scientific reports

OPEN

A better performing algorithm for identification of implausible growth data from longitudinal pediatric medical records

Kylie K. Harrall[1,2]✉, Sarah M. Bird[2,3], Keith E. Muller[1], Lauren A. Vanderlinden[4], Maya E. Payton[5], Anna Bellatorre[2], Dana Dabelea[2,4,6] & Deborah H. Glueck[2,6]

Tracking trajectories of body size in children provides insight into chronic disease risk. One measure of pediatric body size is body mass index (BMI), a function of height and weight. Errors in measuring height or weight may lead to incorrect assessment of BMI. Yet childhood measures of height and weight extracted from electronic medical records often include values which seem biologically implausible in the context of a growth trajectory. Removing biologically implausible values reduces noise in the data, and thus increases the ease of modeling associations between exposures and childhood BMI trajectories, or between childhood BMI trajectories and subsequent health conditions. We developed open-source algorithms (available on *github*) for detecting and removing biologically implausible values in pediatric trajectories of height and weight. A Monte Carlo simulation experiment compared the sensitivity, specificity and speed of our algorithms to three published algorithms. The comparator algorithms were selected because they used trajectory information, had open-source code, and had published verification studies. Simulation inputs were derived from longitudinal epidemiological cohorts. Our algorithms had higher specificity, with similar sensitivity and speed, when compared to the three published algorithms. The results suggest that our algorithms should be adopted for cleaning longitudinal pediatric growth data.

Errors in measures of pediatric height and weight occur frequently[1–3]. Height and weight values abstracted from medical records often contain values that seem biologically implausible (See definitions in the Glossary, in Table 1)[4]. Single values may be biologically implausible if they exceed standardized population norms for particular ages. Values may also be deemed biologically implausible if they appear anomalous in the context of an observed growth trajectory. Examples of biologically implausible values in trajectories include decreases in height across periods of growth, or sudden and improbable weight gains or losses.

We describe novel, free, open-source algorithms to remove biologically implausible values in height and weight. For clarity when comparing our algorithms to other published algorithms, we will refer to our algorithms as the Harrall algorithms. The Harrall algorithms identify anomalous values by using information from the trajectories of growth in height and weight, under four assumptions. First, weight values must be for children over 2 years of age, although height values can be cleaned starting at birth. Second, the algorithms assume that there can be only one true value of height or weight at a single time point. Third, the algorithms assume that height must be increasing across childhood. Fourth, the algorithms assume that absolute weight velocity between two subsequent weight points cannot exceed the absolute magnitude of slope which would occur if a child moved from the CDC[5] 97th percentile in weight at the first age considered, to the 97th percentile in weight at the age corresponding to the second point. Although the algorithms require no manual curation, the algorithms provide data describing biologically implausible and plausible variables, in case investigators wish to review the results.

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida School of Medicine, Gainesville, FL, USA. [2]Lifecourse Epidemiology of Adiposity and Diabetes Center, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [3]Department of Biostatistics, Colorado School of Public Health, Anschutz Medical Campus, Aurora, CO, USA. [4]Deparment of Epidemiology, Colorado School of Public Health, Anschutz Medical Campus, Aurora, CO, USA. [5]Urban Institute, Washington, DC, USA. [6]Department of Pediatrics, University of Colorado School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. ✉email: KylieHarrall@ufl.edu

Algorithmic cleaning An approach which uses computer code to assess, and flag or remove values which are deemed implausible according to pre-specified rules

Biologically implausible value or anomalous value A measure of height or weight which seems to be erroneous, either because (1) it exceeds normative values for humans, (2) the value is inconsistent with normative theories of human growth, or (3) both

Body mass index The square of height in meters, divided by weight in kilograms

Cap A potentially biologically implausible value. If sequential points in the trajectory are connected by line segments, the line segment before a cap point has a positive slope, and the line segment after a cup point has a decrease in slope

Cup A potentially biologically implausible value. If sequential points in the trajectory are connected by line segments, the line segment before a cup point has a negative slope, and the line segment after a cup point has a positive slope

Externally defined limits An approach for defining normative bounds for values in height and weight, by using aggregate population data published by organizations

Manual curation An approach which uses human experts to assess, flag, and potentially remove biologically implausible values

Normative bounds Any limit, usually defined from aggregate population data, defined so that observed data which exceeds the limit is considered to be biologically implausible

Outlier A measure that falls outside of biologically normative bounds. The bounds are usually derived by studying large populations

Repeat An identical measure at two different but consecutive ages

Sensitivity The rate at which an algorithm identifies as correct values which are, in truth, correct

Specificity The rate at which an algorithm identifies as incorrect values which are, in truth, incorrect

Table 1. Glossary.

The design goal for developing the algorithms was to provide a systematic and reproducible way to clean longitudinal measurements of height and weight in pediatric cohorts. The Harrall algorithms are computer-based, and require no manual curation. Computer-based algorithms, like the ones we propose, are suitable for rapid processing of data with large numbers of participants, large numbers of repeated measures per participant, or both.

Obtaining accurate height and weight values is crucial for studies which seek to link anthropometric measures and health conditions. Assessing height and weight in longitudinal epidemiological studies allows computation of body mass index (BMI), a measure of body size. Many cohort studies assess associations between prenatal or early life exposures and subsequent childhood BMI trajectories[6–9]. Examples include assessment of associations between gestational diabetes[6], sex steroids[10], or dietary patterns[7] and BMI later in development. Similarly, many studies assess associations between childhood BMI trajectories and the development of subsequent health conditions. Cohen et al.[11], for example, studied links between BMI growth in early life and hepatic fat in childhood.

To assess accuracy of the Harrall algorithms for cleaning height and weight, we compared the Harrall algorithms to three published algorithms: Daymont et al.[1], Shi et al.[2] and Phan et al.[3] (see Table 2 for a brief description of the three algorithms). Algorithms were included for comparison if the developers provided open-source code, provided a verification study, and accounted for longitudinal data. The inclusion criteria meant that many algorithms were omitted from the comparison. Several algorithms reviewed by Lawman et al.[4] did not take into account longitudinal data[12–20] and others did not provide open-source code[21–23]. The algorithm by Yang and Hutchon[24] had open-source code, but did not validate the cleaning method.

