

RESEARCH ARTICLE

Open Access



# Novel metrics for growth model selection

Matthew R. Grigsby<sup>1</sup>, Junrui Di<sup>2</sup>, Andrew Leroux<sup>2</sup>, Vadim Zipunnikov<sup>2</sup>, Luo Xiao<sup>3</sup>, Ciprian Crainiceanu<sup>2</sup> and William Checkley<sup>1,2\*</sup>

## Abstract

**Background:** Literature surrounding the statistical modeling of childhood growth data involves a diverse set of potential models from which investigators can choose. However, the lack of a comprehensive framework for comparing non-nested models leads to difficulty in assessing model performance. This paper proposes a framework for comparing non-nested growth models using novel metrics of predictive accuracy based on modifications of the mean squared error criteria.

**Methods:** Three metrics were created: normalized, age-adjusted, and weighted mean squared error (MSE). Predictive performance metrics were used to compare linear mixed effects models and functional regression models. Prediction accuracy was assessed by partitioning the observed data into training and test datasets. This partitioning was constructed to assess prediction accuracy for backward (i.e., early growth), forward (i.e., late growth), in-range, and on new-individuals. Analyses were done with height measurements from 215 Peruvian children with data spanning from near birth to 2 years of age.

**Results:** Functional models outperformed linear mixed effects models in all scenarios tested. In particular, prediction errors for functional concurrent regression (FCR) and functional principal component analysis models were approximately 6% lower when compared to linear mixed effects models. When we weighted subject-specific MSEs according to subject-specific growth rates during infancy, we found that FCR was the best performer in all scenarios.

**Conclusion:** With this novel approach, we can quantitatively compare non-nested models and weight subgroups of interest to select the best performing growth model for a particular application or problem at hand.

## Background

Childhood growth modeling plays an important role in understanding and surveilling health outcomes at both individual and population levels. Specific uses include predicting health outcomes based on current trajectories (e.g. failure to thrive, obesity, stunting, wasting) and understanding associations between growth outcomes and childhood exposures (e.g. environmental, gestational, disease) [1, 2]. Many types of statistical approaches have been proposed to model growth measurements as functions of age and related baseline covariates [3–11].

Frequently used statistical models such as linear mixed effects, quantile regression, and functional principal components methods provide great modeling flexibility and are often able to address key features of growth data such as sparsity of sampling, cross-sectional skewness, and smoothness of growth trajectories [12–14].

Comparing models requires an objective criterion that can be uniformly applied to all of them. Nested models can be compared via metrics such as the likelihood ratio test (LRT) or F-test, and penalization for parametrization with the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). However, comparing non-nested models is complicated because not all models optimize the same objective functions. Therefore, a comprehensive model selection strategy among

\*Correspondence: wcheckl1@jhmi.edu

<sup>1</sup> Division of Pulmonary and Critical Care, School of Medicine, Johns Hopkins University, 1830 E. Monument Street, 5th Floor, Baltimore, MD 21287, USA

Full list of author information is available at the end of the article



competing, often non-nested, models necessitates development of a universal selection criterion.

We propose a novel approach based on modifications of the mean squared error, including normalization, age-stratification, and weighting for subject-specific growth rates. These methods differ from those mentioned above in that they measure model predictive performance rather than model fit. Quantifying predictive accuracy at the subpopulation level is critically important in auxology applications. For example, subpopulations representing lower quantiles of growth often contain children who are either stunted or faltering and may require special attention. In such scenarios, model choice may necessarily be driven disproportionately by accuracy of predicting outcomes among said subpopulations. These proposed modifications are centered on an idea of using out-of-sample prediction accuracy as universal measures of model performance.

**Methods**

**Study setting**

This analysis used data collected in the CONTENT study, located in the two peri-urban communities of Pampas de San Juan Miraflores and Nuevo Paraíso. Both were high density populations located approximately 25 km south of Lima [15]. The original purpose of this study was to examine the impact of *Helicobacter pylori* on child growth using World Health Organization Multi-centre Growth Reference Study standards for calculating height and weight Z scores [15]. Further characterization of these regions can be found in previous publications [15, 16].

**Study design**

Data was collected longitudinally between May 2007 and February 2011 [15, 16]. Children were not included if they had severe disease requiring hospitalization, were part of a multiple pregnancy, had a birth weight less than 1500 grams, and/or their parents had intentions of moving during the period of the study [15]. Data was collected at birth with follow up lasting until the age of 24 months. Additional information on study design, including more specific details on information collected, can be found in the original publication [15].

**Biostatistical models**

When studying growth-related health outcomes and exposures, height and weight are usually collected at multiple time points to assess individual growth trajectories [1, 4, 17–20]. Notable features of longitudinal data include within-subject correlation, heterogeneity of individual baseline, and dynamic growth [21]. In this study, we employ traditional growth models such as linear

mixed effects (LME), as well as less well known techniques such as functional concurrent regression (FCR) and functional principal component analysis (fPCA) [13, 21–24]. For simplicity, we used height as our growth outcome in this study. Let  $Y_{ij}$  denote the height of child  $i$  at time point  $j$ , and  $t_{ij}$  is the corresponding age for child  $i$  at time point  $j$ , where  $i = 1, 2, \dots, 215$ , and  $j = 1, 2, \dots, m_i$ . Sex effect was included in LME and FCR models, and we denote  $X_i$  to be the sex for subject  $i$ . Even though linear regression with truncated cubic splines is well known and simple to implement, Grajeda et al. showed they were inaccurate when modeling longitudinal growth because they did not account for the nature of repeated measurements clustered within subjects and because the assumption on independence between measurements was violated [21].

**Parametric, linear mixed effects model**

Inclusion of subject-specific random effects is a convenient way to account for subject level clustering and is easy to implement in most statistical software packages [3, 13, 21, 25].

Since growth exhibits a pronounced non-linear association with age, population mean growth is modeled using truncated cubic splines with knots at 3, 6, 12, and 18 months. Random slopes and intercepts were used to capture the heterogeneity in growth curves. Specifically, random intercepts depict shifts (up or down) of subject-level growth from the population-level intercept, while random slopes represent subject-level growth velocity around the population prediction.

