scientific reports

Check for updates

OPEN

Additive quantile mixed effects modelling with application to longitudinal CD4 count data

Ashenafi A. Yirga[1]✉, Sileshi F. Melesse[1], Henry G. Mwambi[1] & Dawit G. Ayele[2]

Quantile regression offers an invaluable tool to discern effects that would be missed by other conventional regression models, which are solely based on modeling conditional mean. Quantile regression for mixed-effects models has become practical for longitudinal data analysis due to the recent computational advances and the ready availability of efficient linear programming algorithms. Recently, quantile regression has also been extended to additive mixed-effects models, providing an efficient and flexible framework for nonparametric as well as parametric longitudinal forms of data analysis focused on features of the outcome beyond its central tendency. This study applies the additive quantile mixed model to analyze the longitudinal CD4 count of HIV-infected patients enrolled in a follow-up study at the Centre of the AIDS Programme of Research in South Africa. The objective of the study is to justify how the procedure developed can obtain robust nonlinear and linear effects at different conditional distribution locations. With respect to time and baseline BMI effect, the study shows a significant nonlinear effect on CD4 count across all fitted quantiles. Furthermore, across all fitted quantiles, the effect of the parametric covariates of baseline viral load, place of residence, and the number of sexual partners was found to be major significant factors on the progression of patients' CD4 count who had been initiated on the Highly Active Antiretroviral Therapy study.

Abbreviations

AMM	Additive mixed model
QR	Quantile regression
AQM	Additive quantile model
AQMM	Additive quantile mixed model
GAMLSS	Generalized additive model for location, scale, and shape
CAPRISA	Centre of the AIDS Programme of Research in South Africa
HIV	Human immunodeficiency virus
AIDS	Acquired immune deficiency syndrome
CD4	Cluster of difference 4 cell (t-lymphocyte cell)
VL	Viral load refers to the number of HIV copies in a milliliter of blood (copies/ml)
STD	Sexually transmitted diseases
ART	Antiretroviral therapy
ARV	Antiretroviral (drug)
HAART	Highly active antiretroviral therapy
WHO	World Health Organization

Parametric models relate the mean of a response variable to a linear combination of covariate effects and focus on the response's average properties[1]. Nevertheless, there are inevitable occasions when such parametric models fail, and data analysis must turn to more flexible, nonparametric models[2]. Parametric models also assume a distribution for the outcome variable as opposed to purely nonparametric models. However, most of the vast literature on nonparametric regression also deals with the estimation of conditional mean models. In addition, the conventional assumption of nonparametric regression theory that there is additive, independently, and identically distributed (*iid*) error around a smooth underlying conditional mean function is highly implausible in certain data settings[2]. Thus, as in the parametric context, nonparametric methods are usefully complemented

[1]School of Mathematics, Statistics, and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Private Bag X01, Scottsville 3209, South Africa. [2]Institute of Human Virology, School of Medicine, University of Maryland, Baltimore, MD 21201, USA. ✉email: ashu3argaw@gmail.com

by nonlinear estimation of families of conditional quantile functions that relax the independence assumption[2]. The use of parametric and nonparametric regression models for analyzing patients' CD4 count in most applications implies that the estimated effects describe the average CD4 count. However, it is of even great interest to examine the quantile of the outcome distribution, such as the lower ($\leq 25\%$) quantile, which identifies patients at higher risk of developing illnesses.

Quantiles, commonly symbolized by the Greek letter τ, are location and scale parameters simultaneously. For a given $\tau \in (0, 1)$, the τ^{th} quantile is the value of a random variable, where $\tau \times 100\%$ of its value lies below it. In other words, it is the value where at most $(1 - \tau) \times 100\%$ of the value lies above. Thus, τth quantiles close to 0.5-quantile give the median, which is a well-known location parameter. On the other hand, τth quantiles close to zero or one give an idea of the scale. For instance, the interquartile range (IQR) is defined as the 0.75 quantile minus the 0.25 quantile: $IQR = Q_3 - Q_1$.

Quantile regression (QR) solutions are computed for a selected number of quantiles, typically the three quantiles along with two extreme quantiles, that is, for $\tau = \{0.05, 0.25(Q_1), 0.5(Q_2), 0.75(Q_3), 0.95\}$. This necessitates the search for a suitable compromise between the amount of output to manage and the results to interpret and summarize. Although in many practical applications of QR, the focus is on estimating a subset of quantiles, however, it is worth noticing that it is possible to attain estimates across the entire interval of conditional quantiles; in particular, the set: $\{\beta_\tau : \tau \in (0, 1)\}$[2].

QR is a versatile statistical method with many applications that complement mean regression[3,4]. Thus, it emerged as an effective analytic technique in numerous study areas of science due to its competence to drive inferences about individuals that rank below or above the conditional population mean and/or focused on features of the response beyond its central tendency[4–13]. QR is specifically appropriate for the parameters' heterogeneous effect as it yields inferences that can be legitimate irrespective of the true underlying distribution[4,14]. QR techniques look further into the data, get more information, and become more important[15]. By fitting models for more percentiles, one can detect the covariates' heterogeneous effects at the conditional distribution of the response, rather than just the conditional mean. That is especially useful when valuable information lies at the bottom or top quantiles. "QR also enjoys several properties, including equivariance to monotone transformations and robustness to outliers"[2,16]. A semiparametric extension of quantile regression models with different types of nonlinear effects included in the model equation leads to an additive quantile regression model (AQM)[12]. Such a model may reveal systematic differences in dispersion, tail behavior, and other features for covariates[2].

Additive mixed models (AMMs), an extension of additive models, have been developed precisely to incorporate linear and nonlinear effects, as well as random terms when the data are sampled according to longitudinal designs[4,17]. AMMs have been integrated into QR methods to obtain robust results, not only focused on features of the longitudinal outcome at its central tendency that may not be the best location to characterize the data specifically when the errors are non-normally distributed, and the location-shift hypothesis of the normal model is violated but also at conditional quantiles of the longitudinal outcome with no assumption about the response or errors distribution apart from the distribution is restricted to have the τth quantile to be zero. Thus, additive quantile mixed models, which have gained popularity recently as a general method for longitudinal data, bring a comprehensive and more complete picture of the nonparametric as well as the parametric effects[1,4].

CD4 cell count levels signify the well-being of an individual immune system (body's natural defense system against pathogens, infections, and illnesses). The CD4 cell counts of a person who does not have HIV can be between 500 and 1500 per cubic millimeter. Individuals living with HIV who have a CD4 count over 500 but whose immune response is still strong are usually in good health. However, individuals living with HIV who have a CD4 count below 200 are at high risk of developing severe illnesses and death[18,19].

With the CD4 count at deficient levels, patients' immunity is weak. If HIV-infected patients are not on treatment or not virally suppressed, they become vulnerable to acquire opportunistic infections (OIs), making them at risk of the new and ongoing coronavirus disease 2019 (COVID-19) infection and underlying illness[18]. The best strategy to avoid these infections and diseases is by enhancing the immune function level through HAART, a combination of multiple antiretroviral (ARV) drugs. HAART's fundamental goal is to prolong or stop the progression to AIDS and loss of life for those infected with HIV by suppressing and preventing the virus from making copies of itself. When the virus's level (viral load) in the blood is low or undetectable, there is less damage to the body's immune system and fewer HIV infection complications. Even though HIV treatment is prescribed for all individuals living with HIV, it is particularly critical for patients with low CD4 count to start treatment sooner rather than later and adhere to the treatment schedule[18,20]. While researchers believe that early diagnosis and effective treatment are essential to effective control, more research is needed to understand better the adaptive, innate, and host responses that alter viral load set-point and consequently prognosis and infectiousness[18,20].

The need for good and better health is one of each human being's fundamental rights without qualification of race, religion, gender, political conviction, financial, or social condition. Women's health includes their emotional, social, and physical welfare and is determined by these factors and the economic setting of their lives, as well as by biology. However, health issues evade the longer part of women. In national and universal forums, women have emphasized that equality, the sharing of family duties, development, and peace are necessary conditions to achieve good health all through the life cycle. Women are biologically and socially more vulnerable to HIV infection, especially in developing countries[21–24].