The three algorithms chosen for comparison require some manual curation, make strong assumptions, or both. The algorithms described by Shi et al.[2] and Phan et al.[3] both require manual curation for some issues. Shi et al.[2] recommend that all flagged values undergo manual inspection before removal from the dataset. Phan et al.[3] require individual trajectories undergo manual inspection if the standard deviation of model residuals exceeds a threshold, or if over 40% of values in the individual trajectory are flagged. Manual curation uses human experts to assess, flag, and potentially remove biologically implausible values. The algorithms from Daymont et al.[1], Shi et al.[2] and Phan et al.[3] are all based on parametric assumptions about the biological models for growth in height or weight. Among other checks, the algorithm from Daymont et al.[1] evaluates deviations from a weighted moving average. Comparison to a weighted moving average evaluates distance from a line parallel to the x-axis, with

Anomalies algorithm	Approach	Testing method
Daymont et al.[1]	Sequential removal of improbable values from individuals using standard deviations exceeding an iteratively determined cut point for each of three weighted moving averages. This method accounts for variable intervals from the point of interest using inverse-distance weighting. Separate calculations handle unit errors and switches, very extreme values, repeats over time, duplicates, height decreases of over 3 cm, evaluation of single or pairs of measurements, and participants with high error loads	Comparison to a collection of datasets from multiple practices in the United States, with implausible values removed, and then errors added, and comparison to a smaller dataset independently manually curated by expert clinician opinion with consensus discussion to resolve discrepancies
Phan et al.[3]	Flagging of improbable growth values, followed by robust linear regression, with regression diagnostic flagging leading to manual inspection	Comparison to large single dataset, with manual curation using expert clinician opinion
Shi et al.[2]	Removal of largest of jack-knife studentized residuals from linear models, with removal of improbable growth values	Comparison with single large dataset with the addition of simulated outliers and implausible values

Table 2. Literature review of algorithms to remove biologically implausible data for height and weight, which had open-source code, published a verification against a gold standard, and used algorithms which took into account longitudinal data.

the intercept at the value of the weighted average. The algorithm from Shi et al.[2] compares z-scores to linear models in age, and raw values to square root models in age. Phan et al.[3] makes similar parametric assumptions about growth, assuming that both height and weight are linear functions of age. All three sets of assumptions challenge results by other authors[25,26], who have suggested that growth is better described by non-linear, non-polynomial ogive models.

Daymont et al.[1], Shi et al.[2] and Phan et al.[3] evaluated their approaches, at least in part, by running their algorithms on either a single large sample of data, or a sample formed by combining data from multiple practices in the United States. Daymont et al.[1] and Phan et al.[3] compared their results to expert clinician opinions. Daymont et al.[1] and Shi et al.[2] evaluated performance in identifying induced errors. Manual curation, as used by Daymont et al.[1] and Phan et al.[3], does not necessarily provide a gold standard. The opinions of experienced curators may provide a copper or silver, rather than gold standard. Even with the best training and experience, it is difficult to correctly remove erroneous growth records, or to confidently identify accurate measures. Adding simulated errors to existing large sample of data, as Shi et al.[2] did, provides correct identification of the errors added to the large sample of data, but not of errors that were in the original large sample of data.

Our manuscript addresses existing gaps in approach and validation methodology. The Harrall algorithms make no global parametric assumptions, and instead implement a semi-parametric approach by applying rules of local ordering and smoothness. Additionally, no manual curation is required. The validation study uses Monte Carlo simulation to create both correct and anomalous data. Thus, the validation study is able to compare the novel and published algorithms to known correct and incorrect values, permitting unbiased estimation of sensitivity and specificity.

This study had three aims. First, describe novel algorithms in SAS[27] and R[28] for computer-based detection and removal of improbable values of height and weight across childhood. Second visually illustrate the utility of the algorithm by cleaning a longitudinal sample of pediatric growth data, taken from the Exploring Perinatal Outcomes among Children (EPOCH) cohort[29]. Third, use a Monte Carlo simulation to compare estimates of sensitivity, specificity and run time (a measure of computational speed) between the Harrall algorithms, and the algorithms suggested by Daymont et al.[1], Shi et al.[2] and Phan et al.[3].

Results

The Harrall algorithm seeks to flag and remove biologically implausible values. To visually illustrate utility of the new approach, the algorithm was used to clean height trajectories for participants in the EPOCH cohort[29]. Trajectories for a single EPOCH[29] participant before (left) and after (right) cleaning appear in Fig. 1. In this example, the novel height algorithm removed one of ten observed values. The value which was removed was identified because it led to an improbable decrease in height (negative slope), followed by a rapid, and biologically implausible increase in height in a short period of time. Removing the biologically implausible value produces a smoother trajectory for the participant pictured (Fig. 1, right).

As an additional visual illustration of the utility of the Harrall algorithms, height trajectory values for all participants in EPOCH before (left) and after (right) cleaning appear in Fig. 2. The dataset includes data on 604 participants, with an average of 18 measures of height per participant (range 1–53, s.d. = 8.2). The algorithm removed data from 118 of the total 604 participants. Among participants who had points removed, each participant had an average of 19 points removed (min = 5, 25th percentile = 12, median = 17, 75th percentile = 24, max = 40). Of the points removed, 54% of the records were repeats in height across time. After cleaning, 604 participants remained, with an average of 14 longitudinal measures of height per participant (range = 1–34, s.d. = 5.7). The set of cleaned height trajectories shown in Fig. 2, on the right, shows qualitatively that the novel algorithm removes biologically implausible values, as well as rapid decreases and increases in height.

Table 3 gives sensitivity and specificity values for the Monte Carlo experiment comparing the Harrall algorithms to those of Daymont et al.[1], Shi et al.[2] and Phan et al.[3]. Sensitivity values for height show little differences between the algorithms for the experimental conditions studied, with sensitivities of around 0.99 for the algorithms of Daymont et al.[1], Harrall and Shi et al.[2] and of about 0.95 for the algorithm of Phan et al.[3]. By contrast, there is a clear ordering in specificity for height, with, in order, the algorithms of Harrall (0.928), and Daymont et al.[1] (0.903) showing much higher specificity than the algorithms of Phan et al.[3] (0.400) and Shi et al.[2] (0.332). Values for sensitivity for weight appear in a tight range around 0.99 for all algorithms considered. There are clear differences in specificity for weight, with, in order, the algorithms of Harrall (0.908), demonstrating higher values than the algorithms of Daymont et al.[1] (0.786), Shi et al.[2] (0.309) and Phan et al.[3] (0.127). Multivariate analysis of variance results demonstrate significant differences between the four algorithms for height sensitivity, height specificity, weight sensitivity, and weight specificity (all $p < 0.0001$; results not shown).

Table 4 shows results for sensitivity and specificity from the Monte Carlo experiment stratified by the proportion of participants with at least one biologically implausible value. Simulated participants had none, some or many biologically implausible values. Details on the experimental conditions appear in the Methods section. For all algorithms, for both height and weight, sensitivity and specificity values did not change appreciably with the percentage of participants with biologically implausible values, for the experimental conditions considered.