Although standard LME models are intended to account for within subject correlation, it has been shown that, in growth data, random intercept and slope models may have autocorrelated residuals [21]. Therefore, we used a continuous autoregressive error of order one to model the correlation structure between pairs of measurements for any subject. The model is formulated as

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 t_{ij} + b_{1i} t_{ij}) + \beta_2 t_{ij}^2 + \beta_3 t_{ij}^3 + \sum_{k \in \{3,6,12,18\}} \gamma k (t_{ij} - k)_+^3 + \alpha_1 X_i + \epsilon_{ij}$$

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim MVN \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} \right)$$

$$\begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{im_i} \end{bmatrix} \sim MVN \left( \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \dots & \rho^{|t_{i1}-t_{im_i}|} \\ \vdots & \ddots & \vdots \\ \rho^{|t_{i1}-t_{im_i}|} & \dots & 1 \end{bmatrix} \right)$$

where  $\beta$ 's and  $\gamma$ 's represent the fixed effects of time and age on height, while  $b_{0i}$  and  $b_{1i}$  represent the random intercepts and slopes, respectively. We assume independence between subjects.

**Nonparametric, functional models**

It has been noted that some parametric models may not be sufficiently flexible to fully capture the non-linearity in individual growth trajectories [24]. Therefore, non-parametric approaches have gained popularity in recent years to deal with longitudinal data. One reason to think of repeated measurements as functions at different time points is because the derivatives could be of interest as well (e.g. growth rates of children). Two functional approaches are discussed next.

Functional principal component analysis has become a first-line approach to analyzing functional or longitudinal data [22, 26–29]. It involves non-parametric estimation of the covariance structure and identifying the dominant features (eigenfunctions) of the covariance matrix. Subjects' random effects are a linear combination of a relatively small number of the eigenfunctions. This allows for increased complexity in the shape of estimated subject-level trajectories, but typically requires more parameters to be estimated than with LME models. Fast Covariance Estimations (FACES) was developed as a fast bi-variate smoothing method for the covariance operator which has been proved to be widely reliable and computationally efficient [30]. A newer version of FACES was designed to handle sparse functional data with a revised bivariate smoother, and a fast algorithm for approximating the leave-one-subject-out cross validation for selection of the smoothing parameter [31]. The model can be expressed as

$$Y_{ij} = f_0(t_{ij}) + b_i(t_{ij}) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

where  $f(\cdot)$  is a smooth mean function and  $b_i(\cdot)$  is generated from a zero-mean Gaussian process with covariance operator  $C(s, t) = Cov(b_i(s), b_i(t))$ . Detailed methods to model and estimate  $C(s, t)$  as tensor-product splines and to predict subject  $i$ 's growth curve  $X_i(t) = f(t_i) + u_i(t_{ij})$  can be found in Xiao et al. [31].

Functional principal component analysis is a way to examine functional variability, however, it is not directly comparable to LME models since it does not take into account effects of other covariates such as gender. As a generalization, we will consider functional concurrent regression (FCR) as a more natural extension of both LME and fPCA because they include time invariant gender fixed effects which correspond with the LME models, but also utilize benefits of modeling growth data as

a complex function similar to fPCA. Functional concurrent regression models were introduced and developed in recent years [24, 32–39]. The comparable FCR model to the LME model specified above can be expressed as

$$Y_{ij} = f_0(t_{ij}) + \alpha_1 X_i + b_i(t_{ij}) + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

where  $f_0(\cdot)$  is a smooth estimate of the average population growth curve,  $\alpha_1$  is the time-invariant fixed sex effect, and  $b_i(\cdot)$  models the subject-specific random functional deviation of subject  $i$  and is generated from a zero-mean Gaussian process with covariance function  $C(s, t)$ . Furthermore,  $b_i(\cdot)$  and  $(\epsilon_{i1}, \dots, \epsilon_{im_i})$  are assumed to be mutually independent across subjects. Smoothing parameters can be selected using either restricted maximum likelihood or generalized cross validation as described by Wood et al. [40, 41]. From a modelling perspective, it is notable that fPCA is a special case of FCR without effects from covariates other than time. The addition of fixed effects in this context is non-trivial. Details on the FCR estimation procedure are further described by Leroux et al. and an accompanied R package [42, 43].

**Definition of comparison criteria metrics**

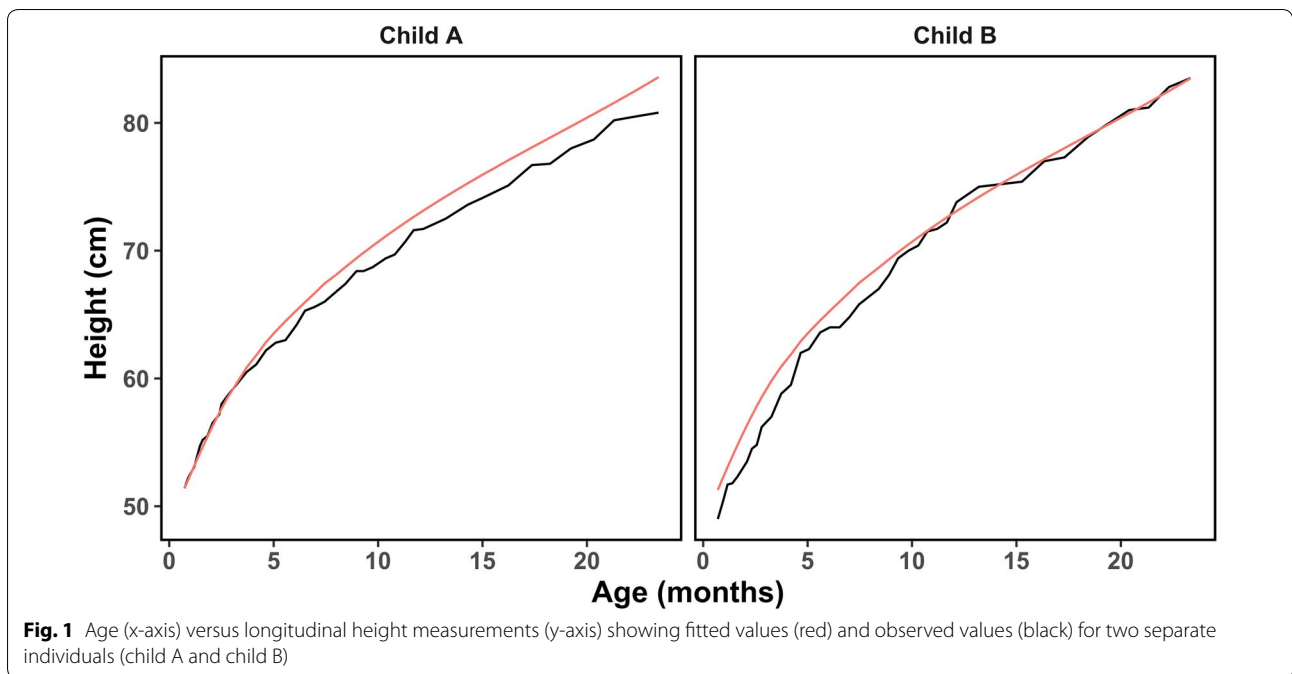
In this section, we introduce three metrics to perform growth model comparison. Let  $\hat{Y}_{ij}^k$  be the fitted value obtained from model  $k$ , and  $i = 1, \dots, 215$ ,  $j = 1, \dots, m_i$ , and  $k = 1, \dots, 3$ .

