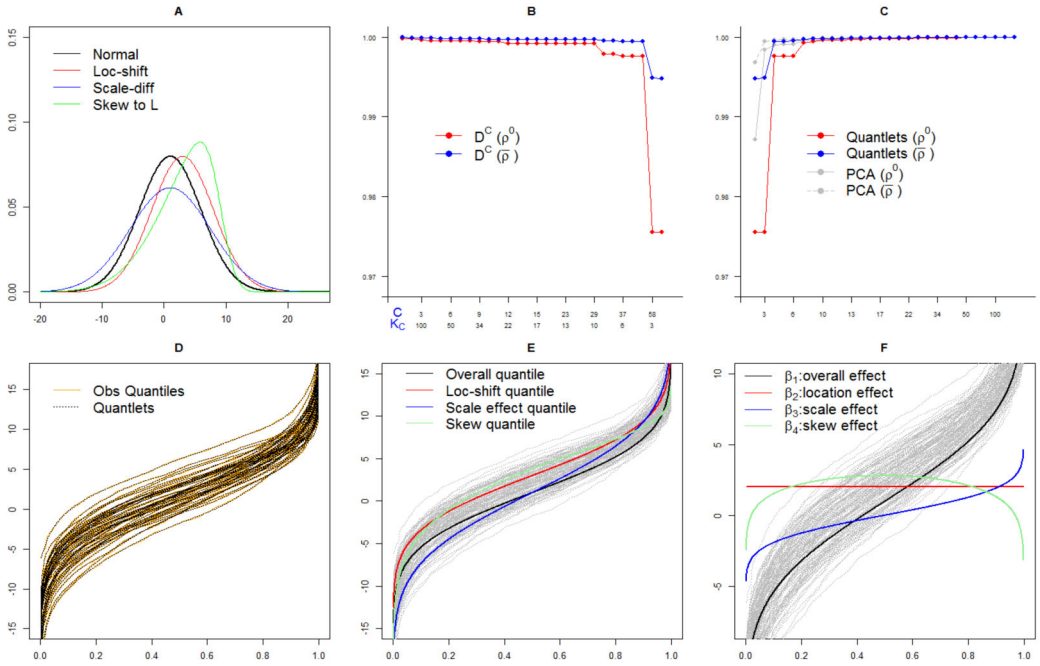

Figure 5:

Simulated data in the skewed normal scenario and their *quantlet* representations:(A) density functions of the population, (B) the near-lossless criterion varying with the different number of basis functions, (C) the concordance correlation varying with the cumulative number of the *quantlets*, and compared with principal components (D) the relation between empirical quantile functions and *quantlet* fits, (E) mean quantile functions by group and (F) quantile functional regression coefficients.