Table 5 shows average run time in seconds for each experimental condition. Run time reflects the clock time between starting the algorithm and producing values. Overall, the run time is comparable between the algorithms, with all algorithms taking seconds to run. Run times increase for the Harrall algorithm as the percentage of participants with at least one biologically implausible value increases. The increase reflects additional iterative cycles of cleaning needed to remove a higher percentage of biologically implausible values. However, the run time remains in seconds, no matter the percentage of biologically implausible values. In general, for both height and weight, the Shi et al.[2] algorithm is fastest, followed, in order, by the algorithms of Harrall, Daymont et al.[1] and Phan et al.[3].

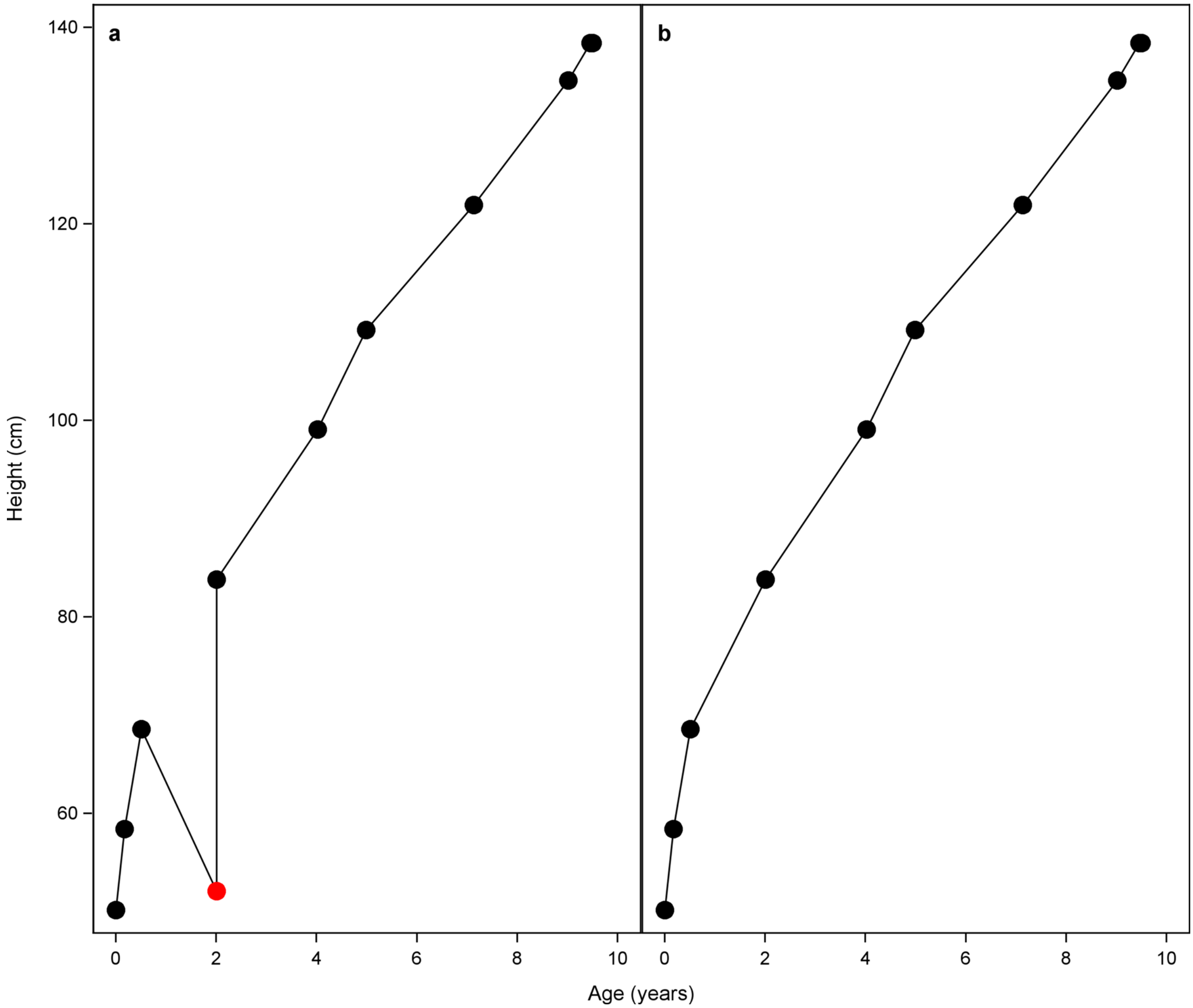

Figure 1. Height trajectory for a single participant, before (left) and after (right) cleaning with the Harrall algorithm. Biologically implausible values, which were removed are shown in red, with retained values in black.

Discussion

In this report, we provided new methods for removing biologically implausible values for height and weight in pediatric populations. We also provided free and open-source R[28] and SAS[27] code to implement the algorithms. Both a visual demonstration and a simulation study were used to evaluate the utility of the novel algorithms. The visual demonstration showed smoother height trajectories after removal of biologically implausible values with the Harrall algorithm. The simulation study compared the novel algorithm to three published algorithms[1–3]. The novel and published algorithms had similar sensitivity values for both height and weight, across all experimental conditions examined. However, the Harrall algorithms had better specificity for height, and much better specificity for weight. Run times were similar for the novel and existing algorithms. Because algorithms were compared on similar testbeds, with known correct and known biologically implausible data, it was possible to compare the algorithms head-to-head against a gold standard.

Figures 1, 2 visually demonstrate the appeal of the Harrall height algorithms. The graph of height data, with biologically implausible values removed, shows smoother spaghetti plots of height growth over development, both individually (Fig. 1) and for all participants together (Fig. 2). Because the EPOCH data were observed, and not simulated, there is no gold standard. The better-looking spaghetti plots are aesthetic improvements, but may not represent truly cleaner data.

