Annals of Epidemiology xxx (2018) 1–8

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

# New approach for the identification of implausible values and outliers in longitudinal childhood anthropometric data

Joy Shi, MSc [a], Jill Korsiak, MSc [a], Daniel E. Roth, MD, PhD [a, b, *]

[a] Centre for Global Child Health and SickKids Research Institute, Hospital for Sick Children, Toronto, ON, Canada
[b] Department of Pediatrics, Hospital for Sick Children and University of Toronto, Toronto, ON, Canada

ARTICLE INFO

*Article history:*
Received 6 June 2017
Accepted 8 January 2018
Available online xxx

*Keywords:*
Longitudinal growth data
Jackknife residuals
Biologically implausible values
Outliers

ABSTRACT

*Purpose:* We aimed to demonstrate the use of jackknife residuals to take advantage of the longitudinal nature of available growth data in assessing potential biologically implausible values and outliers.
*Methods:* Artificial errors were induced in 5% of length, weight, and head circumference measurements, measured on 1211 participants from the Maternal Vitamin D for Infant Growth (MDIG) trial from birth to 24 months of age. Each child's sex- and age-standardized z-score or raw measurements were regressed as a function of age in child-specific models. Each error responsible for a biologically implausible decrease between a consecutive pair of measurements was identified based on the higher of the two absolute values of jackknife residuals in each pair. In further analyses, outliers were identified as those values beyond fixed cutoffs of the jackknife residuals (e.g., greater than +5 or less than −5 in primary analyses). Kappa, sensitivity, and specificity were calculated over 1000 simulations to assess the ability of the jackknife residual method to detect induced errors and to compare these methods with the use of conditional growth percentiles and conventional cross-sectional methods.
*Results:* Among the induced errors that resulted in a biologically implausible decrease in measurement between two consecutive values, the jackknife residual method identified the correct value in 84.3% −91.5% of these instances when applied to the sex- and age-standardized z-scores, with kappa values ranging from 0.685 to 0.795. Sensitivity and specificity of the jackknife method were higher than those of the conditional growth percentile method, but specificity was lower than for conventional cross-sectional methods.
*Conclusions:* Using jackknife residuals provides a simple method to identify biologically implausible values and outliers in longitudinal child growth data sets in which each child contributes at least 4 serial measurements.

## Introduction

Child growth in stature and body dimensions is a continuous and dynamic process that is optimally studied through longitudinal follow-up. As such, many epidemiologic studies are designed to collect repeated anthropometric measures of each child across successive time points to describe growth patterns, identify predictors, and assess associations with later health outcomes [1–3]. Standardized procedures for the measurement and collection of anthropometry data have been established to maximize data quality [4,5], and advanced analytical strategies are employed to accommodate the longitudinal nature of the data collected in such studies [6]; similarly, rigorous quality control procedures during data cleaning are warranted to identify outliers and implausible values.

For cross-sectional studies in which each child only contributes one set of measurements, identification of outliers and implausible values is limited to the use of fixed cutoffs that were derived from a reference population, such as those established for the WHO Child Growth Standards [7] or from the observed distributional properties of the study population. However, with serial anthropometric measurements collected in the same individual, the growth trajectory of each individual provides another basis upon which to assess the biological plausibility of any given measurement for that individual. For example, a decrease in length or height between two successive time points is biologically implausible and thus indicates

* Corresponding author. Hospital for Sick Children, 686 Bay Street, Toronto, ON, Canada M5G0A4.
*E-mail address:* daniel.roth@sickkids.ca (D.E. Roth).

https://doi.org/10.1016/j.annepidem.2018.01.007
1047-2797/Crown Copyright © 2018 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

that at least one of the values was incorrectly measured or recorded. However, without additional information, it is often challenging to determine which of the two values induced the implausible trajectory. Many longitudinal studies use only conventional cross-sectional approaches to identify outliers; when the longitudinal nature of the data is taken into consideration, the methods used to identify these implausible values are often not described, or exclusions are made on a case-by-case basis [8].

Yang et al. [9] recently described the application of conditional growth percentiles for systematically identifying outliers and implausible values in longitudinal childhood anthropometric data. In their example, a hierarchical model of serial weight measurements as a function of age was constructed to estimate an individual's weight percentile at time *t*, while conditioning on the individual's weight percentile at time $t{-}1$. This approach implies that the plausibility of a given measurement is solely based on the preceding measurement (and therefore cannot be applied to an individual's first measurement), yet the expected value is contingent on the overall growth trajectory of the study population.