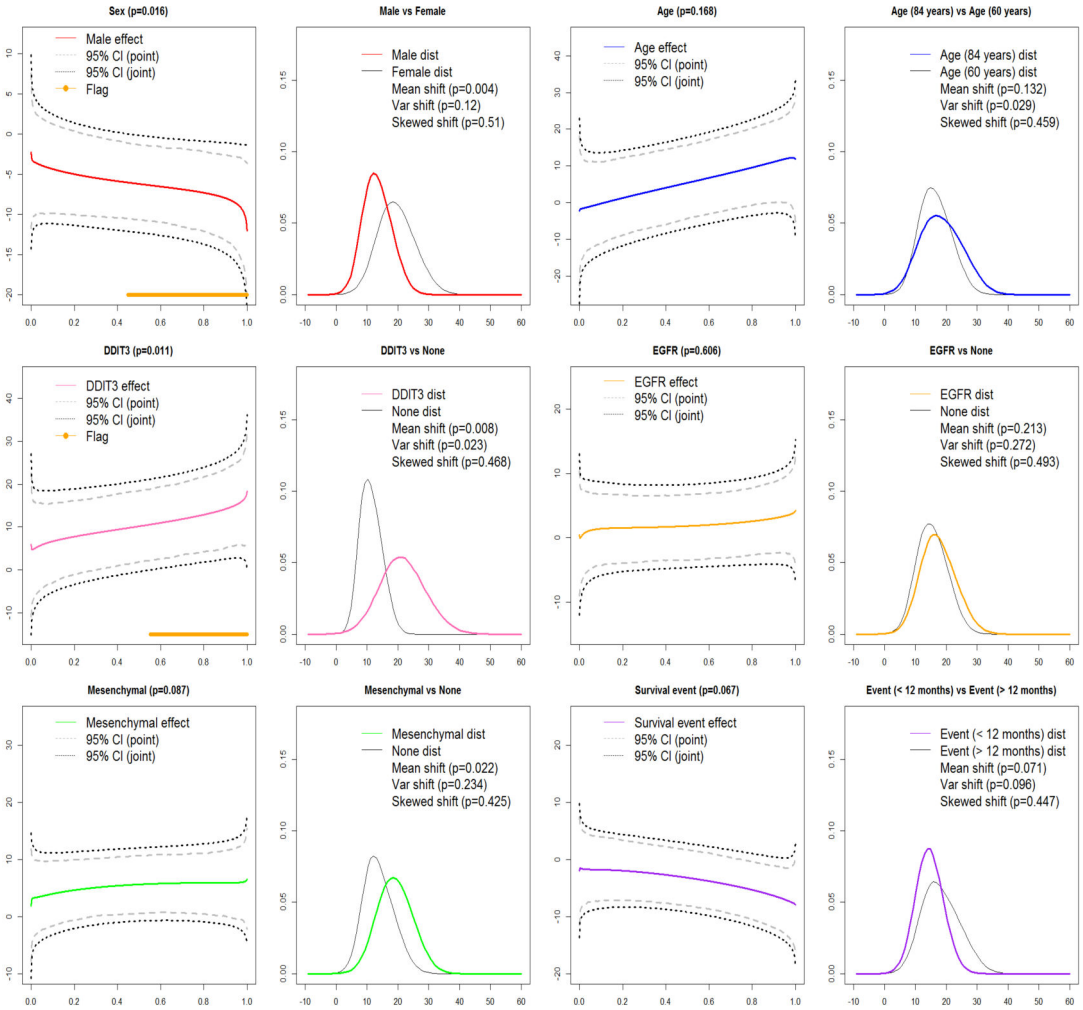

Figure 6:

Posterior inference for functional coefficients for T1-post contrast image: for each covariate (6), the left panel includes posterior mean estimate, point and joint credible bands, GBPV in heading along with SimBas less then .05 (orange line), and the right panel includes predicted densities for the two levels of the covariate along with the posterior probability scores for the moment different testings.

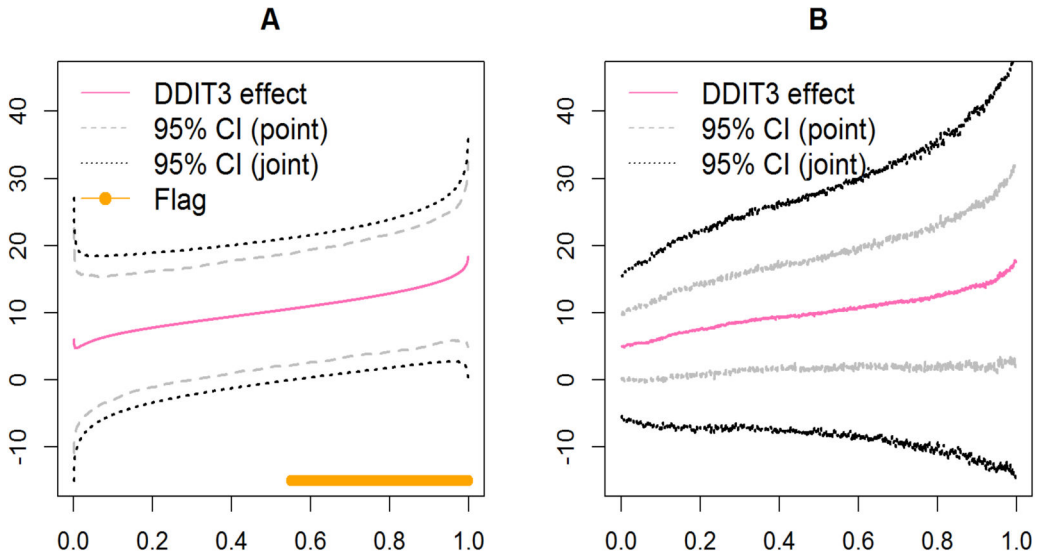

Figure 7:

Comparison between quantlet and naive approaches for DDIT3 status for (A) quantlet approach with sparse regularization and (B) the naive *one-p-at-a-time* quantile functional regression approach.