**Mean squared error (MSE)**

The first widely used selection metric is a subject specific mean squared error defined as

$$MSE_i^k = \frac{1}{m_i} \sum_{j=1}^{m_i} (Y_{ij} - \hat{Y}_{ij}^k)^2$$

where  $i = 1, \dots, n$ . Subject-specific MSEs can be combined to evaluate the performance of models on subpopulations of interest. One of the key limitations of using the MSE is demonstrated in Fig. 1. The black lines show observed growth curves of two selected children (child A and B), while the red lines show predicted growth curves. Child A has more error among larger height values, while child B tends to have increased error among smaller values. Un-normalized MSEs may disproportionately favor the most recent, almost always the largest, observations. As a result, the MSE for child A is inflated and greater than that of child B. However, normalization revealed that child A has lower overall error compared to child B. The scale of the data is not always consistent among subjects. Thus, subjects with larger measurements might



dominate the comparisons when using metrics that contain original measure units such as MSE [44]. Moreover, subjects with larger changes can bring more difficulty in comparison when using MSE [45]. Despite these problems, practitioners and academicians still tend to rely this kind of absolute error measurement [44–46]. We next introduce three modifications to the MSE that better account for specifics of child growth data.

**Normalized mean squared error (nMSE)**

It has been widely accepted that using relative error measurements which are unit-free can improve comparison performance and account for differences in measurement units as well as heteroscedasticity, thus providing fairer comparisons of predictive models [45, 47–49]. Subject-specific normalized mean squared error adds localized normalization and is defined as

$$nMSE_i^k = \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{(Y_{ij} - \hat{Y}_{ij}^k)^2}{Y_{ij}^2}$$

which can be considered as percentage errors. The error expressed in percentages gives a more robust metric of goodness-of-fit that can be uniformly applied across a wide age span.

**Age-stratified mean squared error (aMSE)**

Age-stratified mean squared error performs age-stratification and calculates within-strata subject-specific MSEs. It is defined as

$$aMSE_{is}^k = \frac{1}{m_i} \sum_{j \in Age_s} (Y_{ij} - \hat{Y}_{ij}^k)^2$$

**Weighted mean squared error (wMSE)**

It is also possible to create a metric using the MSE, nMSE, or aMSE that weights subgroups of interest. For example, we weighted individuals based on their growth velocity between 3 and 12 months so that slower growing individuals carried more weight. We used the following equations

$$(Height\ velocity)_i = \frac{h_{ti12} - h_{ti3}}{t_{i12} - t_{i3}}$$

$$wMSE_i^k = \frac{1}{m_i} \sum_{j=1}^{m_i} \frac{(Y_{ij} - \hat{Y}_{ij}^k)^2}{(Quartile\ of\ height\ velocity)_i}$$

where  $h_{ti}$  are height values at the corresponding  $t_i$  (time points) closest to 3 and 12 months of age. We calculated quartile of height velocity based on each child’s growth velocity relative to all others in our sample. Therefore, the height velocities of children in the 0th–24th, 25th–49th, 50th–74th, and 75th–100th would be assigned values of 1, 2, 3, and 4, respectively. This is one example of weighting specific individuals; one could also weight other subject-specific metrics of interest (e.g. those with poorer outcomes).

**Model comparisons**

Four common scenarios in growth modeling were considered: forward, backward, in-range, and new individual

prediction (Fig. 2). Forward prediction represents the scenario where missingness happens in the later stages of growth and the goal is to use data from the earlier stages to predict missingness in later stages. Backward prediction is opposite to forward; missingness in the early stages is predicted using data from later stages. In-range prediction happens where missingness takes place inside of the monitoring period, and new individual prediction occurs when there is missingness for an entire individual.

Error was measured by holding out a portion of the data (out-of-sample), fitting models to in-sample data, and then measuring predictive accuracy on the observations held out. With forward, backward, and in-range, analysis was performed by randomly selecting 50% of the children and subsequently holding out 10, 20, and 50% of their data. For new individual prediction, we randomly selected 10, 20, and 50% of the children to hold out. Primary analysis was performed using the 20% method, with the 10 and 50% used for comparison in sensitivity analysis.

Model performance will be presented as median and interquartile range (IQR) of MSE, nMSE, wMSE, or aMSE for each of the three model types (i.e. LME, fPCA, and FCR).