HIV/AIDS and other sexually transmitted diseases (STD) have a devastating effect on women's health, mostly young ladies. The consequences of HIV/AIDS go beyond women's health to include their families' economic support and livelihoods. Thus, the social, development, and health consequences of HIV/AIDS and other sexually transmitted diseases have strong gender dimensions that cannot be ignored[23–25]. Understanding the changing epidemiology of HIV using statistical disease models will allow the clinician to decide who may be at high risk and clarify the application of rules to avoid sequential HIV transmission[18,20,26,27]. Although antiretroviral (ARV) recommendation presently remains the same for all individuals living with HIV, examining the progression of

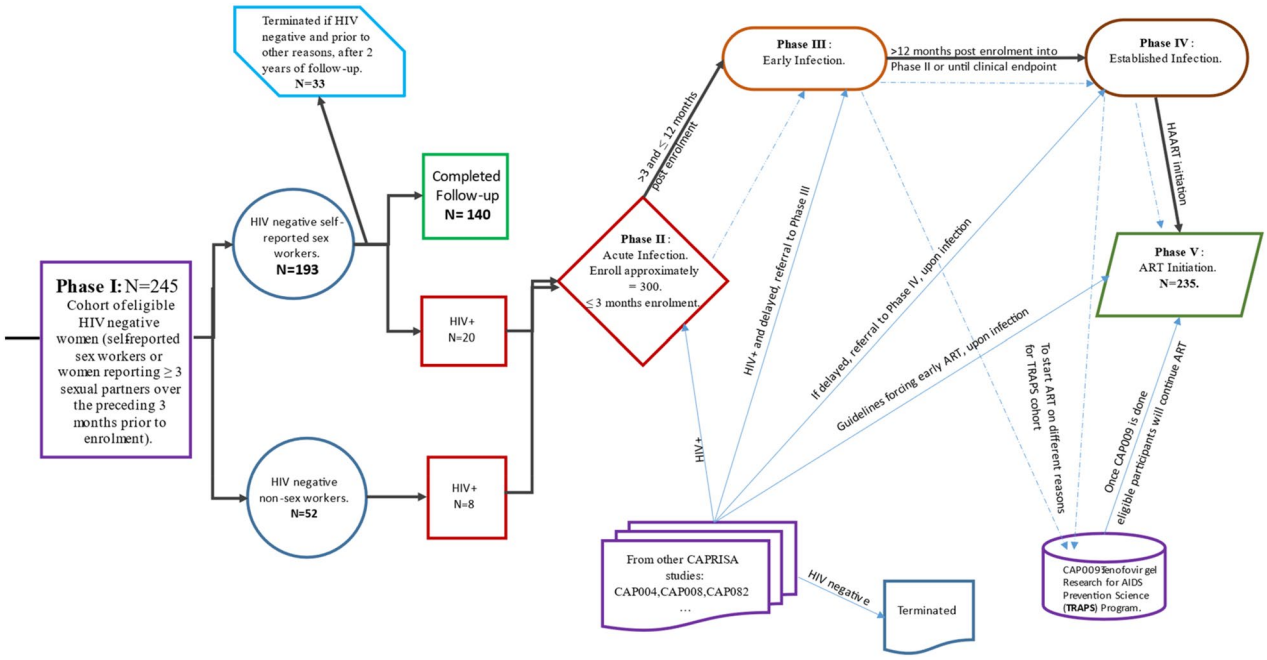

Figure 1. Diagrammatic overview of the CAPRISA 002 AI cohort study design.

CD4 count or evolution of the viral load using data-driven models will allow the clinician to interpret potential information accurately and cope with misdirection or distortion of the information due to patient-specific effects[18,26–28]. This study is a continuation of our previous work in Yirga et al.[18]. This study aims to analyze the longitudinal CD4 count of HIV-infected patients involved in a CAPRISA study using AQMM and justify how the method evolved can be used to attain robust nonparametric as well as parametric effects at various locations of the conditional distribution that brings a comprehensive and more complete picture of the covariate effects. The use of AQMM has many advantages. Additive nonparametric effects models are not new in the applied statistics literature. To implement these methods, Koenker et al.[47] introduce smoothing penalties for total variation, especially for the nonparametric components of the model. Researchers are also eager to learn what are the factors influencing the CD4 count (high or low) in HIV studies. AQMMs are the best way to answer this question.

Materials and methods

Data description. This study used data from the Centre for the AIDS Programme of Research in South Africa (CAPRISA). The CAPRISA study was effected at the Doris Duke Medical Research Institute (DDMRI) at the Nelson R Mandela School of Medicine of the University of KwaZulu-Natal in Durban, South Africa[18,29]. Between August 2004 and May 2005, CAPRISA introduced a cohort study registering high-risk HIV-negative women to a follow-up study with an intense ongoing examination. Women infected with HIV were recruited into the CAPRISA 002 Acute Infection (AI) study and then followed up carefully to study disease progression and CD4/viral load evolution[18,20,29–32].

Once HIV-infected women were enrolled in CAPRISA's AI Phase II study, their CD4 count and viral load were measured and assessed regularly. When their CD4 count $\leq$ 350 cells/mm^3 for more than two consecutive visits between six months or if they are with AIDS-defining illness (WHO clinical stage 3–5), they would be referred to a public government clinic for ARV treatment. However, according to the South African National Department of Health, these patients would only start HAART once their CD4 count is $\leq$ 200 cells/mm^3, until 2015. With effect from the 1st of January 2015, according to the National Department of Health, the criteria to start HIV patients on early initiation of ART is CD4 count of 500 cells/mm^3 or less than that[20]. HIV-infected women in Phase II–IV were followed up until they are started HAART. After that, they would be transitioned to Phase V and followed up for a minimum of five years, or eligible participants would be offered to join immediately into Phase V[33]. After the five years of follow-up have been accomplished, participants would be offered an optional annual follow-up for up to fifteen extra years to patients who recurred in Phase V[33]. Figure 1 illustrates the screening and enrolment process of the study data set. One can find further detail on the study population's design, development, and procedures here[29–33].

Consent for publication. Not applicable.

Methods

Parametric regression models typically use a linear function to connect the conditional values of the response variable to the covariates. In real-world applications, however, biased or invalid results might result from such a linearity assumption. Many studies use nonlinear assumptions between variables[34–37]. One may consider various

modeling techniques when dealing with nonlinearity. The most popular nonparametric models, smoothing splines, and transformation models use parameters such as sampling designs (cross-sectional or longitudinal), outcomes (discrete or continuous), distribution assumptions (parametric or nonparametric), and so on[2]. In choosing which method to follow, the amount of effort expended during the investigation may have a significant influence. Likewise, lacking theory or programming can lead to a certain decision being made over another[2].

Nonparametric regression permits the presumption of linearity to be relaxed[34,35,38] and limits the analysis to smooth and continuous functions[39]. Nonparametric regression, also known as scatter smoothing, aims to distinguish the best regression function according to the data distribution instead of estimating the parameters[39].

The nonparametric regression model is given by.

$$y = \sum_{i=1}^{n} f_i(x_i) + \varepsilon_i, \tag{1}$$

where the function $f_i(\cdot)$ is unknown, and commonly assumed that the errors are normally and identically distributed: $\varepsilon_i \sim NID\left(0, \sigma^2\right)$[39]. Several methods have been introduced to model nonparametric regression models; however, the most used techniques that have been extended to QR are local polynomial regression[40] and smoothing splines[41,42]: for further details, see Wu and Zhang[34], Fox[38], Davino et al.[39], Craig and Ng[43], Koenker et al.[44], Koenker[45], Cleveland and Loader[46], or Koenker et al.[47].