We propose an alternative longitudinal approach to assess for outliers and implausible values using growth trajectories that are fit to each individual's anthropometric data, in which the identification of outliers and implausible values does not depend on the distribution of measurements in the whole study population. In this study, we aimed to use a longitudinal growth data set with artificially induced errors to demonstrate how jackknife residuals of linear models of z-scores or raw growth data as a function of age can be used to identify biologically implausible values and outliers. We further aimed to compare the sensitivity and specificity of jackknife residuals to alternative methods of data cleaning, including conditional growth percentiles and conventional cross-sectional approaches.

## Methods

### *Data source*

Anthropometric data collected from infants enrolled in the Maternal Vitamin D for Infant Growth (MDIG) trial were used. Data collection including anthropometry is ongoing; therefore, data available up to January 26, 2017 were used for this study. The MDIG trial methods have been previously described [10]. In brief, the MDIG trial is a randomized placebo-controlled dose-ranging trial of vitamin D supplementation during pregnancy and lactation in Dhaka, Bangladesh. Length, weight, and head circumference measurements are scheduled in tri-monthly intervals from birth to 24 months of age, plus an additional measurement taken at 2, 4, 6, or 8 weeks of age chosen through random assignment, with some variability between infants in actual timing of measurement collection. Age- and sex-standardized z-scores for length, weight, and head circumference measurements were generated using a combination of growth references: the Intergrowth-21st Newborn Size standards; the Intergrowth-21st International Postnatal Growth Standards for Preterm Infants; and the World Health Organization (WHO) Child Growth Standards (Supplementary Material Methods). Data cleaning of any natural occurring errors in the data set was not conducted to preclude biasing the results in favor of one method over another.

### *Simulation of implausible values and outliers*

Simulated outliers and implausible values were randomly generated in the existing data set by deliberate introduction of data errors. In primary analyses, we randomly selected 5% of all encounters for error induction, in which values were randomly shifted upward or downward in relation to the original observed values. The magnitude of the errors (i.e. differences between original and shifted values) followed a normal distribution of mean = 0 with a standard deviation based off of the derived standard deviation of the raw anthropometric measurements in the WHO Multicentre Growth Reference Study that is the basis for the WHO child growth standards [7]. As such, the standard deviation of the errors varied by type of measurement, sex, and age of the infant. In sensitivity analyses, we used error rates of 10% and 15%, and standard deviations of 2- and 3-times the age- and sex-specific standard deviation for the corresponding anthropometric measure. A Monte Carlo approach was used, in which each scenario for the varying error rates and error generation methods was simulated 1000 times.

### *Identifying biologically implausible values*

Jackknife residuals were applied to identify the incorrect measurement in instances where errors introduced into the anthropometric data set resulted in a biologically implausible decrease in a child's length, weight, or head circumference from one time point to the next. Jackknife (or externally studentized) residuals, $r_{(-i)}$, are generated from regression residuals, $e_i$, that are scaled by a function of the mean squared error with the $i^{\text{th}}$ observation deleted, $MSE_{(-i)}$, and the leverage, $h_i$:

$$r_{(-i)} = \frac{e_i}{\sqrt{MSE_{(-i)}(1-h_i)}} \tag{1}$$

Jackknife residuals are expected to follow a *t* distribution with $(n{-}k{-}2)$ degrees of freedom, where *n* is the number of observations and *k* is the number of parameters in the fitted model, thereby giving the distribution a mean of 0 and standard deviation slightly greater than 1. Given that the $k^{th}$ observation is an outlier, the jackknife residuals of other observations will shrink toward zero due to an overestimation of MSE, whereas the jackknife residual of the $k^{th}$ observation will not. As such, jackknife residuals respond more strongly to the presence of a single outlier than does the standardized residual [11].

Any decrease in raw length or head circumference measurements was considered to be biologically implausible, whereas a decrease of greater than 15% in the raw measurements for weight was considered biologically implausible. This analysis addressed instances in which an error in the data set could be clearly identified, but the exact time point at which the error occurred was not as easily discerned. Therefore, these analyses were limited to children for whom there was at least one implausible decrease induced in anthropometric measures.

All individuals with a biologically implausible decrease between any two measurement time points were first identified based on the criteria listed previously. Separately for each child, linear regression was used to fit a straight line through the individual's sex- and age-standardized z-score of the corresponding anthropometry measurement as a function of age:

$$Z_{ij} = \beta_{0i} + \beta_i \cdot t_{ij} + \varepsilon_{ij} \tag{2}$$

where "$i$" denotes the $i^{\text{th}}$ individual and "$j$" denotes the $j^{\text{th}}$ time point. For raw measurements, each individual's measurements were regressed on the square root of age ($t^{½}$) to model a curvilinear relationship in which growth rates vary with age [12]:

$$Y_{ij} = \beta_{0i} + \beta_i \cdot t_{ij}^{1/2} + \varepsilon_{ij} \tag{3}$$

Each measurement of a given individual is assessed for adequate fit to the modeled trajectory for that individual using jackknife