We did not produce spaghetti plots illustrating the EPOCH[29] weight data for the following reason. Height increases monotonically with age. By contrast, weight may fluctuate between consecutive time points, and from month to month. Both increases and decreases are plausible. Looking at growth curves for height provides information, because it shows the removal of implausible decreases in height. Looking at growth curves for weight for a particular example shows little information, as small increases or decreases may reflect changes between

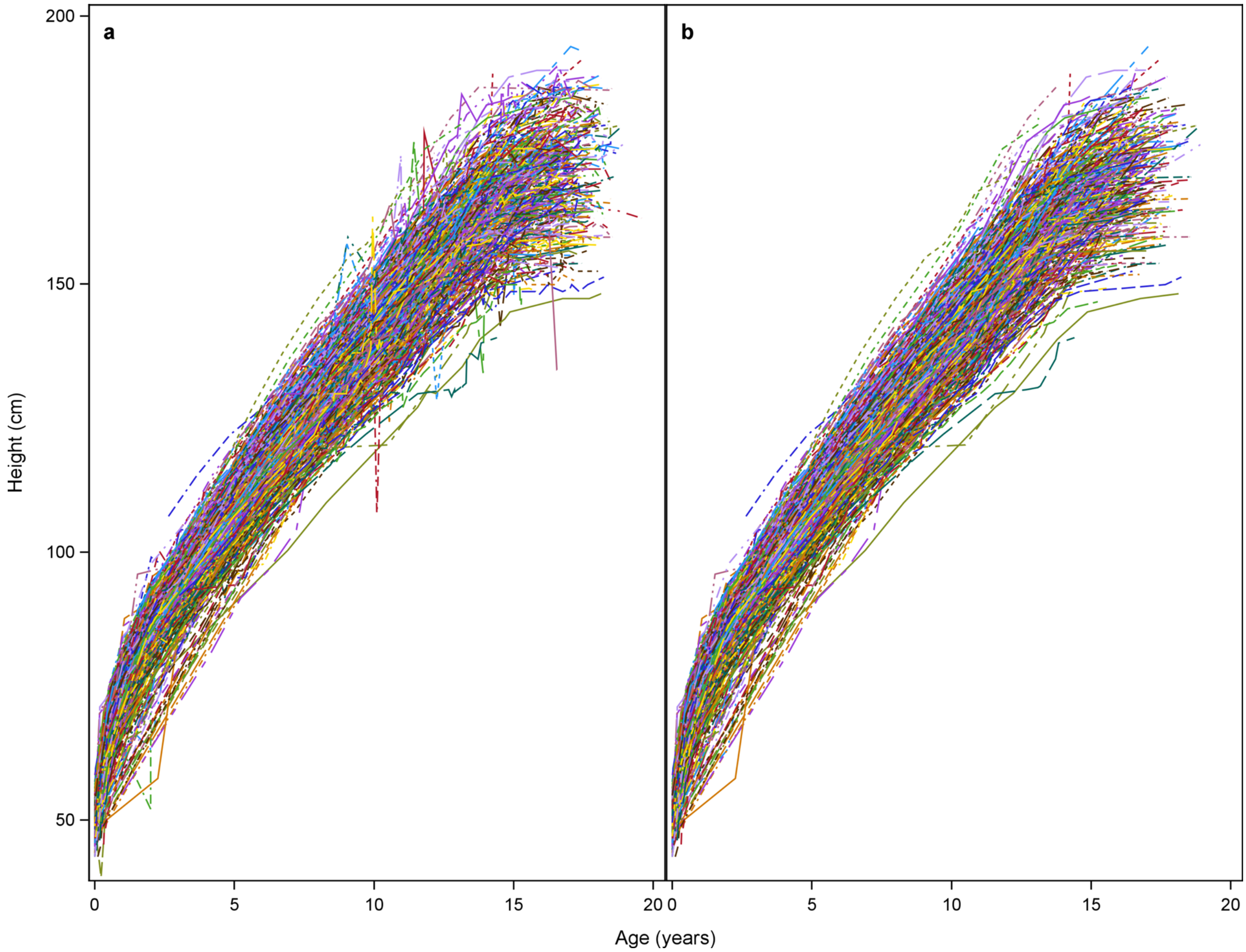

Figure 2. Spaghetti plots of repeated measures of heights from the EPOCH study, before (left) and after (right) cleaning with the Harrall algorithm. The removal of biologically implausible values appears to smooth the trajectories.

		Algorithm			
		Daymont et al.[1]	Harrall et al	Phan et al.[3]	Shi et al.[2]
Height	Average sensitivity (s.d.)	0.996 (2.91×10^{-3})	0.992 (7.46×10^{-3})	0.949 (1.85×10^{-2})	0.995 (3.70×10^{-3})
	Average specificity (s.d.)	0.903 (1.67×10^{-2})	0.928 (1.44×10^{-2})	0.400 (3.24×10^{-2})	0.332 (2.91×10^{-2})
Weight	Average sensitivity (s.d.)	0.995 (1.82×10^{-3})	0.999 (1.63×10^{-3})	0.993 (2.24×10^{-3})	0.999 (9.78×10^{-4})
	Average specificity (s.d.)	0.786 (7.07×10^{-2})	0.908 (3.15×10^{-2})	0.127 (6.85×10^{-2})	0.309 (4.46×10^{-2})

Table 3. Average sensitivity and specificity, and standard deviation (s.d.) for height and weight. The number of replicates for calculations of sensitivity was 9000: for specificity, 6000.

true measurements, or may reflect an error or more than one error in measurement. Thus, we reasoned that results from the simulation study would provide better overall information about the algorithmic performance.

The Monte Carlo study (Tables 3, 4) shows significant, but not clinically relevant differences between the novel and published algorithms in sensitivity for both height and weight. The significance of the hypothesis test results reflects the size chosen for the simulation design. Because there are 9000 replicates in the Monte Carlo study, the standard deviations are very small, and the p-values achieve significance. Differences in sensitivity in tenths or hundredths probably make no difference for scientists. The results for specificity for weight demonstrate both statistical and clinical significance. The Harrall algorithm achieved specificity of 0.908 for weight, compared to 0.309 for Shi et al.[2] and 0.127 for Phan et al.[3]. These differences are both statistically and operationally significant. The Harrall algorithm detects about nine in ten incorrect values for weight, compared to about one in three for the algorithm of Shi et al.[2], and about one in ten for the algorithm of Phan et al.[3]. For height, Harrall (0.928) and