The parametric QR model is given by.

$$Y_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \varepsilon_{\tau i}, \quad i = 1, \ldots, n, \quad 0 < \tau < 1, \tag{2}$$

where Y_i is the response variable, x_i's are covariates, $\beta_{\tau i}$'s are the quantile specific linear effects, and $\varepsilon_{\tau i}$ is a random variable assumed to be an unknown error term on which no specific distributional assumptions are made except that the distribution is restricted to have the τth quantile to be zero[12,48,49]. For this reason, the parametric QR model aims at describing the quantile function $Q_{Y_i}(\tau|\boldsymbol{x}_i)$ of the continuous outcome Y_i conditional on covariate vector $\boldsymbol{x}_i$ at a given quantile τ, and this can be expressed as follows

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = F_{Y_i}^{-1}(\tau|\boldsymbol{x}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \varepsilon_{\tau i}, \quad \text{with } Q_{\varepsilon_{\tau i}}(\tau|\boldsymbol{x}_i) \sim F_{\tau i}, \tag{3}$$

where $F_{\tau i}$ is subject to $F_{\tau i}(0) = \tau$, $F_{Y_i}^{-1}(\cdot)$ is the inverse cumulative distribution function of Y_i. For a comprehensive overview of QR, see, for example, Koenker[2], Konker and Basset[3], Buchinsky[5], Yu et al.[9], or Koenker and Hallock[50].

As much as the parametric QR assumptions enjoy a simple model structure, convenience of interpretation, and lower computational cost, it is not flexible enough and hence carries the risk of model misidentifications for complex problems[51]. Nonparametric QR has become a viable alternative to avoid restrictive parametric assumptions. Koenker et al.[47] explored nonparametric QR in spline models (quantile smoothing splines), which they defined as solutions to

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \rho_\tau\left(y_i - f(x_i)\right) + \lambda \left(\int_0^1 |f''(x)|^p dx\right)^{1/p}, \tag{4}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}, p \geq 1$, is the so-called *check (loss) function,* the parameter $\tau \in (0, 1)$ controls the quantile of interest, and $\lambda \in \mathbb{R}^+$ is a smoothing parameter[3,47].

As closely analogous to the parametric QR model (3), Koenker[2] generalized nonparametric QR models as

$$Q_{Y_i}(\tau|\boldsymbol{x}_i) = f(\boldsymbol{x}_i, \beta_i(\tau)) \tag{5}$$

Then, Koenker[2] formulated the τth nonparametric QR estimator as

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}(\tau) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\left(y_i - f(\boldsymbol{x}_i, \beta(\tau))\right) \tag{6}$$

Several techniques were proposed for nonparametric QR modelings, such as Bivariate quantile smoothing spline[52] and Kernel quantile regression[53]. However, nonparametric QR is an important yet challenging topic that needs to be addressed in-depth[51]. One can find a brief account of nonparametric QR strategies in numerous studies; see, for example, Koenker[2] or Davino et al.[39]. To account for the nonlinearity relationships between quantiles of the outcome and covariates, Rigby and Stasinopoulos[54] also proposed generalized additive models for location, scale, and shape (GAMLSS). GAMLSS enables additional flexibility to fit the covariates' nonlinear effects; however, they do not result in easily interpretable expressions for the quantiles. They are based on specifying distinct distributional parameters[12]. Instead, additive quantile regression models (AQMs) allow for the inclusion of nonlinear covariate effects and give more flexibility[12].

Additive models, introduced by Hastie and Tibshirani[41], Stone[55], and Breiman and Friedman[56], are flexible regression tools that manipulate linear as well as nonlinear terms. The nonlinear terms in additive models are modeled through smoothing splines[4]. They provide programmatic approaches for nonparametric (nonlinear in parameters) regression modelings; by restricting nonlinear covariate effects to be composed of low-dimensional additive pieces so that we can overcome some of the worst aspects of the notorious curse of dimensionality[11]. The literature on additive models is vast[17,41,55,57,58]. However, most of the work has been done based on estimating conditional mean functions. The additive quantile regression model (AQM) provides an attractive framework

for parametric as well as nonparametric regression illustrations focused on features of the response beyond its central tendency[4,11,12].

Fenske et al.[12] defined the τth AQMs that extend the linear predictor, $\boldsymbol{x}_i^{'}\boldsymbol{\beta}_\tau$, with a sum of nonlinear functions of continuous covariates, $\sum f_{\tau j}(\cdot)$, as follows.

$$Q_{Y_i}(\tau|\boldsymbol{x}_i,\boldsymbol{z}_i) = \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} + \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i) + \varepsilon_{\tau i}, \quad j = 1,\ldots,q, \tag{7}$$

where $f_{\tau j}$ denote generic functions of covariates $\boldsymbol{z}_i$ for the ith observation, and allows for the inclusion of different model terms such as nonlinear effects (smooth functions) of z_k, $f_\tau(z_k)$, and varying coefficient terms, $z_k^{'}f_\tau(z_k)$, where the effect of the covariate $z_k^{'}$ varies smoothly over the domain of z_k according to some functions of f_τ. However, the underlying assumption of the error term, $\varepsilon_{\tau i}$, remains the same as in the QR model (3); see Fenske et al.[12] for more details.

AQM estimates the additive effect using linear programming algorithms as in the conventional QR model[12]. However, in the AQM case, determining adequate numbers and the position of knots is challenging. To avoid these challenges, Fenske et al.[12] used penalty methods such as quantile smoothing splines of Koenker et al.[47]. Thus, the minimization problem of AQM that consists of extra penalty term is given by[12]:

$$\underset{f_\tau}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}_{\tau i} - \sum_{j=1}^{q} f_{\tau j}(\boldsymbol{z}_i) \right) - \lambda \boldsymbol{v}\left(f_\tau^{'}\right), \tag{8}$$

where $\boldsymbol{v}\left(f_\tau^{'}\right) = sup \sum_{i=1}^{n-1} \left| f_\tau^{'}(z_{i+1}) - f_\tau^{'}(z_i) \right|$, represents the total variation of the derivation $f_\tau^{'} : [a,b] \to \mathbb{R}$, where the *sup* is taken over all partitions $a \le z_1 < \ldots < z_n < b$, and λ is a tuning parameter that controls the smoothness of the estimated function also known as "total variation regularization": see Koenker[2], Fenske et al.[12], or Koenker et al.[47] for more details.

Fenske et al.[1] proposed extending AMMs to the QR model for longitudinal data that consists of fixed individual-specific intercepts and slopes modeled through penalized splines of Ruppert et al.[59]. However, their model did not include random-effect terms and did not allow for individual-specific effects to have a general covariance structure[4]. The version of Geraci[4] additive QR model for longitudinal data includes linear and nonlinear terms, as well as multiple random effects to account for the correlation at the individual level with a general variance–covariance matrix and allow for automatic smoothing selection within a mixed model framework of Ruppert et al.[59]. Thus, as pointed out by Geraci[4], because of the following two basic ideas, his model was shown to have superior performance compared with the approach of Fenske et al.[1]: the first point is regarding the ith unit effects, which he assumed to be random instead of fixed so that the covariance structure between effects can be introduced; the second point is that instead of prior specification, the nonparametric term's smoothing is automatically estimated from the data[4].

Geraci[4] defined the τth additive QR model for longitudinal data as

$$Q_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{p} f_\tau^k\left(x_{ijk}\right) + z_{ij}^{'}u_{\tau,i}, \tag{9}$$

$$j = 1,\ldots,n_i, \quad i = 1,\ldots,m, \quad \tau \in (0,1),$$

where $x_{ij}^{'}$ is the jth row of a known $n_i \times p$ matrix $\boldsymbol{X}_i$, $z_{ij}^{'}$ is the jth row of a known $n_i \times q$ matrix $\boldsymbol{Z}_i$, y_{ij} is the jth observation of the response vector $\boldsymbol{y}_i = \left(y_{11},\ldots,y_{1n_i}\right)$ for the ith unit, $f_\tau^k(\cdot)$ is a τ-specific, centered, twice-differentiable smooth function of the kth component of $\boldsymbol{x}$, and $u_{\tau,i}$ is a $q \times 1$ vector of values that collects ith unit random effects associated with z_{ij} and its distribution is assumed to depend on a τ-specific parameter[4].