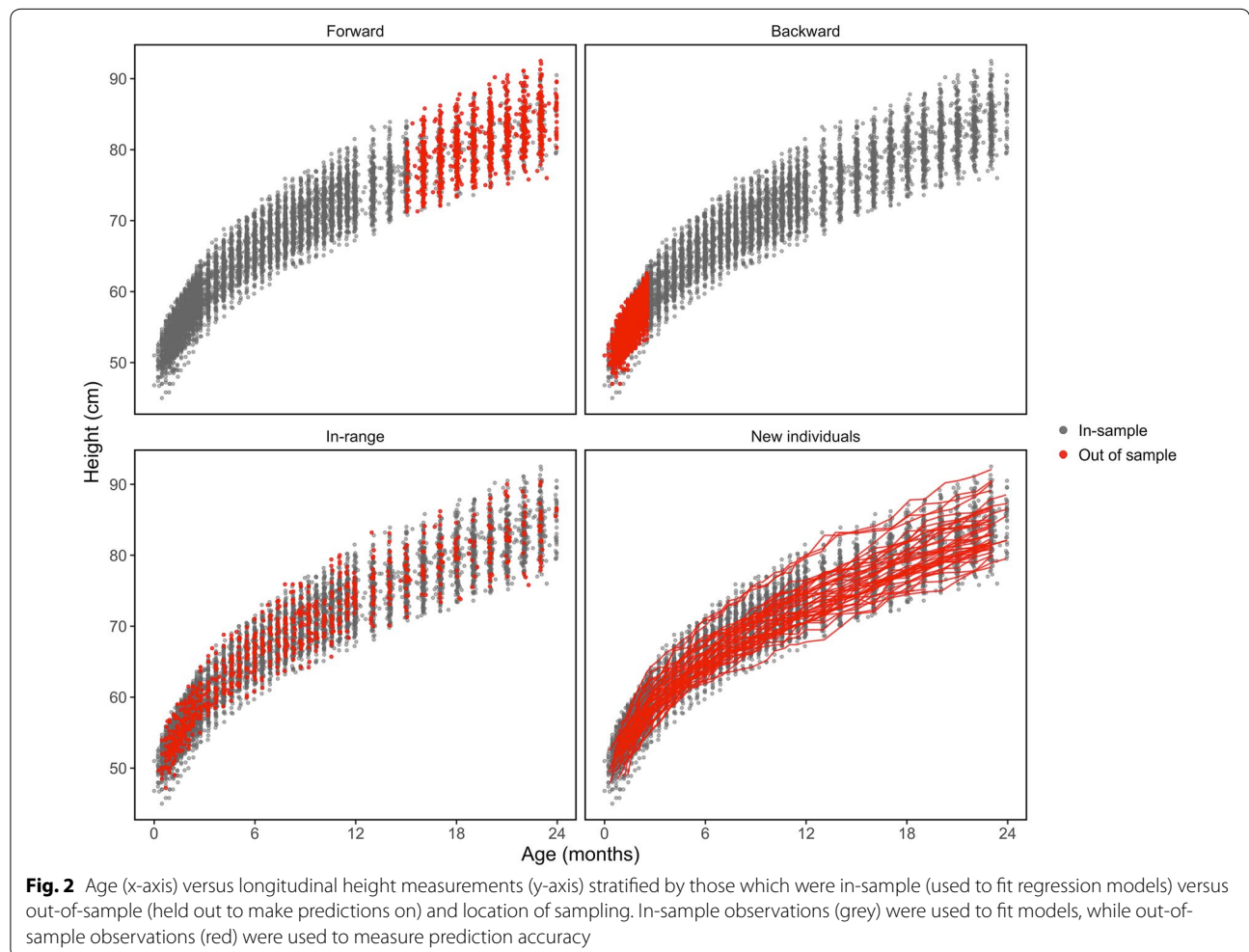
**Results**

**Population characteristics**

The final sample included 215 with complete data out of 304 eligible children. Eleven (3.6%) had incomplete anthropometric data and 78 (25.7%) did not have follow up past the age of 1 year. There were 39 observations per individual on average, with males representing 49% (n = 106) of the sample. Median lag between observations was 14 days (interquartile range 11–27).

**In-range**

In-range prediction error was lowest with fPCA when using nMSE, lowest with FCR when using wMSE, and the same when using MSE (Tables 1, 2, and 3). Median nMSE ranged from 10.1E–5 (IQR 5.6E–5 to 18.1E–5) for LME



to  $5.6E-5$  ( $3.7E-5$  to  $7.7E-1$ ) for FCR. Distributional properties of prediction error for each model, metric, and sampling method can be seen in Fig. 3. Using aMSE, we saw varying distributions again with in-range prediction. FCR and fPCA performed similarly well followed by LME (Fig. 4).

**Forward**

Model error with forward prediction, using median nMSE, ranged from 0.79 (IQR 0.39–1.74) for LME to 0.62 (0.32–1.42) for fPCA. Model error was lowest for fPCA when utilizing MSE and nMSE, but FCR slightly outperformed fPCA with wMSE (Tables 1, 2, and 3). There were similar results using aMSE with fPCA performing best, followed by FCR and LME (Fig. 4). Intra-strata comparison showed a trend, with FCR, fPCA, and LME performing better in ages 12–18 months compared to ages 18–24 months.

**Backward**

Backward prediction revealed similar results. fPCA performed best using MSE and nMSE, but FCR performed best with wMSE (Tables 1, 2, and 3). Median nMSE

**Table 1 Median and interquartile range for MSE stratified by location of prediction and model type**

	Backward	Forward	In-range	New individuals
LME	0.58 (0.29, 1.40)	0.79 (0.39, 1.74)	0.41 (0.27, 0.73)	0.39 (0.29, 0.47)
FCR	0.49 (0.26, 0.92)	0.71 (0.29, 1.29)	0.23 (0.17, 0.36)	0.17 (0.13, 0.22)
fPCA	0.48 (0.24, 0.90)	0.62 (0.32, 1.42)	0.24 (0.17, 0.35)	0.18 (0.14, 0.22)

Best performing models are in italics. Error was measured by holding out a portion of the data (out-of-sample), fitting models to in-sample data, and then measuring predictive accuracy on the observations held out. With forward, backward, and in-range, analyses were performed by randomly selecting 50% of the children and subsequently holding out 20% of their data