Table 1:

Types of regression based on response type and objective function.

Response (·)	Objective function $E((\cdot)\|X)$	Objective function $F^{-1}_{(\cdot)}(p\|X)$
scalar Y	classic regression	quantile regression
function $Y(t)$	functional regression	functional quantile regression
quantile function $Q(p)$	quantile functional regression*	quantile functional quantile regression

Table 2:

Results for Simulation 1: Area and coverage for the joint 95% confidence intervals: (A) naive quantile regression approach, (B) naive quantile functional regression approach, (C) principal component method, (D) *quantlet* space without sparse regularization,(E) *quantlet* space with sparse regularization, and (F) Gaussian *quantlet* space approach.

Type	A	B	C	D	E	F
$\beta_1(p)$	2.693 (1.000)	1.533 (1.000)	1.010 (0.999)	1.088 (0.999)	0.941 (1.000)	0.961 (1.000)
$\beta_2(p)$	3.998 (1.000)	2.160 (1.000)	1.454 (1.000)	1.533 (1.000)	1.360 (1.000)	1.392 (1.000)
$\beta_3(p)$	3.903 (1.000)	2.169 (1.000)	1.467 (1.000)	1.574 (1.000)	1.350 (1.000)	1.419 (1.000)
$\beta_4(p)$	3.751 (1.000)	2.186 (1.000)	1.441 (1.000)	1.515 (1.000)	1.359 (1.000)	1.369 (0.373)

Table 3:

Simulation 1: Testing for conditional moment statistics in simulation: (A) naive quantile regression approach, (B) naive quantile functional regression approach, (C) principal component method, (D) *quantlet* space without sparse regularization, (E) *quantlet* space with sparse regularization, (F) Gaussian *quantlet* space approach, and (G) feature extraction approach, where the values in this table are the posterior probability scores derived by its corresponding method for each test (the first column).

H_0	True	A	B	C	D	E	F	G
$\mu_1 = \mu_3$	$\mu_1 = \mu_3$	0.000	0.000	0.193	0.214	0.191	0.214	0.205
$\mu_2 = \mu_4$	$\mu_2 = \mu_4$	0.000	0.000	0.449	0.462	0.438	0.462	0.438
$\sigma_1 = \sigma_3$	$\sigma_1 \neq \sigma_3$	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\sigma_2 = \sigma_4$	$\sigma_2 = \sigma_4$	0.000	0.002	0.420	0.413	0.411	0.159	0.187
$\xi_1 = \xi_3$	$\xi_1 = \xi_3$	0.013	0.374	0.499	0.484	0.494	0.493	0.389
$\xi_2 = \xi_4$	$\xi_2 \neq \xi_4$	0.000	0.000	0.000	0.000	0.000	0.505	0.000

Table 4:

Posterior probability score of difference tests for the GBM data set: (B) naive quantile functional regression approach, (E) *quantlet* space with sparse regularization, and(G) feature extraction approach, where the values in this table are the posterior probability scores derived by its corresponding method for each different test between treatment and reference groups in the top row.

Test	$\mu_T = \mu_R$			$\sigma_T = \sigma_R$			$\xi_T = \xi_R$		
Method	**B**	**E**	**G**	**B**	**E**	**G**	**B**	**E**	**G**
Sex	0.000	0.004	0.028	0.000	0.120	0.064	0.346	0.510	0.547
Age	0.000	0.132	0.308	0.000	0.029	0.026	0.176	0.459	0.003
DDIT3	0.000	0.008	0.027	0.000	0.023	0.036	0.344	0.468	0.418
EGFR	0.000	0.213	0.453	0.000	0.272	0.403	0.368	0.493	0.467
Mesenchymal	0.000	0.022	0.040	0.000	0.234	0.433	0.071	0.425	0.191
Survival$_{12}$	0.000	0.071	0.160	0.000	0.096	0.034	0.312	0.447	0.969