Geraci[4] considered a spline model of the type: $f_\tau(x) \approx \sum_{h=1}^{H} \nu_{\tau,h} B_h(x)$, to model nonlinear functions of the components of $\boldsymbol{x} = (x_1,\ldots,x_s,x_{s+1},\ldots,x_p)^{'}$ that consists of the first s terms of nonlinear functions and $p-s$ linear functions. The B_h's denote the *basis functions* (ν_τ), h's represent the corresponding τ-specific coefficients of B_h's and H indicates the number of knots[4]. The approximated quantile function from the model (9) is then expressed as follows[4]:

$$Q^*_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{k=1}^{s}\sum_{h=1}^{H_k} \nu_{\tau,hk}\, B_h^{(k)}\left(x_{ijk}\right) + \sum_{k=s+1}^{p} \beta_{\tau,k}\, x_{ijk} + z_{ij}^{'}u_{\tau,i} \tag{10}$$

In matrix notation, the ith unit of expression (10), which is then called additive quantile mixed model (AQMM), is given by[4]

$$Q^*_{\boldsymbol{y}_i|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \boldsymbol{F}_i\boldsymbol{\beta}_\tau + \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} + \boldsymbol{B}_i\boldsymbol{\nu}_\tau, \tag{11}$$

where $B^{(k)}\left(x_{ijk}\right)$ is considered as $H_{k\times 1}$ vector of values taken by the kth spline evaluated at x_{ijk}, $\nu_{\tau,k} = (\nu_{\tau,1},\ldots,\nu_{\tau,H_k})^{'}$ considered as the $H_{k\times 1}$ vector of spline coefficients for the kth covariate, and $H = \sum_k H_k$. Furthermore, $\boldsymbol{B}_i$ and $\boldsymbol{\nu}_\tau$, defined, respectively, as the $n_i \times H$ matrix with rows $\left(B^{(1)}\left(x_{ij1}\right)^{'},\ldots,B^{(s)}\left(x_{ijs}\right)^{'}\right)$ and

Variable	Descriptive measures						
	Mean	Median	Minimum	Maximum	$Q_{0.25}$	$Q_{0.75}$	IQR
SQRT_CD4 count (cells/μL)	23.26	22.98	5	44	20	26.19	6.19
Baseline VL (cells/**mL**)	130,730.33	26,600	1 (undetected)	5,510,000	5080	113,000	107,920
Age (Years)	27.15	25	18	59	22	30	8
Baseline BMI	28.98	26.84	17.89	54.89	23.33	32.96	9.63

Table 1. Descriptive statistics for non-categorical variables.

Variable	Total	Variable	Total
Place of residence		**Number of sexual partners**	
Rural (*reference*)	105 (44.7%)	No partner (*reference*)	43 (18.3%)
Urban	130 (55.3%)	Stable partner	182 (77.4%)
Educational level		Many partners	10 (4.3%)
Primary schools (*reference*)	11 (4.7%)	**Number of women**	235
Secondary schools	224 (95.3%)		

Table 2. Baseline descriptive statistics for categorical variables.

$\left(\boldsymbol{v}_{\tau,1}^{'},\ldots,\boldsymbol{v}_{\tau,s}^{'}\right)^{'}$, $\boldsymbol{F}_i$ is the $n_i \times (p-s+1)$ matrix with rows $(1, x_{ij(s+1)},\ldots,x_{ijp})'$ and $\boldsymbol{\beta}_\tau = (\beta_{\tau,0}, \beta_{\tau,s+1},\ldots,\beta_{\tau,p})'$[4].

The objective function of AQMM, where the vectors $\boldsymbol{u}_{\tau,i}$ and $\boldsymbol{v}_\tau$ are assumed to follow zero-centered multivariate Gaussian distributions with variance–covariance matrices $\boldsymbol{\Sigma}_\tau$ and $\boldsymbol{\Phi}_\tau = \oplus_{k=1}^{s}\phi_{\tau,k} I_{H_k}$, respectively, with selecting $\rho_\tau(\boldsymbol{r}) = \sum_{j=1}^{n} r_j\{\tau - I(r_j < 0)\}$ for a vector $\boldsymbol{r}=(r_1,\ldots,r_n)'$, is given by Geraci[4] as

$$\sum_{i=1}^{M}\rho_\tau\left(\boldsymbol{y}_i - \boldsymbol{F}_i\boldsymbol{\beta}_\tau - \boldsymbol{Z}_i\boldsymbol{u}_{\tau,i} - \boldsymbol{B}_i\boldsymbol{v}_\tau\right) + \sum_{i=1}^{M}||u_{\tau,i}||^2_{\Sigma_\tau^{-1}} + \sum_{k=1}^{s}\phi_{\tau,k}^{-1}||v_{\tau,k}||^2, \tag{12}$$

where "$\boldsymbol{u}_{\tau,i}$'s are assumed to be independent for different i (but may have a general covariance matrix) and are independent of $\boldsymbol{v}_\tau$, and $\phi_{\tau,k}$'s determine the amount of smoothing for the nonparametric terms"[4]. Minimizing the objective function of expression (12) proceeds as the same as minimizing the objective function of quantile mixed-effects models[49,60,61] where the asymmetric Laplace distribution with a location parameter μ, scale parameter $\sigma > 0$, and skewness parameter $\tau \in (0,1)$[60,62–64], are employed as *quasi-likelihood* for the fidelity term[4]. Further discussion of AQMM is provided by Geraci[4].

Ethical approval and consent to participate. The study was approved by the Research Ethics Committee of the University of KwaZulu-Natal (E013/04), the University of the Witwatersrand (MM040202), and the University of Cape Town (025/2004). All participants provided written informed consent. All methods were performed following the relevant guidelines and regulations expressed in the Declaration of Helsinki.

Results

Geraci[4] illustrated the full range of AQMM that is described above. The purpose of this analysis is to model the CD4 count of patients from KwaZulu-Natal, South Africa, as part of a comprehensive study of HIV/AIDS. The results of this study illustrate longitudinal CD4 counts among HIV-infected patients enrolled in the CAPRISA 002 AI study by employing an AQMM. The median age of our sample of 235 women was 25 years. Our sample consisted of 7019 measurements on 235 women from 18 to 59 years of age. There were multiple visits for all participants, ranging from 2 to 61, with a median of 29.

Tables 1 and 2 show descriptive measures for the variables studied. Low (upper) quantiles are those where at least 25% (75%) of the observations are at or below it, or 75% (25%) are at or above it[2]. In Table 1, it is shown that the median BMI for the participants was 26.84 (range 17.89–54.89). The median square root CD4 count and baseline viral load were 22.98 cells/mm^3 and 26,600 copies, respectively. Of a total of 235 women, 105 (44.7%) lived around Vulindlela (rural area), and 130 (55.3%) lived around eThekwini (Durban, urban area) in KwaZulu-Natal, South Africa (see Table 2). The majority of the women, 182 (77.4%), were in a stable partnership, 224 (95.3%) completed secondary school (Table 2), and most of them (78.8%) were self-reported sex workers[18,29,31]. Additional details are available here[29–32] concerning the CAPRISA 002 AI study. We analyze this study data set intending to explain the different conditional distribution of the CD4 count by considering two covariates entered as nonparametric additive effects: time and baseline BMI; as well as discrete (baseline viral load), continuous (age), and categorical covariates (place of residence, educational level, and the number of sexual partners) entered in the model as parametric effects (see Tables 1, 2). Figure 2 shows observed square root transformed CD4 counts