		Error proportion	Algorithm			
			Daymont et al.[1]	Harrall et al	Phan et al.[3]	Shi et al.[2]
Height	Average sensitivity (s.d.)	None	0.997 (9.76×10^{-4})	>0.999 (2.43×10^{-4})	0.971 (3.18×10^{-3})	1.00 (0)
		Some	0.995 (2.77×10^{-3})	0.991 (5.06×10^{-3})	0.945 (9.45×10^{-3})	0.995 (1.55×10^{-3})
		High	0.994 (3.56×10^{-3})	0.986 (6.39×10^{-3})	0.931 (1.13×10^{-2})	0.992 (1.88×10^{-3})
	Average specificity (s.d.)	None				
		Some	0.906 (1.74×10^{-2})	0.931 (1.51×10^{-2})	0.386 (3.28×10^{-2})	0.322 (3.13×10^{-2})
		High	0.900 (1.52×10^{-2})	0.925 (1.31×10^{-2})	0.414 (2.51×10^{-2})	0.342 (2.26×10^{-2})
Weight	Average sensitivity (s.d.)	None	0.995 (1.74×10^{-3})	>0.999(8.63×10^{-5})	0.992 (2.25×10^{-3})	1.00 (0)
		Some	0.995 (1.84×10^{-3})	0.999 (1.86×10^{-3})	0.993 (2.19×10^{-3})	0.999 (8.45×10^{-4})
		High	0.995 (1.88×10^{-3})	0.999 (1.85×10^{-3})	0.993 (2.18×10^{-3})	0.999 (1.01×10^{-3})
	Average specificity (s.d.)	None				
		Some	0.787 (7.12×10^{-2})	0.908 (3.23×10^{-2})	0.130 (6.94×10^{-2})	0.307 (4.60×10^{-2})
		High	0.784 (7.01×10^{-2})	0.894 (3.07×10^{-2})	0.125 (6.75×10^{-2})	0.310 (4.30×10^{-2})

Table 4. Algorithm-specific average sensitivity and specificity, by proportion of people with at least one repeat, erroneous measurement, or both. Sensitivity and specificity are expressed as fractions to three significant figures. Standard deviations (s.d.) are expressed in scientific notation with three significant figures.

		Algorithm			
	Error proportion	Daymont et al.[1]	Harrall et al	Phan et al.[3]	Shi et al.[2]
Height	None	10.42 (4.39×10^{-1})	1.36 (8.07×10^{-2})	21.26 (9.35×10^{-1})	2.39 (0.128)
	Some	12.11 (5.29×10^{-1})	9.97 (6.77)	20.66 (1.48)	2.24 (0.207)
	High	12.88 (5.51×10^{-1})	11.57 (7.37)	20.38 (1.28)	2.40 (0.133)
Weight	None	8.46 (4.48×10^{-1})	1.33 (7.90×10^{-2})	21.04 (9.46×10^{-1})	2.38 (0.128)
	Some	8.74 (4.26×10^{-1})	5.06 (4.24)	20.44 (1.49)	2.24 (0.202)
	High	8.84 (4.21×10^{-1})	5.41 (4.16)	20.15 (1.30)	2.39 (0.129)

Table 5. Average timing in seconds per dataset (N participants = 100, average # observations per participant = 29.83 (s.e. = 0.0063). Timing is expressed in seconds with three significant figures. Standard deviations are expressed in scientific notation with three significant figures, in seconds. Timing reflects run time rather than processing time.

Daymont et al.[1] (0.903) algorithms are relatively close in specificity, with both having higher specificity than the algorithms of Phan et al.[3] (0.400) and Shi et al.[2] (0.332).

We used sensitivity and specificity as metrics for the simulation study. Because both sensitivity and specificity are proportions, they describe the percentage of correct values which are correctly kept, and the proportion of incorrect values which are correctly removed. For any study, the proportion of incorrect and correct values for weight and height may differ. It is impossible to know a priori how many, and what percent of each sort of measure will be removed.

Removing biologically implausible values can reduce analytic problems for modeling series of repeated measurements of height and weight. Removing large peaks and dips can increase convergence of models. Removing noise improves precision and can remove bias in model results.

The source of errors in measurements in height and weight is not clear, although one can speculate. Errors can arise from unit errors, such as measurements in pounds or inches instead of kilograms or centimeters. Errors can arise from medical staff skipping height measurements in a busy clinic. Errors can arise from wiggly children, bouncing on scales and squirming while being measured. Errors can arise from infrequent calibration of scales. Measurement errors can arise from digit preferences by clinic or study staff. And errors can arise from transcription, or translation mistakes.

Scientists must make affirmative and transparent choices as to whether to clean or not to clean data. In any observed set of data, the probability structure which generated the true data and the errors is impossible to detect. Because the probability structure is unknown, it is unclear whether removing biologically implausible values will create bias in models fit to the data. Without cleaning data, it is difficult to fit models with pediatric height or weight as predictors or outcomes. The high variance of dirty data may lead to difficulties with convergence of models. Even if modeling is possible, conclusions from hypothesis testing with dirty data may or may not match conclusions from hypothesis testing with cleaned data. If the conclusions do not match, we advocate using the clean data, because conclusions drawn from biologically implausible values may be incorrect. If the conclusions from clean and dirty data match, we still recommend using the clean data, because the lower variance of cleaned data will give tighter confidence intervals. In some cases, it may be reassuring to report the conclusions from the dirty data as a sensitivity analysis. Another possible approach is to fit measurement error models, such as the nonlinear models of Caroll et al.[30].

The manuscript used Monte Carlo simulation to generate both clean and dirty data. Using simulated data to assess algorithm performance relies on opinion. Subjective questions that must be addressed include: Is this an appropriate model for the underlying data? Are these appropriate parameters? Does the frequency and regularity of the data represent real-world data? Is this capturing not just underlying patterns of growth but typical variation (from minor illnesses, etc.)—and if not, is that important? Does the mechanism for introducing errors mimic the type, variety, and severity of errors found in real-world data? Do the features varied during simulation represent all of the important aspects of underlying data or errors that can impact algorithm importance? Sometimes the answer to one of these questions is clearly no. But there are no objective criteria to determine when the answer is yes.

The manuscript had several limitations. First, good performance on a Monte Carlo study reflects good performance on data generated under certain assumptions, and thus may not guarantee good performance on other data. In particular, the simulated data matched observed data from the EPOCH study[29] in terms of the average and variance of the number of observations per year of life. The algorithm may not have performed as well in a set of data with less frequent measurements. In addition, a Monte Carlo study can only demonstrate superiority on the experimental conditions considered, rather than all possible experimental conditions. However, we sought to mirror as accurately as possible both normal human growth, and the errors added to the data. We assumed a Preece-Baines[25] nonlinear growth model, and added repeats, cups and caps at the frequency observed in the EPOCH cohort[29]. The conditions we selected were representative of real observed data in terms of the percentage of participants with at least one repeat or biologically implausible value, the size of the errors, and the total number of errors. Second, the algorithm did not use an error tolerance to limit removal of small decreases in height. Because of errors in measurements, such decreases may be common. It is possible that future improvements of the algorithm will add this feature. Third, we used percentiles from the CDC[5] clinical growth charts for age to flag unreasonable changes in weight across time. Weight thresholds were calculated as the absolute magnitude of the difference between the 97th percentile of weight at the age one year after the index age, and the 97th percentile of weight at the index age, irrespective of the time difference between two subsequent points. The threshold used is in some sense arbitrary, as the choice of percentiles was made to remove the most extreme data. Fourth, the Harrall algorithms require at least three data points for every participant. Fifth, if there are several measurements at the same time point, the algorithm will select the first one, which may subsequently be removed if anomalous. A future version of the algorithm may permit better error handling for multiple measures at the same time point. Finally, the algorithm is not appropriate to use in cases of severe illness, when skeletal height may decrease, or when rapid weight gain or loss is expected, as after bariatric surgery, treatment with GLP1-R agonists, or in the context of severe illnesses like cancer.