**Table 2 Median and interquartile range for nMSE, stratified by location of prediction and model type**

	Backward	Forward	In-range	New individuals
LME	18.19 (9.77, 43.90)	11.67 (6.25, 24.59)	10.14 (5.59, 18.08)	8.55 (6.75, 12.21)
FCR	17.44 (8.11, 31.11)	10.84 (4.47, 18.78)	5.57 (3.73, 7.67)	3.85 (3.02, 4.78)
fPCA	16.17 (8.05, 32.35)	9.78 (4.72, 19.86)	5.72 (3.81, 7.98)	4.10 (3.17, 4.76)

Best performing models are in italics. All values in Table 2 were multiplied by  $10^5$  to help better visualize performance differences. Error was measured by holding out a portion of the data (out-of-sample), fitting models to in-sample data, and then measuring predictive accuracy on the observations held out. With forward, backward, and in-range, analyses were performed by randomly selecting 50% of the children and subsequently holding out 20% of their data

**Table 3 Median and interquartile range for wMSE, stratified by location of prediction and model type**

	Backward	Forward	In-range	New individuals
LME	0.23 (0.12, 0.68)	0.39 (0.17, 0.76)	0.18 (0.11, 0.39)	0.17 (0.12, 0.27)
FCR	0.19 (0.10, 0.28)	0.24 (0.12, 0.45)	0.10 (0.06, 0.16)	0.07 (0.05, 0.12)
fPCA	0.21 (0.12, 0.45)	0.25 (0.10, 0.62)	0.12 (0.07, 0.19)	0.08 (0.06, 0.13)

Best performing models are in italics. Error was measured by holding out a portion of the data (out-of-sample), fitting models to in-sample data, and then measuring predictive accuracy on the observations held out. With forward, backward, and in-range, analyses were performed by randomly selecting 50% of the children and subsequently holding out 20% of their data

ranged from  $18E-5$  (IQR  $9.8E-5$  to  $43.9E-5$ ) for LME to  $16.2E-5$  ( $8.1E-5$  to  $32.3E-5$ ) for fPCA. All predicted points in backward prediction fell between ages 0–6 months. Therefore, the aMSE did not stratify the data and was interpreted as the standard MSE.

**New individuals**

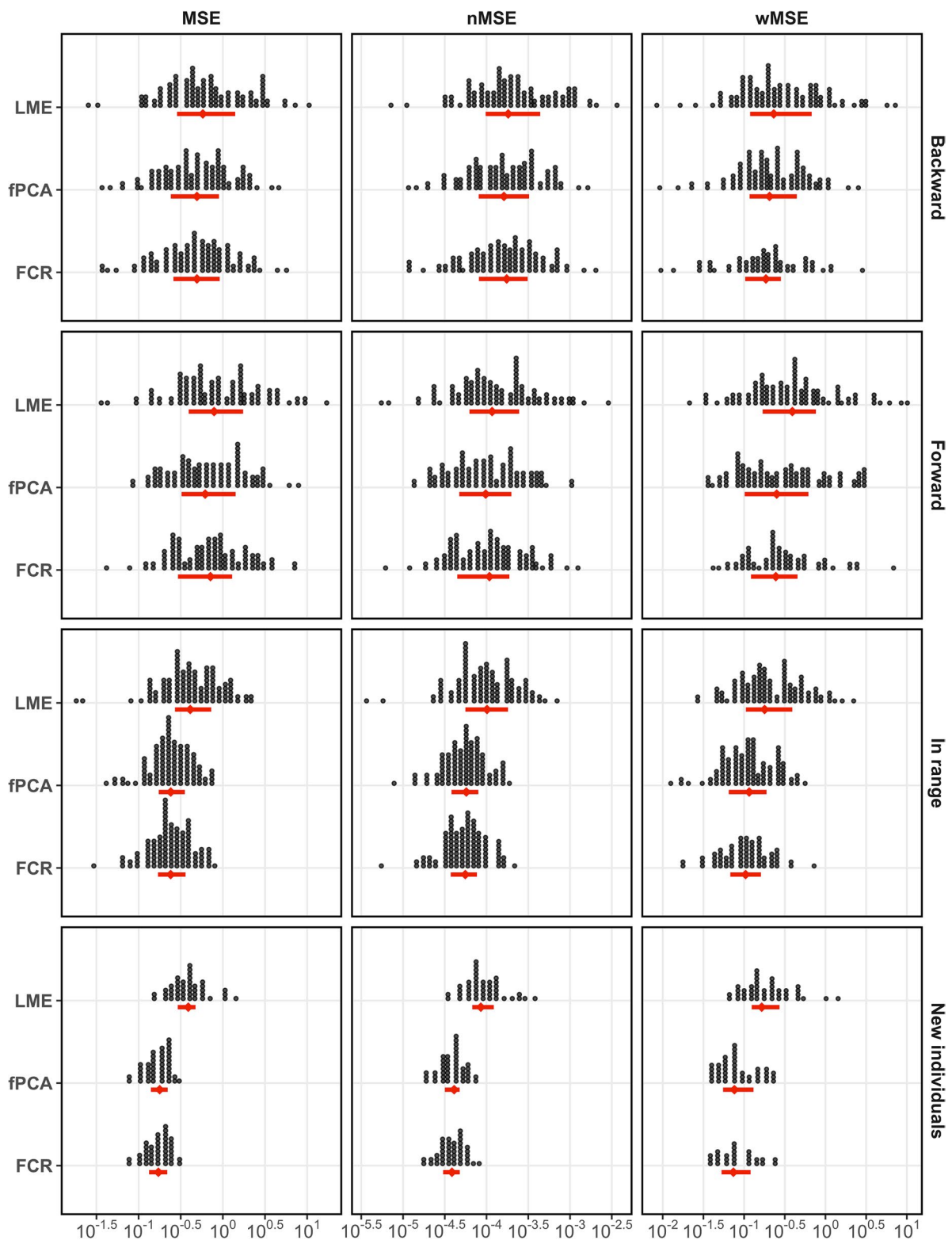
When predicting in-range on new individuals, FCR slightly outperformed fPCA for all metrics and sampling methods. Median nMSE ranged from  $8.6E-5$  (IQR  $6.6E-5$  to  $12.2E-5$ ) for LME to  $3.9E-5$  ( $3.0E-5$  to  $4.8E-5$ ) for FCR. Error distributions using aMSE were consistent with the above findings, with FCR and fPCA performing best followed by LME (Fig. 4). Between-strata differences were more apparent for LME, with LME showing less error at higher age ranges.