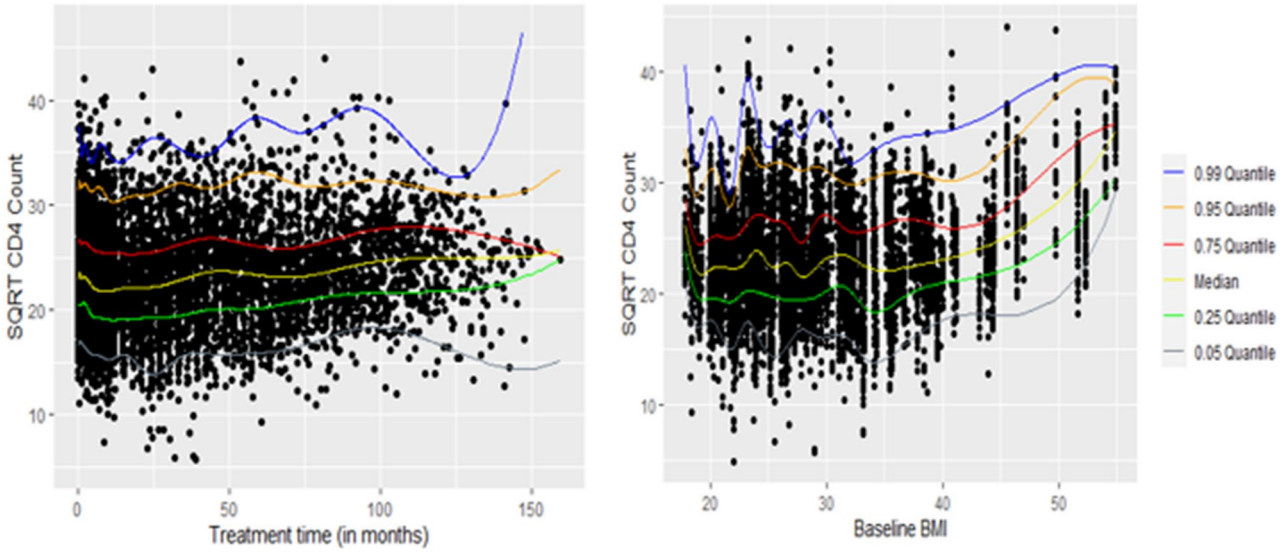

Figure 2. Observed CD4 counts (square root transformed) by time and baseline BMI.

by treatment time and baseline BMI, respectively, for a total of 7019 observations. The nonlinear patterns, which connect the sample quantiles, are estimated conditionally on time and baseline BMI for six quantile levels. The curves (nonlinear patterns) suggest the requirement of some degree of smoothing (Fig. 2).

Following the AQMM of Geraci[4], we used a transformed continuous form of the outcome (i.e., square root CD4 count) for fitting purposes. Thus, the proposed τ^{th} AQMM form of our study, using expression (10), can be specified as

$$\begin{aligned} Q^{*}_{y_{ij}|\boldsymbol{u}_i,\boldsymbol{x}_i,\boldsymbol{z}_i}(\tau) = \beta_{\tau,0} + \sum_{h=1}^{H_1} \nu_{\tau,1}\, B_h^{(1)}(time_i) + \sum_{h=1}^{H_2} \nu_{\tau,2}\, B_h^{(2)}(BMI_i) + \beta_{\tau,1}\, ART_i \\ + \beta_{\tau,2}\, VL_i + \beta_{\tau,3}\, residence_i + \beta_{\tau,4}\, education_i + \beta_{\tau,5}\, partner_i \\ + \beta_{\tau,6}\, age_i + u_{\tau,0} + u_{\tau,1}(time_i), \end{aligned} \tag{13}$$

where y_{ij} is the square root transformed form of the outcome ($\sqrt{CD4count}$) at the jth time point for the ith subject, time is the time variable measured in months from the start of the study, BMI indicates the patient's baseline BMI, ART is the dichotomous HAART initiation (0 = pre-ART, 1 = post-ART), VL is patient's baseline viral load, the residence is patient's place of residence, education is the educational level of participants, partner indicates the number of sexual partners of the participant, age is participant age at enrolment, $u_{\tau,0}$ indicates the random intercept, and $u_{\tau,1}$ indicates the random slope. The symbol τ specifies the quantile of interest; we made the estimation at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 to get the complete picture of the effects.

Geraci[4] employed the AQMM in the R package *lqmm* as an ad-on to fit additive quantile mixed models. As the same as the smooth terms' specification in the R package *mgcv*[17], one can enter continuous covariates within the *s* (smooth) function to control the model smoothness using splines when fitting AQMM[4]. Furthermore, the shrinkage smoothers obtained using the *bs* option inside the *s* command in the R package *mgcv* are constructed so that smooth terms can be penalized away altogether, not contribute to the model[17,65]. Thin plate smoother provides statistical and computational efficiency, stable optimal approximations (especially for large data sets), and can be constructed for smooths of more than one covariate at a time[4,66]. Thus, it was used as a shrinkage spline to fit the proposed model (13). The remaining parametric terms in the *aqmm* function[4] are specified the same way as in other R linear mixed model fitting functions such as *lqmm ()* and *lme4 ()*. The output is separated into two parts: Parametric part that includes estimated fixed effects, with their standard errors (SE), in parentheses, and significant mixed effect representation of smoothing splines (see Table 3). Since the smooth coefficients are mostly uninterpretable, we focus on their variances to evaluate the spline coefficients' penalty at various quantiles (see Table 4 and Supplementary information). However, their estimated smoothed effects are depicted in Fig. 3. Table 4 also presents the estimated variance of the random effects from the fitted model (13).

According to Table 3, the age effect is positive and significant at the bottom, median, and at $\tau = 0.75$ quantile levels (see also Supplementary information). On the other hand, the effect of education on square root CD4 count does not seems to be significant across all quantiles after the patient had been initiated on HAART. The square root CD4 count across all quantiles is affected by post-HAART initiation as expected. A significant positive effect of HAART initiation on CD4 cell counts is observed at the median quantile and becomes roughly constant at higher quantiles (see Table 3 and Supplementary information). In addition, patients with stable sexual partners showed significant improvements in their CD4 cell count across all quantiles. The CD4 cell count is significantly lowered in patients who have many sexual partners, especially at the bottom ($\tau = 0.05$) and at the top ($\tau = 0.95, 0.99$) quantiles (Table 3).

Fixed effects	$\widehat{Q}_{0.05}$ (SE)	$\widehat{Q}_{0.25}$ (SE)	$\widehat{Q}_{0.5}$ (SE)	$\widehat{Q}_{0.75}$ (SE)	$\widehat{Q}_{0.95}$ (SE)
Intercept	16.004 (0.6634)***	19.647 (0.4749)***	21.204 (0.5340)***	24.167 (1.0536)***	29.379 (0.6324) ***
Age	0.0398 (0.0156)**	0.0209 (0.0114)	0.0418 (0.0052)***	0.0331 (0.0078)***	0.0203 (0.0178)
Secondary school	− 0.4491 (0.5731)	− 0.4734 (0.4101)	− 0.0165 (0.6619)	0.0385 (1.0677)	0.8323 (0.5574)
Post HAART	0.7430 (0.0879)***	1.5296 (0.0598)***	1.5968 (0.0402)***	1.5292 (0.0546)***	1.7007 (0.1322)***
Baseline VL	− 3.83e−06 (8.42e−07)***	− 2.09e−06 (2.69e−07)***	− 1.79e−06 (2.41e−07)***	− 1.57e−06 (1.60e−07)***	− 1.70e−06 (2.21e−07)***
Urban	− 0.50002 (0.1668)**	0.2499 (0.0545)***	0.0998 (0.0334)**	0.1275 (0.1436)	− 0.8846 (0.2216)***
Stable partner	0.6135 (0.1655)***	0.3046 (0.1549)	0.5424 (0.1140)***	0.4907 (0.1594)**	0.6339 (0.2960)*
Many partners	− 2.2771 (0.2707)***	− 0.7858 (0.2589)**	− 0.8432 (0.1091)***	− 1.1719 (0.2569)***	− 3.6497 (0.4451)***
Results of the smooth terms					
s (Time)	− 2.5075 (0.5426)***	− 2.3766 (0.5549)***	− 2.1985 (0.4735)***	− 2.2829 (0.4999)***	− 2.3324 (0.4373)***
s (Baseline BMI)	5.4382 (1.0786)***	5.6868 (1.1094)***	5.5767 (1.3014)***	5.7904 (1.2077)***	5.2604 (1.0753)***

Table 3. Parameter estimates followed by results of the smoothing terms from the AQMM for the CAPRISA 002 AI study data across different quantiles. *Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.', 0.1 ' ', 1. The reference categories are given in Table 2.