The Harrall algorithms do not use population quantiles from the CDC to remove single extreme points for each participant, although the CDC weight-for-age quantiles are used to generate cutpoints for implausible growth velocity in weight. The use of CDC quantiles may represent a limitation, particularly for researchers studying populations inside of the United States. Using source data collected outside the study to define biologically implausible values is often called an externally driven cleaning approach[4]. Both the CDC[5] and the World Health Organization (WHO)[31] organizations provide percentiles of pediatric height and weight by age and sex. Our rationale for not using population quantiles to declare single points implausible was based on results from Freedman et al., who showed that this method often removed true values[32]. Freedman et al.[32] tested the validity of using WHO percentiles to remove extreme values of childhood height, weight, and BMI in the National Health and Nutrition Examination Survey (NHANES) study. The group found two major issues when assessing outliers flagged by the WHO standards. First, almost all of the outliers sat at the upper range of the measurement's distribution. By removing outliers from only the upper range of the distribution, cleaning using quantiles may lead to bias in model inference[32]. Second, at least 75% of the participants who were identified as having high biologically implausible values of height also had leg lengths above the 95th percentile[32], suggesting that the removed values may have in fact correctly represented real data from proportionally large youth.

Since the Harrall algorithm cannot identify anomalies at the first or last point of the trajectory, researchers may consider using the CDC or WHO population quantiles to flag improbable values prior to running the algorithm. In general, we discourage doing that. Our rationale is that the Harrall algorithm removes anomalous values in the context of the entire trajectory of growth for a single child. Thus, any values that remain are contextually valid, even if apparently improbably large or improbably small, in comparison to population quantiles.

The manuscript also has several strengths. The algorithm is based on very few assumptions. The algorithm for height is based on the assumption that skeletal height increases throughout childhood. In addition, the code is available, both in R[28] and SAS[27], and released as open-source, copy-left under the GNU[33] public license. This means that the code can be used or modified by others, as long as they do not charge for the software. The goal is to allow wide use of the software, and application to other datasets involving pediatric height and weight.

Another strength of the Harrall algorithms is that they are computer-based. Because computer-based algorithms are fast, they are increasingly useful in studies with large sample sizes. The increasing use of electronic medical records and the rise of large epidemiological groupings of cohorts, such as ECHO[34] and HELIX[35], means that sample sizes are increasing. Large sample sizes means that manual curation takes more time, and costs more. In addition, computer-based algorithms provide replicable results. By contrast, since manual curation is based on expert opinion, rather than an algorithm, two curators, or even the same curator making a second review may get a different answer. In addition, manual curation is sometimes conducted by reviewers with little training, and thus may not even represent expert opinion.

It is our hope that computer-based algorithmic curation of pediatric height and weight data will improve accuracy and reduce the bias in studies of growth. Understanding which data are biologically implausible will allow researchers to design systems to limit the sources of measurement error. We hope to extend the algorithmic approach to other longitudinal biological measures in humans, in which rapid changes are improbable.

Methods

Algorithm

Software

Both the height and weight algorithms are available in SAS[27] (Version 9.4, Cary, NC), and R[28]. Code and example input and output files are provided in Online Supplementary File 2, and available for free download from https:/github.com/HarrallKK/HarrallAlgorithms.

Multiple measurements at a single time point

For the algorithm, the assumption is made that only one measurement may occur at a single time point. The algorithm assesses multiple measurements that co-occur at the same timepoint, and removes all but one of the repeated measurements. If the dataset includes measurements from multiple sources, for example research visits and medical record abstraction, the algorithm can be set to preferentially select data from one source over the others. If no preference is selected, the first observation in the dataset for the timepoint in question is selected.

Identification of repeats, caps or cups

A repeat is an identical measure at two different consecutive ages. Repeats are identified when the slope between two consecutive measurements equals zero. If sequential points in the trajectory are connected by line segments, the line segment before a cup point has a negative slope, and the line segment after a cup point has a positive slope. The line segment before a cap point has a positive slope, and the line segment after a cup point has a negative slope.

Flagging of caps and cups

The algorithm cleans data for trajectories with 3 or more points. First, the algorithm identifies the sign and magnitude of the slope of each line segment. It then uses separate subroutines to flag and remove points identified as a cup, cap, or repeat. The resulting set of line segments does not usually form a line. What subroutine is used depends on the number of points in the trajectory, and the signs and magnitudes of each slope. The subroutines are designed so that the maximum absolute slope of the set of slopes computed is the smallest possible. Because there are a combinatorial number of possible subroutines, which differ based on the number of points in the trajectory, the logic of each subroutine is not described here. The open-source code allows the reader to examine, and if desired, adjust each subroutine.

Decision rule for removal of flagged values

Flagged biologically implausible values are handled differently for height and weight trajectories. Based on the assumption that children do not decrease in height, all caps and cups are removed from trajectories of height. Weight is only removed if the difference exceeds a threshold (Supplementary Table A4). Thresholds were derived using CDC[5] clinical growth charts, which specify United States population-based percentiles of weights-for-age. Thresholds were computed as the absolute differences between the 97th percentile of weight at the age one year after the index age, and the 97th percentile of weight at the index age.

Iterative cleaning

The algorithm is iterative. The first iteration identifies and removes repeats. Subsequent iterations assess for, and remove, flagged caps, cups, and additional repeats that arise during the cleaning process. The iterative process ends when a maximum number of iterations is completed, or when no further caps, cups, or repeats exceed detection thresholds.

EPOCH study

Cohort participants were the offspring of singleton pregnancies, born at a single hospital in Denver, Colorado, USA, between 1992 and 2002, who had biological mothers who were members of the Kaiser Permanente of Colorado Health plan. Height and weight values were obtained from medical record abstraction, and from two in-person research visits at approximately 10 and 16 years of age, at which height and weight were measured in light clothing and without shoes by trained staff using standardized, calibrated instruments. The height and weight values thus represented a mix of two approaches for obtaining anthropometric data. The study was approved by the Colorado Multiple Institutional Review Board. Caregivers provided written informed consent and children provided written assent before entering the study. All procedures were performed following the relevant guidelines and regulations.