**Sensitivity analyses**

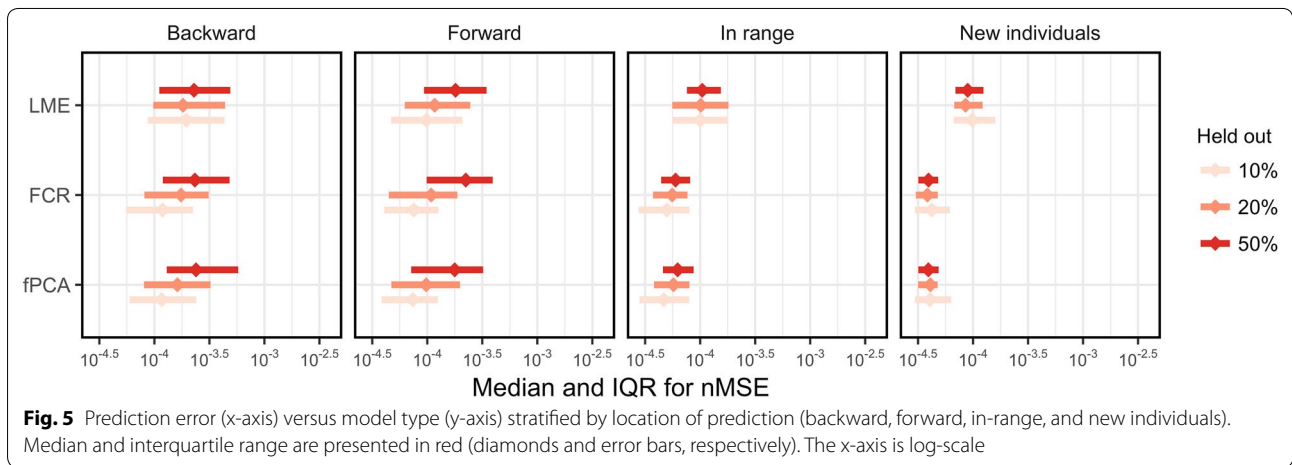
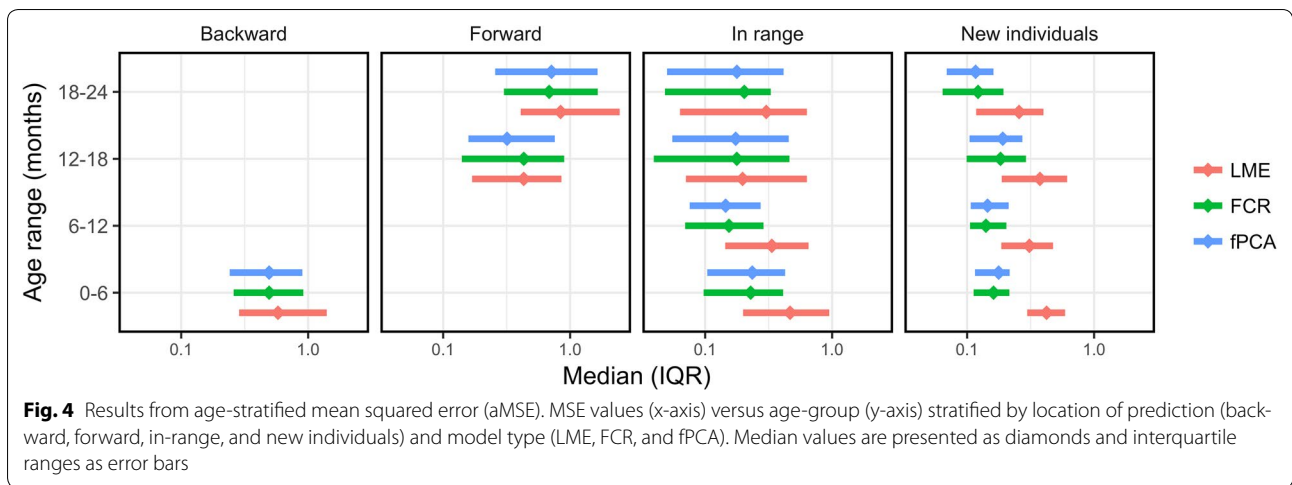
As seen in Fig. 5, prediction error in backward, forward, and in-range tended to be larger with increased number of observations held out. This trend was not as apparent when predicting on new individuals.

**Discussion**

This analysis demonstrates how to compare growth models (both nested and non-nested) by measuring prediction error via nMSE, wMSE, and aMSE. Each metric is subject-specific and can be used in a variety of real world situations. Sampling techniques can be adjusted to replicate exact scenarios of interest. Utilizing the nMSE and aMSE addresses the issue of the MSE favoring larger measurements. Furthermore, the aMSE can illuminate intra-age group performance differences and the wMSE demonstrates the ability to weight specific subgroups of interest, potentially helping to further detect performance gaps between growth models.



**Fig. 3** Prediction error (x-axis) versus model type (y-axis) stratified by prediction error metric (MSE, nMSE, and wMSE) and location of prediction (backward, forward, in-range, and new individuals). Median and interquartile range are presented in red (diamonds and error bars, respectively) and individual observations are presented in black (dots). The x-axis is log-scale with different ranges for MSE, nMSE, and wMSE due to differences in scale of prediction error values



Based on the results of this study, functional models outperformed traditional linear models in all scenarios. Even when utilizing proven techniques with LME (i.e. truncated cubic splines and autoregressive correlation correction), FCR and fPCA performed better in all scenarios tested [21, 50]. The difference in prediction error between FCR and fPCA in most situations was relatively small. Employing the wMSE revealed a shift in the best performing model when predicting backward, forward, and in-range. In these situations, the MSE and nMSE preferred fPCA as the best performing model while the wMSE showed FCR outperforming fPCA (Tables 1, 2, and 3). While functional models consistently outperformed LME, it seems they were more sensitive to the proportion of data removed when predicting backward and in-range (Fig. 5). One possible explanation is that shapes of curves are well defined for LME with cubic splines; however, for functional approaches, it is more difficult to predict trajectories of growth curves with limited amount of data.