Results across different quantiles							
	$\widehat{Q}_{0.05}$	$\widehat{Q}_{0.25}$	$\widehat{Q}_{0.5}$	$\widehat{Q}_{0.75}$	$\widehat{Q}_{0.85}$	$\widehat{Q}_{0.95}$	$\widehat{Q}_{0.99}$
Variance of the random effects							
$\hat{\sigma}_0$(Intercept)	0.02748	0.8687	0.0354	0.2453	0.3454	0.0467	0.0033
$\hat{\sigma}_0$(Time)	8.104e−18	1.929e−16	3.328e−17	5.451e−17	7.671e−17	1.044e−17	2.963e−18
Variance of the smooth terms							
$\hat{\phi}_{Time}$	8.796	28.94	36.74	30.28	21.92	10.13	2.669
$\hat{\phi}_{BaselineBMI}$	1876.501	6463.83	7823.81	6290.32	4979.39	2183.69	576.902

Table 4. Estimated variance of the random effects and smooth terms from the AQMM for the CAPRISA 002 AI study data.

Furthermore, we found a clear indication, at the bottom ($\tau = 0.05$) and more extreme quantiles ($\tau = 0.85, 0.95, 0.99$), that there is a significant negative effect of patients who were residing around the urban area on their CD4 cell count (see Table 3 and Supplementary information). Table 3 also shows that the negative effect of baseline viral load on the CD4 cell count is higher at the lower quantiles than at the median and higher quantiles (see also, Supplementary information). In addition, R package *aqmm()* sample outputs using CAPRISA 002 AI study data at $\tau = 0.25, 0.75, 0.85$, and 0.99 can be found in Supplementary information.

The variance of the first smooth term ($\hat{\phi}_{Time}$) indicates a stronger penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at the top quantiles (Table 4). Similarly, the variance of the second smoother ($\hat{\phi}_{BaselineBMI}$) shows a strong penalty on the spline coefficients at $\tau = 0.25, 0.5, 0.75, 0.85$ quantiles than at the bottom and at more extreme quantiles. Table 4 shows that the random effects' variances have roughly constant variability of subject linear trends across the fitted quantiles (see, also, Supplementary information).

Based on the seven fitted quantile levels ($\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95, 0.99$), Fig. 3 depicts the two estimated smoothed covariate effects on patients' CD4 counts. Patients enrolled in the CAPRISA 002 AI study exhibit nonlinear time effects on CD4 counts that are prominent at all quantile levels. As the quantile increases, its effect becomes stronger. However, it is after several treatment visits that such progress towards higher CD4 counts occurs. Consequently, the progression is slow until about 50 months, then it increases steadily thereafter across all quantile levels (Fig. 3).

Furthermore, overall fit quantile levels, the significant smoothed baseline BMI effect on patients' CD4 counts is roughly constant for patients with a baseline BMI of about 40 but gradually improves from there. Because of this, patients with low BMI need to be monitored carefully before and after HAART initiation. Despite this, physicians should not ignore patients with high BMI. According to our studies and other findings, a plausible explanation may be that BMI may affect drug metabolism and, thus, the progress of HAART and its immunological responses[20,67,68]. Moreover, higher levels of BMI have a more significant effect than lower levels (Fig. 3).

Discussion and conclusion

As a cutting-edge statistical method for modeling percentiles of response variables conditioned on respective covariates, quantile regression is the most widely used. While regression for medians may be seen as more robust than regression for the mean, QR, a generalization of median regression, allows better exploration of data by allowing the modeling of conditional quantiles at low or high extents, such as the 5th and 95th percentiles. As

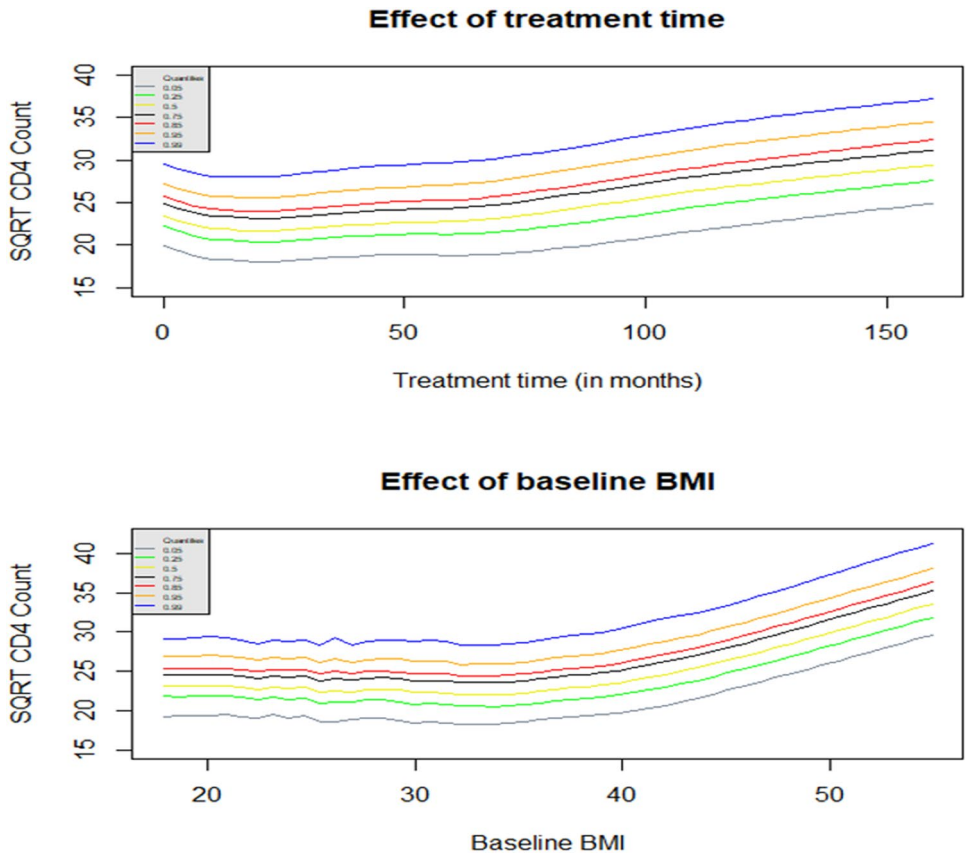

Figure 3. Predicted smoothed covariate effects on the square root CD4 count of HIV-infected patients recurred in the CAPRISA 002 AI study at various quantiles using AQMM.

a result, QR is becoming more common in clinical, biomedical, and other health-related research. Mean-based regression is used to formulate mixed-effects models and their estimated effects on the response variable. In some cases, this centrality-based inference method may not be the optimal method for dealing with the data since the data may not adequately represent their distribution. It has recently been demonstrated that QR has the potential to be extended to a mixed-effects modeling setting, even though QR was initially developed in a univariate setting[48,60,61]. Studies of quantile mixed-effects models have received increasing attention[15,48,60,61,69–76].

Quantile mixed-effects models have been extended to additive models to obtain robust results across various quantile levels of the longitudinal outcome, which brings a rigorous covariates' effect[74–76]. The additive version of the quantile mixed-effects model has gained a great deal of popularity, as discussed above; because it offers an efficient and flexible framework for nonlinear and linear longitudinal forms of data analysis focused on features of the outcome beyond its central tendency[1,4,11,12,47,73,75,76].

In this study, we investigated the effect of multivariate additive quantile mixed models of Geraci[4] on the longitudinal CD4 count of HIV-infected patients across different quantile levels according to parametric and nonparametric covariate effects. By using this recently developed model, robust results are obtained, not only at the central location of the longitudinal outcome that may not be the best place to analyze the data but also at different points of the conditional distribution that gives an inclusive and more complete picture of the parametric as well as the nonparametric covariate effects.