Data simulation methods

Aims, design and number of replicates

Monte Carlo simulation methods were reported using an approach suggested by Morris et al.[36]. The aim of the simulation was to compare average sensitivity and specificity for each algorithm on a testbed with known correct and incorrect values. A further aim was to evaluate whether the sensitivity and specificity of the algorithms varied with two factors: (1) the proportion of errors, including cups, caps and repeats, and (2) the relative position of the population height or weight curve within the percentiles of the CDC[5] clinical growth charts. Three levels were chosen for each factor, for a total of 9 = 3* 3 cross-classified levels. The levels are described below. With 1000 replicates simulated for each cross classified level, the sample size for the experiment was 9000 = 9*1000. Using 1,000 realizations made the half width of the asymptotic 95% confidence interval for estimates of sensitivity and specificity no more than 0.031.

Simulation of correct height and weight data

Correct data for both height and weight was generated by using the Preece-Baines[25] model, with three parameter sets described in Supplemental Table A1. Random individual deviations from the population line were generated using the SAS[27] software *normal* function. The number and spacing of the data per year of life were simulated to mirror the mean and standard deviation of the number and spacing of the EPOCH data, shown in Supplemental Table A2.

Inclusion of known biologically implausible values

The probability of occurrence, number and magnitude of the biologically implausible values were randomly varied, as a function of input parameters described in Supplemental Table A3. Caps and cups were included by adding (or subtracting) values from simulated correct values. Repeats were included by substituting repeat values for previously simulated correct values.

The proportions of errors

In order to mirror errors in an observed data set, we conducted the following analysis of the EPOCH[29] data for both height and weight. We used the Harrall algorithms to clean both height and weight data. Comparing the resulting cleaned and deleted values allowed estimation of the probabilities of errors shown in Supplemental Table A3.

The three levels considered in the simulation experiment corresponded to adding no errors, making the probability of at least one error of each sort equivalent to that derived from the EPOCH[29] study and shown in Supplemental Table A3, and doubling the probability of at least one of each sort of error. The number of errors, given that person had at least one error, and the size of the errors added was held constant.

The relative position of the true height and weight curves

The three sets of parameters for the Preece-Baines[25] models shown in Supplemental Table A1 were chosen so that the resulting family of curves stayed within the bounds created by the 5th and 95th quantile curves of the CDC[5] clinical growth charts, and represented relatively low (i.e., closer to the 5th percentile than the median), medium (i.e., close to the median), and high (i.e., closer to the 95th percentile than the median) heights or weights respectively.

Sensitivity and specificity

Sensitivity and specificity were computed using the formulae in Table 6. Specificity was only computed if the study participant had at least one biologically implausible value. The denominator of specificity is the number of biologically implausible values: if this were zero, specificity cannot be calculated, as division by zero is not allowed. *Average sensitivity and specificity* were computed by averaging the sensitivity and specificity computed for each replication of the experiment.

Timing

Total run time was assessed on a Dell 5280 Precision tower, with an Intel® Xeon®W 2123 central processing unit, 3.60 GHz processor with 32.0 GB of RAM (31.7 GB usable), running 64-bit Windows 10 Enterprise.

Statistical methods

We used multivariate analysis of variance (MANOVA) to assess whether the novel approach differed from the three previously published approaches[1–3] on six outcomes: (1) sensitivity for height, (2) specificity for height, (3) sensitivity for weight, (4) specificity for weight, (5) processing time for cleaning height and 6) processing time for cleaning weight. We assessed normality of the jackknifed studentized residuals. Each replicate of the Monte Carlo experiment was treated as an independent sampling unit. The total sample size for assessing sensitivity is 9000. Because the first factor of the experiment introduces no errors, the total sample size for assessing specificity is 6000.

Each separate general linear multivariate model used the respective measures for the four algorithms as outcomes, and an intercept as the only predictor. This gave the sample averages as the parameter estimates. We

		Truth	
		Correct value	**Incorrect or anomalous value**
Algorithmic result	Value declared correct	N_{11}	N_{12}
	Value declared incorrect	N_{21}	N_{22}

Table 6. Errors and successes in detection for a cleaning algorithm, either overall, or per participant. Sensitivity is defined as $N_{11}/(N_{11} + N_{21})$. Specificity is defined as $N_{22}/(N_{12} + N_{22})$.

used the Hotelling-Lawley trace statistic to conduct two-sided tests at Bonferroni-corrected Type I error rate of 0.05/6 = 0.0083.

Data availability

The code to generate the data and the results are available at https:/github.com/HarrallKK/HarrallAlgorithms. Interested researchers who wish to compare their approaches to ours, or to other published approaches are encouraged to cite this manuscript, and freely use the code. De-identified pediatric height and weight data from the EPOCH study may be made available, upon reasonable request, and after review of EPOCH publications committee, by inquiry to Anna.Bellatorre@cuanschutz.edu.

Received: 7 February 2024; Accepted: 1 August 2024
Published online: 06 August 2024