A limitation of the MSE is its tendency to be inflated by outliers. Using subject-specific estimates partially addresses this, but there is still the possibility of having outliers within subjects. Sensitivity analysis should be performed to assess whether more robust (outlier-insensitive) approaches are necessary. There are a few other limitations to this study. First, even though we used a variety of sampling strategies, they do not comprehensively represent real world situations. There are more scenarios that could not be included in this analysis, such as predicting backward and forward on new individuals as well as choosing different hold out percentages. Second, aMSE can be less useful in certain situations. For example, age-stratification may not be needed when predicting over a relatively short age range or if data is sparse with fewer observations in each age group.

Our study also has some potential strengths. First, the proposed method is a novel approach of transforming the subject-specific MSE (i.e. nMSE, aMSE, and wMSE) to assess prediction error differences between



both nested and non-nested growth models. Alternative methods such as AIC, BIC, F-test, and the LRT only work for nested models. Second, our approach is flexible, allowing adaptation to specific real-world situations. The ability to weight subgroups of interest and adapt the age ranges used with aMSE contributes to this. Third, the CONTENT dataset is of high quality and high resolution. There were very few outliers regarding growth trends and the average number of observations per child was approximately 40 within a 2-year span. Finally, this analysis employed modern growth modeling techniques. FCR, fPCA, and LME are proven effective techniques for longitudinal growth modeling [13, 22, 24, 26–29, 32–39, 51–53].

## Conclusion

Subject-specific normalized mean squared error, age-stratified mean squared error, and weighted mean squared error are useful metrics for comparing both nested and non-nested growth models. We applied these metrics to three competing modeling methods and demonstrated the ability to weight subgroups of interest and evaluate performance gaps.

## Authors' contributions

MG, JD, and AL were responsible for conducting analysis. MG, JD, AL, VZ, and WC were responsible for data interpretation and creating the first draft of the manuscript. LX provided mentoring and assistance with coding statistical models in R. WC and CC provided mentoring. All authors read and approved the final manuscript.

## Author details

<sup>1</sup> Division of Pulmonary and Critical Care, School of Medicine, Johns Hopkins University, 1830 E. Monument Street, 5th Floor, Baltimore, MD 21287, USA.

<sup>2</sup> Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. <sup>3</sup> Department of Statistics, North Carolina State University, Raleigh, NC, USA.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

This study was approved by the European Union Ethics Committee, A.B. PRISMA and Universidad Peruana Cayetano Heredia, Lima, Peru, and the Johns Hopkins Bloomberg School of Public Health, Baltimore, USA.

## Funding

Our work was supported by the Bill and Melinda Gates Foundation [OPP1114097].

## Publisher's Note

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

Received: 16 May 2017 Accepted: 14 February 2018

Published online: 23 February 2018

## References

- Fairley L, Petherick ES, Howe LD, Tilling K, Cameron N, Lawlor DA, et al. Describing differences in weight and length growth trajectories between white and Pakistani infants in the UK: analysis of the Born in Bradford birth cohort study using multilevel linear spline models. *Arch Dis Child*. 2013;98:274.
- Regnault N, Gillman MW. Importance of characterizing growth trajectories. *Ann Nutr Metab*. 2014;65(2–3):110–3.
- Borghesi E, De Onis M, Garza C, Van den Broeck J, Frongillo E, Grummer-Strawn L, et al. Construction of the world health organization child growth standards: selection of methods for attained growth curves. *Stat Med*. 2006;25(2):247–65.
- Howe LD, Tilling K, Matijasevich A, Petherick ES, Santos AC, Fairley L, et al. Linear spline multilevel models for summarising childhood growth trajectories: a guide to their application using examples from five birth cohorts. *Stat Methods Med Res*. 2016;25(5):1854–74.
- Richard SA, McCormick BJ, Miller MA, Caulfield LE, Checkley W. Modeling environmental influences on child growth in the MAL-ED cohort study: opportunities and challenges. *Clin Infect Dis*. 2014;59(suppl 4):S255–60.
- Tilling K, Macdonald-Wallis C, Lawlor DA, Hughes RA, Howe LD. Modelling childhood growth using fractional polynomials and linear splines. *Ann Nutr Metab*. 2014;65(2–3):129–38.
- Berkey CS, Reed RB. A model for describing normal and abnormal growth in early childhood. *Hum Biol*. 1987;56:973–87.
- Count EW. Growth patterns of the human physique: an approach to kinetic anthropometry: Part I. *Hum Biol*. 1943;15(1):1–32.
- Guo S, Roche AF, Fomon SJ, Nelson SE, Chumlea WC, Rogers RR, et al. Reference data on gains in weight and length during the first two years of life. *J Pediatr*. 1991;119(3):355–62.
- Karlberg J, Jalil F, Lam B, Low L, Yeung C. Linear growth retardation in relation to the three phases of growth. *Eur J Clin Nutr*. 1994;48:S25–43 (**Discussion S-4**).
- Wingerd J. The relation of growth from birth to 2 years to sex, parental size and other factors, using Rao's method of the transformed time scale. *Hum Biol*. 1970;42:105–31.
- Koenker R. *Quantile regression*. Cambridge: Cambridge University Press; 2005.
- Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics*. 1982;38:963–74.
- Yao F, Müller H-G, Wang J-L. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*. 2005;100(470):577–90.
- Jaganath D, Saito M, Gilman RH, Queiroz DM, Rocha GA, Cama V, et al. First detected *Helicobacter pylori* infection in infancy modifies the association between diarrheal disease and childhood growth in Peru. *Helicobacter*. 2014;19(4):272–9.
- Checkley W, Gilman RH, Black RE, Lescano AG, Cabrera L, Taylor DN, et al. Effects of nutritional status on diarrhea in Peruvian children. *J Pediatr*. 2002;140(2):210–8.
- Checkley W, Epstein LD, Gilman RH, Black RE, Cabrera L, Sterling CR. Effects of *Cryptosporidium parvum* infection in Peruvian children: growth faltering and subsequent catch-up growth. *Am J Epidemiol*. 1998;148(5):497–506.
- Checkley W, Epstein LD, Gilman RH, Cabrera L, Black RE. Effects of acute diarrhea on linear growth in Peruvian children. *Am J Epidemiol*. 2003;157(2):166–75.
- Roth DE, Perumal N, Al Mahmud A, Baqui AH. Maternal vitamin D3 supplementation during the third trimester of pregnancy: effects on infant growth in a longitudinal follow-up study in Bangladesh. *J Pediatr*. 2013;163(6):1605–11.
- Tilling K, Davies N, Windmeijer F, Kramer MS, Bogdanovich N, Matush L, et al. Is infant weight associated with childhood blood pressure? Analysis of the promotion of breastfeeding intervention trial (PROBIT) cohort. *Int J Epidemiol*. 2011;40(5):1227–37.