A series of AQMM at $\tau = 0.05, 0.25, 0.5, 0.75, 0.85, 0.95$, and 0.99 were performed, and the results were discussed. According to the results, patients' CD4 count is markedly increased after HAART initiation, and their baseline viral load shows a negative effect on the progression of their CD4 count over time, as we would have expected. All fitted quantiles of the response variable were affected by a significant nonlinear relationship between time and baseline BMI. Study results suggest that, across all fitted quantile levels, the patient's education level does not significantly influence the progression of CD4 counts over time. All but the most extreme quantiles of HIV-positive patients showed a significant difference in the CD4 count regardless of their age. In addition, CD4 cell recovery was found to be significant across all quantiles among patients with a stable sexual partner. Contrary to this, HIV-infected patients with many sexual partners during the treatment period showed a negative effect on CD4 cell count across all fitted quantile levels.

As we expected, the patient's CD4 count increased significantly after HAART was initiated, and their baseline viral load also showed a significant negative effect on the patient's CD4 count over time. Baseline BMI and time were also significant nonlinear effects in our analysis. Further, patients with higher BMIs at baseline have improved CD4 cell count over time after treatment. Despite this, higher BMI patients should not be ignored

clinically. This study instead suggests that BMI can influence drug metabolism and, consequently, the immunological responses to HAART. According to the nonlinear time effect, patients' CD4 counts are not increasing rapidly over time. The growth starts after multiple treatment visits. Hence, the study suggests that HIV patients who are not clinically and immunologically stable on HAART could experience increased risks if exposed to COVID-19, especially if they are not on HAART immediately after HIV exposure.

One can estimate the covariate effects over the grid $\tau \in (0, 1)$ as per the analysis aspects. An investigator, however, should be cautious when using AQMM since the method needs some adjustment to control the estimation algorithm and demands more computing time to estimate the random effects[4]. For instance, for this study, it took 2–3 h to fit the proposed model (13) at a single τ as like Geraci[4]. To overcome this computational burden, Geraci[4] suggested the necessity of further improvement to the AQMM. As the studied data set is an ongoing study, there is a plan to extend AQMM application to genetics in future work since it produces satisfactory results.

Data availability

The dataset used for this study can be obtained by requesting Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics Unit, CAPRISA, Email: Nonhlanhla.Yende@caprisa.org) on reasonable request.

Received: 19 May 2021; Accepted: 19 August 2021
Published online: 09 September 2021

References

1. Fenske, N., Fahrmeir, L., Hothorn, T., Rzehak, P. & Höhle, M. Boosting structured additive quantile regression for longitudinal childhood obesity data. *The International Journal of Biostatistics* **9**(1), 1–18 (2013).
2. Koenker, R. *Quantile Regression* (Cambridge University Press, Cambridge, 2005).
3. Koenker, R. & Bassett, Jr G. Regression quantiles. *Econ. J. Econ. Soc.* **46**(1), 33–50 (1978).
4. Geraci, M. Additive quantile regression for clustered data with an application to children's physical activity. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **68**(4), 1071–1089 (2019).
5. Buchinsky, M. Recent advances in quantile regression models: A practical guideline for empirical research. *J. Hum. Resour* . **33**(1), 88–126 (1998).
6. Koenker, R. & Geling, O. Reappraising medfly longevity: A quantile regression survival analysis. *J. Am. Stat. Assoc.* **96**(454), 458–468 (2001).
7. Peterson, M. D. & Krishnan, C. Growth charts for muscular strength capacity with quantile regression. *Am. J. Prev. Med.* **49**(6), 935–938 (2015).
8. Sherwood, B., Wang, L. & Zhou, X. H. Weighted quantile regression for analyzing health care cost data with missing covariates. *Stat. Med.* **32**(28), 4967–4979 (2013).
9. Yu, K., Lu, Z. & Stander, J. Quantile regression: Applications and current research areas. *J. R. Stat. Soc. Ser. D (The Statistician)* **52**(3), 331–350 (2003).
10. Cade, B. S. & Noon, B. R. A gentle introduction to quantile regression for ecologists. *Front. Ecol. Environ.* **1**(8), 412–420 (2003).
11. Koenker, R. Additive models for quantile regression: Model selection and confidence bandaids. *Braz. J. Probab. Stat.* **25**(3), 239–262 (2011).
12. Fenske, N., Kneib, T. & Hothorn, T. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Stat. Assoc.* **106**(494), 494–510 (2011).
13. Yirga, A. A., Ayele, D. G. & Melesse, S. F. Application of quantile regression: Modeling body mass index in Ethiopia. *Open Public Health J.* **11**(1), 221–233 (2018).
14. Winkelmann, R. Reforming health care: Evidence from quantile regressions for counts. *J. Health Econ.* **25**(1), 131–145 (2006).
15. Huang, Q., Zhang, H., Chen, J. & He, M. Quantile regression models and their applications: A review. *J. Biom. Biostat.* **8**, 354. https://doi.org/10.4172/2155-6180.1000354 (2017).
16. Gilchrist, W. *Statistical Modelling with Quantile Functions* (Chapman and Hall/CRC, London, 2000).
17. Wood, S. N. *Generalized Additive Models: An Introduction with R* (CRC Press, London, 2017).
18. Yirga, A. A., Melesse, S. F., Mwambi, H. G. & Ayele, D. G. Negative binomial mixed models for analyzing longitudinal CD4 count data. *Sci. Rep.* **10**(1), 1–15 (2020).
19. AIDSMAP. CD4 cell counts|aidsmap (2021). https://www.aidsmap.com/about-hiv/cd4-cell-counts.
20. Yirga, A. A., Melesse, S. F., Mwambi, H. G. & Ayele, D. G. Modelling CD4 counts before and after HAART for HIV infected patients in KwaZulu-Natal South Africa. *Afr. Health Sci.* **20**(4), 1546–1561 (2020).
21. World Health Organization. *Women, Ageing, and Health: A Framework for Action: Focus on Gender* (2007).
22. World Health Organization. *AIDS Epidemic Update: December 2009* (WHO Regional Office Europe, 2010).
23. UN Women. *Message from UN Women's Executive Director for World AIDS Day, the 1st of December 2014* (2014). https://www.unwomen.org/en/news/stories/2014/12/world-aids-day-2014.
24. AMFAR. *The Foundation for AIDS Research. Statistics: Women and HIV/AIDS* (2015). https://www.amfar.org/about-hiv-and-aids/facts-and-stats/statistics--women-and-hiv-aids/.
25. Whelan, D. Gender and HIV/AIDS: Taking stock of research and programmes. In *UNAIDS* (1999).
26. Kassutto, S. & Rosenberg, E. S. Primary HIV type 1 infection. *Clin. Infect. Dis.* **38**(10), 1447–1453. https://doi.org/10.1086/420745 (2004).
27. Cohen, M. S., Shaw, G. M., McMichael, A. J. & Haynes, B. F. Acute HIV-1 infection. *N. Engl. J. Med.* **364**(20), 1943–1954. https://doi.org/10.1056/NEJMra1011874 (2011).
28. Rosenberg, E. S. *et al.* Immune control of HIV-1 after early treatment of acute infection. *Nature* **407**(6803), 523 (2000).
29. Van Loggerenberg, F. *et al.* Establishing a cohort at high risk of HIV infection in South Africa: Challenges and experiences of the CAPRISA 002 Acute Infection study. *PLOS ONE* **3**(4), e1954 (2008).
30. Garrett, N. *et al.* Acceptability of early antiretroviral therapy among South African women. *AIDS Behav.* **22**(3), 1018–1024 (2018).
31. Mlisana, K. *et al.* Rapid disease progression in HIV-1 subtype C-infected South African women. *Clin. Infect. Dis.* **59**(9), 1322–1331 (2014).
32. Moosa, Y. *et al.* Case report: Mechanisms of HIV elite control in two African women. *BMC Infect. Dis.* **18**(1), 1–7 (2018).
33. Karim, S. A., Williamson, C. & Garrett, N. Viral set point and clinical disease progression: The role of immunological, genetic and viral factors over the course of disease and during antiretroviral therapy. CAP002: Acute Infection Study. (An ongoing study) (2017). Accessed 14 Mar 2021. https://www.caprisa.org/Pages/CAPRISAStudies.
34. Wu, H. & Zhang, J.-T. *Non-parametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches* Vol. 515 (Wiley, New York, 2006).
35. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. *Longitudinal Data Analysis* (CRC Press, London, 2008).