References

1. Daymont, C. *et al.* Automated identification of implausible values in growth data from pediatric electronic health records. *J. Am. Med. Inform. Assoc.* **24**, 1080–1087 (2017).
2. Shi, J., Korsiak, J. & Roth, D. E. New approach for the identification of implausible values and outliers in longitudinal childhood anthropometric data. *Ann. Epidemiol.* **28**, 204-211.e3 (2018).
3. Phan, H. T. T. *et al.* Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: Protocol and application to a large patient cohort. *Sci. Rep.* **10**, 10164 (2020).
4. Lawman, H. G. *et al.* Comparing methods for identifying biologically implausible values in height, weight, and body mass index among youth. *Am. J. Epidemiol.* **182**, 359–365 (2015).
5. Centers for Disease Control. *Growth Charts - 2000 CDC Growth Charts - United States*. https://www.cdc.gov/growthcharts/cdc_charts.htm (2022).
6. Hockett, C. W., Harrall, K. K., Glueck, D. H. & Dabelea, D. M. Exposure to gestational diabetes and BMI trajectories through adolescence: Exploring Perinatal Outcomes in Children study. *J. Clin. Endocrinol. Metab.* https://doi.org/10.1210/clinem/dgad278 (2023).
7. Bekelman, T. A. *et al.* Adherence to index-based dietary patterns in childhood and BMI trajectory during the transition to adolescence: The EPOCH study. *Int. J. Obes. (London)* **45**, 2439–2446 (2021).
8. Moore, B. F., Harrall, K. K., Sauder, K. A., Glueck, D. H. & Dabelea, D. Neonatal adiposity and childhood obesity. *Pediatrics* **146**, e20200737 (2020).
9. Hockett, C. W. *et al.* Persistent effects of in utero overnutrition on offspring adiposity: The Exploring Perinatal Outcomes among Children (EPOCH) study. *Diabetologia* **62**, 2017–2024 (2019).
10. Kim, C., Harrall, K. K., Glueck, D. H. & Dabelea, D. Sex steroids and adiposity in a prospective observational cohort of youth. *Obes. Sci. Pract.* **7**, 432–440 (2021).
11. Cohen, C. C. *et al.* Body composition trajectories from birth to 5 years and hepatic fat in early childhood. *Am. J. Clin. Nutr.* **116**, 1010–1018 (2022).
12. World Health Organization Expert Committee. Physical status: The use and interpretation of anthropometry. Report of a WHO Expert Committee. *World Health Organ. Tech. Rep. Ser.* **854**, 1–452 (1995).
13. Field, A. E. *et al.* Relation between dieting and weight change among preadolescents and adolescents. *Pediatrics* **112**, 900–906 (2003).
14. Lobstein, T. J., James, W. P. T. & Cole, T. J. Increasing levels of excess weight among children in England. *Int. J. Obes.* **27**, 1136–1138 (2003).
15. National Health and Nutrition Examination Survey. *NHANES 2001–2002: Body measures data documentation, codebook, and frequencies*. https://wwwn.cdc.gov/nchs/nhanes/2001-2002/BMX_B.htm (2004).
16. Conde, W. L. & Monteiro, C. A. Body mass index cutoff points for evaluation of nutritional status in Brazilian children and adolescents. *J. Pediatr. (Rio J)* **82**, 266–272 (2006).
17. Smith, N. *et al.* Body weight and height data in electronic medical records of children. *Int. J. Pediatr. Obes.* **5**, 237–242 (2010).
18. Youth Risk Behavior Surveillance System. 2013 YRBS data user's guide. https://www.cdc.gov/healthyyouth/data/yrbs/files/2013/pdf/yrbs_2013_national_user_guide.pdf (2012).
19. Centers for Disease Control and Prevention. Cut-offs to define outliers in the 2000 CDC growth charts (2014).
20. Lo, J. C. *et al.* Prevalence of obesity and extreme obesity in children aged 3–5 years. *Pediatr. Obes.* **9**, 167–175 (2014).
21. Kim, J. *et al.* Incidence and remission rates of overweight among children aged 5 to 13 years in a district-wide school surveillance system. *Am. J. Public Health* **95**, 1588–1594 (2005).
22. Sturm, R. & Datar, A. Body mass index in elementary school children, metropolitan area food prices and food outlet density. *Public Health* **119**, 1059–1068 (2005).
23. Lawman, H. G. *et al.* Trends in relative weight over one year in low-income urban youth. *Obesity (Silver Spring)* **23**, 436–442 (2015).
24. Yang, S. & Hutcheon, J. A. Identifying outliers and implausible values in growth trajectory data. *Ann. Epidemiol.* **26**, 77-80.e2 (2016).
25. Preece, M. A. & Baines, M. J. A new family of mathematical models describing the human growth curve. *Ann. Hum. Biol.* **5**, 1–24 (1978).
26. Cole, T. J., Donaldson, M. D. C. & Ben-Shlomo, Y. SITAR—a useful instrument for growth curve analysis. *Int. J. Epidemiol.* **39**, 1558–1566 (2010).
27. SAS Institute Inc. *SAS® 9.4 Language Reference: Concepts* 6th edn. (SAS Institute Inc., 2016).
28. R Core Team. R: A Language and Environment for Statistical Computing. https://www.r-project.org (2023).
29. Crume, T. L. *et al.* Association of exposure to diabetes in utero with adiposity and fat distribution in a multiethnic population of youth: the Exploring Perinatal Outcomes among Children (EPOCH) Study. *Diabetologia* **54**, 87–92 (2011).
30. Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. *Measurement Error in Nonlinear Models: A Modern Perspective* (CRC Press, 2006).
31. The WHO Child Growth Standards. https://www.who.int/tools/child-growth-standards/standards.
32. Freedman, D. S. *et al.* Validity of the WHO cutoffs for biologically implausible values of weight, height, and BMI in children and adolescents in NHANES from 1999 through 2012. *Am. J. Clin. Nutr.* **102**, 1000–1006 (2015).
33. Free Software Foundation. The GNU Public License. https://www.gnu.org/licenses/gpl-3.0.en.html (2023).
34. Gillman, M. W. & Blaisdell, C. J. Environmental influences on Child Health Outcomes, a Research Program of the National Institutes of Health. *Curr. Opinion Pediatr.* **30**, 260 (2018).
35. Vrijheid, M. *et al.* The human early-life exposome (HELIX): Project rationale and design. *Environ. Health Perspect.* **122**, 535–544 (2014).

36. Morris, T. P., White, I. R. & Crowther, M. J. Using simulation studies to evaluate statistical methods. *Stat. Med.* **38**, 2074–2102 (2019).

Acknowledgements

KKH, DHG and KEM were supported by NIH/NIGMS funded grants 5R01GM121081–08, 3R25GM111901 and 3R25GM111901-04S1. DD was supported by NIH/NIGMS funded grant 5R01GM121081–08. The pediatric height data described in Figures 1 and 2 was collected during R01DK068001, for which DD was the principal investigator. The views expressed in this manuscript cannot be inferred to be those of the NIGMS or the NIH.

Author contributions

KKH conceptualized the data cleaning algorithm, wrote the prototype code, and directed the project. SB produced open-source, commented macro code for running the algorithm, and wrote the simulation to compare algorithms. KEM provided informative feedback on early drafts of the manuscript. LAV assisted with running R code to implement one of the published comparator algorithms, and assisted with the translation of the algorithm into R. MEP assisted with the review of the literature. AB assisted in revising the Stata code to implement one of the published comparator algorithms. DD provided mentorship, and served as the principal investigator for the EPOCH study, the source of the inputs for the simulations. DHG edited the manuscript, and directed the project. Funding for the project was provided by grants led by DD, KEM, and DHG. All authors critically reviewed the manuscript, and approved its submission.

Funding

National Institute of General Medical Sciences,5R01GM121081–08,5R01GM121081–08,5R01GM121081–08,5R01GM121081–08,National Institutes of Health,R01DK068001

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/s41598-024-69161-5.

Correspondence and requests for materials should be addressed to K.K.H.

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.