21. Grajeda LM, Ivanescu A, Saito M, Crainiceanu C, Jaganath D, Gilman RH, et al. Modelling subject-specific childhood growth using linear mixed-effect models with cubic regression splines. *Emerg Themes Epidemiol*. 2016;13(1):1.
22. Besse P, Ramsay JO. Principal components analysis of sampled functions. *Psychometrika*. 1986;51(2):285–311.
23. Crainiceanu CM, Ruppert D, Wand MP. Bayesian analysis for penalized spline regression using Win BUGS. *J Stat Softw*. 2005;14:1–24.
24. Guo W. Functional mixed effects models. *Biometrics*. 2002;58(1):121–8.
25. Goldstein H. Efficient statistical modelling of longitudinal data. *Ann Hum Biol*. 1986;13(2):129–41.
26. Besse PC, Cardot H, Ferraty F. Simultaneous non-parametric regressions of unbalanced longitudinal data. *Comput Stat Data Anal*. 1997;24(3):255–70.
27. Kneip A. Nonparametric estimation of common regressors for similar curve data. *Ann Stat*. 1994;22:1386–427.
28. Ramsay JO, Dalzell C. Some tools for functional data analysis. *J R Stat Soc Ser B (Methodol)*. 1991;53:539–72.
29. Staniswalis JG, Lee JJ. Nonparametric regression analysis of longitudinal data. *J Am Stat Assoc*. 1998;93(444):1403–18.
30. Xiao L, Zipunnikov V, Ruppert D, Crainiceanu C. Fast covariance estimation for high-dimensional functional data. *Stat Comput*. 2016;26(1–2):409–21.
31. Xiao L, Li C, Checkley W, Crainiceanu CM. Fast covariance estimation for sparse functional data. [arXiv preprint arXiv:1603.05758](https://arxiv.org/abs/1603.05758); 2016.
32. Cederbaum J, Pouplier M, Hoole P, Greven S. Functional linear mixed models for irregularly or sparsely sampled data. *Stat Model*. 2016;16(1):67–88.
33. Chen H, Wang Y. A penalized spline approach to functional mixed effects model analysis. *Biometrics*. 2011;67(3):861–70.
34. Di C-Z, Crainiceanu CM, Caffo BS, Punjabi NM. Multilevel functional principal component analysis. *Ann Appl Stat*. 2009;3(1):458.
35. Hadjipantelis PZ, Aston JA, Müller H-G, Evans JP. Unifying amplitude and phase analysis: a compositional data approach to functional multivariate mixed-effects modeling of Mandarin Chinese. *J Am Stat Assoc*. 2015;110(510):545–59.
36. Reimherr M, Nicolae D. Estimating variance components in functional linear models with applications to genetic heritability. *J Am Stat Assoc*. 2016;111(513):407–22.
37. Yuan Y, Gilmore JH, Geng X, Martin S, Chen K, J-I Wang, et al. FMEM: functional mixed effects modeling for the analysis of longitudinal white matter tract data. *NeuroImage*. 2014;84:753–64.
38. Zhou L, Huang JZ, Martinez JG, Maity A, Baladandayuthapani V, Carroll RJ. Reduced rank mixed effects models for spatially correlated hierarchical functional data. *J Am Stat Assoc*. 2010;105(489):390–400.
39. Zhu H, Brown PJ, Morris JS. Robust, adaptive functional regression in functional mixed model framework. *J Am Stat Assoc*. 2012;106:1167.
40. Wood SN. *Generalized additive models: an introduction with R*. Boston: CRC Press; 2017.
41. Wood SN, Li Z, Shaddick G, Augustin NH. Generalized additive models for gigadata: modeling the UK black smoke network daily data. *J Am Stat Assoc*. 2017;112:1–12.
42. Leroux A, Xiao L, Crainiceanu C, Checkley W. Dynamic prediction in functional concurrent regression with an application to child growth. *Stat Med*. 2017. <https://doi.org/10.1002/sim.7582>
43. R Core Team. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing; 2017.
44. Carbone R, Armstrong JS. Note. Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners. *J Forecast*. 1982;1(2):215–7.
45. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: empirical comparisons. *Int J Forecast*. 1992;8(1):69–80.
46. Chatfield C. Apples, oranges and mean square error. *Int J Forecast*. 1988;4(4):515–8.
47. Flores BE. A pragmatic view of accuracy measurement in forecasting. *Omega*. 1986;14(2):93–8.
48. Makridakis S. Accuracy measures: theoretical and practical concerns. *Int J Forecast*. 1993;9(4):527–9.
49. Sanders NR. Measuring forecast accuracy: some practical suggestions. *Prod Inventory Manag J*. 1997;38(1):43.
50. Funatogawa I, Funatogawa T, Ohashi Y. An autoregressive linear mixed effects model for the analysis of longitudinal data which show profiles approaching asymptotes. *Stat Med*. 2007;26(9):2113–30.
51. Davidian M, Giltinan DM. *Nonlinear models for repeated measurement data*. Boston: CRC Press; 1995.
52. Lindstrom MJ, Bates DM. Nonlinear mixed effects models for repeated measures data. *Biometrics*. 1990;46:673–87.
53. Ramsay JO. *Functional data analysis*. New York: Wiley; 2006.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