36. Lindsey, J. K. Nonlinear models in medical statistics. *Oxford University Press on Demand* (2001).
37. Davidian, M. & Giltinan, D. M. Nonlinear models for repeated measurement data: An overview and update. *J. Agric. Biol. Environ. Stat.* **8**(4), 387–419 (2003).
38. Fox, J. *Non-parametric Simple Regression: Smoothing Scatterplots, No. 130* (Sage, Thousand Oaks, 2000).
39. Davino, C., Furno, M. & Vistocco, D. *Quantile Regression: Theory and Applications* Vol. 988 (Wiley, New York, 2013).
40. Chaudhuri, P. Global non-parametric estimation of conditional quantile functions and their derivatives. *J. Multivar. Anal.* **39**(2), 246–269 (1991).
41. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models* Vol. 43 (CRC Press, London, 1990).
42. Hendricks, W. & Koenker, R. Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Stat. Assoc.* **87**(417), 58–68 (1992).
43. Craig, S. G. & Ng, P. T. Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *J. Urban Econ.* **49**(1), 100–120 (2001).
44. Koenker, R., Portnoy, S. & Ng, P.. Non-parametric estimation of conditional quantile functions. In Dodge, Y. (Ed) (1992).
45. Koenker, R. Censored quantile regression redux. *J. Stat. Softw.* **27**(1), 1–25 (2008).
46. Cleveland, W. S. & Loader, C. Smoothing by local regression: Principles and methods. In *Statistical Theory and Computational Aspects of Smoothing*. 10–49 (Physica-Verlag HD, 1996).
47. Koenker, R., Ng, P. & Portnoy, S. Quantile smoothing splines. *Biometrika* **81**(4), 673–680 (1994).
48. Liu, Y. & Bottai, M. Mixed-effects models for conditional quantiles with longitudinal data. *Int. J. Biostat.* **5**(1), 28 (2009).
49. Lachos, V. H., Chen, M.-H., Abanto-Valle, C. A. & Azevedo, C. L. Quantile regression for censored mixed-effects models with applications to HIV studies. *Stat. Interface* **8**(2), 203 (2015).
50. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Perspect.* **15**(4), 143–156 (2001).
51. Lin, C. Y., Bondell, H., Zhang, H. H. & Zou, H. Variable selection for non-parametric quantile regression via smoothing spline analysis of variance. *Statistics* **2**(1), 255–268 (2013).
52. He, X., Ng, P. & Portnoy, S. Bivariate quantile smoothing splines. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **60**(3), 537–550 (1998).
53. Li, Y., Liu, Y. & Zhu, J. Quantile regression in reproducing kernel Hilbert spaces. *J. Am. Stat. Assoc.* **102**(477), 255–268 (2007).
54. Rigby, R. A. & Stasinopoulos, D. M. Generalized additive models for location, scale and shape. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **54**(3), 507–554 (2005).
55. Stone, C. J. Additive regression and other non-parametric models. *Ann. Stat.* **13**(2), 689–705 (1985).
56. Breiman, L. & Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **80**(391), 580–598 (1985).
57. Der, G. & Everitt, B. S. *Applied Medical Statistics Using SAS* (CRC Press, London, 2012).
58. Xiang, D. Fitting generalized additive models with the GAM procedure. In *SUGI Proceedings* 256–326 (Cary, NC: SAS Institute, Inc., 2001)
59. Ruppert, D., Wand, M. P. & Carroll, R. J. *Semiparametric Regression, No. 12* (Cambridge University Press, Cambridge, 2003).
60. Geraci, M. & Bottai, M. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics* **8**(1), 140–154 (2007).
61. Galarza Morales, C. E. Quantile regression for mixed-effects models (2015). https://bit.ly/3i7BPyQ.
62. Koenker, R. & Machado, J. A. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**(448), 1296–1310 (1999).
63. Yu, K. & Moyeed, R. A. Bayesian quantile regression. *Stat. Probab. Lett.* **54**(4), 437–447 (2001).
64. Yu, K. & Zhang, J. A three-parameter asymmetric Laplace distribution and its extension. *Commun. Stat. Theory Methods* **34**(9–10), 1867–1879 (2005).
65. Zuur, A. *et al. Mixed Effects Models and Extensions in Ecology with R* (Springer, New York, 2009).
66. Wood, S. N. Thin plate regression splines. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **65**(1), 95–114 (2003).
67. Palermo, B., Bosch, R. J., Bennett, K. & Jacobson, J. M. Body mass index and CD4+ T-lymphocyte recovery in HIV-infected men with viral suppression on antiretroviral therapy. *HIV Clin. Trials* **12**(4), 222–227 (2011).
68. Li, X. *et al.* Predictive effects of body mass index on immune reconstitution among HIV-infected HAART users in China. *BMC Infect. Dis.* **19**(1), 1–9 (2019).
69. Galvao, A. F. Jr. Quantile regression for dynamic panel data with fixed effects. *J. Econ.* **164**(1), 142–157 (2011).
70. Fu, L. & Wang, Y.-G. Quantile regression for longitudinal data with a working correlation model. *Comput. Stat. Data Anal.* **56**(8), 2526–2538 (2012).
71. Lipsitz, S. R., Fitzmaurice, G. M., Molenberghs, G. & Zhao, L. P. Quantile regression methods for longitudinal data with drop-outs: Application to CD4 cell counts of patients infected with the human immunodeficiency virus. *J. Roy. Stat. Soc. Ser. C (Appl. Stat.)* **46**(4), 463–476 (1997).
72. Geraci, M. & Bottai, M. Linear quantile mixed models. *Stat. Comput.* **24**(3), 461–479 (2014).
73. Galarza, C. E., Lachos, V. H. & Bandyopadhyay, D. Quantile regression in linear mixed models: A stochastic approximation EM approach. *Stat Interface* **10**(3), 471 (2017).
74. Yue, Y. R. & Rue, H. Bayesian inference for additive mixed quantile regression models. *Comput. Stat. Data Anal.* **55**(1), 84–96 (2011).
75. Sriram, K., Shi, P. & Ghosh, P. A Bayesian semiparametric quantile regression model for longitudinal data with application to insurance company costs. In *IIM Bangalore Research Paper* 355 (2011).
76. Huang, Y. Quantile regression-based Bayesian semiparametric mixed-effects models for longitudinal data with non-normal, missing and mismeasured covariate. *J. Stat. Comput. Simul.* **86**(6), 1183–1202 (2016).

Acknowledgements

We gratefully acknowledge CAPRISA for giving us access to the CAPRISA 002: Acute Infection Study data. CAPRISA is funded by the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes for Health (NIH), and U.S. Department of Health and Human Services (Grant: AI51794). The authors would also like to thank Dr. Nonhlanhla Yende-Zuma (Head of Biostatistics unit at CAPRISA) for her cooperation, assistance, and technical support.

Author contributions

A.A.Y. obtained the data, did the analysis, and prepared the manuscript. A.A.Y., S.F.M., H.G.M., and D.G.A. planned the research problem. All authors deliberated on the results and consequences and commented on the paper at all stages. All authors contributed extensively to the work presented in this manuscript. All authors read and ratified the ultimate manuscript.

Funding
This work was supported through the DELTAS Africa Initiative and the University of KwaZulu-Natal. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant 107754/Z/15/Z], DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme and the UK government.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-97114-9.

Correspondence and requests for materials should be addressed to A.A.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.