*JAMIA Open*, 5(4), 2022, 1–9
https://doi.org/10.1093/jamiaopen/ooac089
Research and Applications

Research and Applications

# Cleaning of anthropometric data from PCORnet electronic health records using automated algorithms

**Pi-I D. Lin[1], Sheryl L. Rifas-Shiman[1], Izzuddin M. Aris[1], Matthew F. Daley[2], David M. Janicke[3], William J. Heerman[4], Daniel L. Chudnov[5], David S. Freedman[6], and Jason P. Block[1]**

[1]Division of Chronic Disease Research Across the Lifecourse (CoRAL), Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, Massachusetts, USA, [2]Institute for Health Research, Kaiser Permanente Colorado, Aurora, Colorado, USA, [3]Department of Clinical and Health Psychology, University of Florida, Gainesville, Florida, USA, [4]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, Tennessee, USA, [5]MITRE Corporation, McLean, Virginia, USA and [6]Division of Nutrition, Physical Activity, and Obesity, Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Corresponding Author: Pi-I D. Lin, ScD, MS, Division of Chronic Disease Research Across the Lifecourse (CoRAL), Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Suite 401, Boston, MA 02215, USA; p_lin@harvardpilgrim.org

Received 27 July 2022; Revised 30 September 2022; Editorial Decision 5 October 2022; Accepted 14 October 2022

## ABSTRACT

**Objective:** To demonstrate the utility of *growthcleanr*, an anthropometric data cleaning method designed for electronic health records (EHR).
**Materials and Methods:** We used all available pediatric and adult height and weight data from an ongoing observational study that includes EHR data from 15 healthcare systems and applied *growthcleanr* to identify outliers and errors and compared its performance in pediatric data with 2 other pediatric data cleaning methods: (1) conditional percentile (*cp*) and (2) PaEdiatric ANthropometric measurement Outlier Flagging pipeline (*peanof*).
**Results:** 687 226 children (<20 years) and 3 267 293 adults contributed 71 246 369 weight and 51 525 487 height measurements. *growthcleanr* flagged 18% of pediatric and 12% of adult measurements for exclusion, mostly as carried-forward measures for pediatric data and duplicates for adult and pediatric data. After removing the flagged measurements, 0.5% and 0.6% of the pediatric heights and weights and 0.3% and 1.4% of the adult heights and weights, respectively, were biologically implausible according to the CDC and other established cut points. Compared with other pediatric cleaning methods, *growthcleanr* flagged the most measurements for exclusion; however, it did not flag some more extreme measurements. The prevalence of severe pediatric obesity was 9.0%, 9.2%, and 8.0% after cleaning by *growthcleanr*, *cp*, and *peanof*, respectively.
**Conclusion:** *growthcleanr* is useful for cleaning pediatric and adult height and weight data. It is the only method with the ability to clean adult data and identify carried-forward and duplicates, which are prevalent in EHR. Findings of this study can be used to improve the *growthcleanr* algorithm.

**Key words:** electronic health record, automatic data processing, body height, body weight, PCORnet, big data

**LAY SUMMARY**

Anthropometric data from electronic health records (EHR) provides straightforward access to large amounts of data for conducting clinical research. However, data errors can potentially bias study results. In this study, we demonstrate the utility of an automated anthropometric data cleaning method—*growthcleanr*—on more than 12 million pediatric and adult height and weight measurements. Using *growthcleanr*, we flagged 18% of pediatric and 12% of adult measurements for exclusion, mostly because measures were inappropriately carried forward from prior visits for children or were duplicates for both adults and children. After removing the flagged measurements, 0.5% and 0.6% of the pediatric heights and weights and 0.3% and 1.4% of the adult heights and weights, respectively, were considered biologically implausible according to established cut points. We compared the results of *growthcleanr* for pediatrics to other existing pediatric data cleaning methods; *growthcleanr* flagged the most measurements for exclusion but failed to flag some more extreme measurements because *growthcleanr* has a broad inclusion criterion that can be refined for every study. Our study showed that growthcleanr is useful for cleaning pediatric and adult height and weight data. It is the only method with the ability to clean adult data and identify measures that are carried-forward and duplicated, both of which are common issues with EHR data. Findings of this study can be used to improve the *growthcleanr* algorithm.

## INTRODUCTION (BACKGROUND AND SIGNIFICANCE)

Electronic health record (EHR) data are increasingly available in clinical epidemiologic and population health surveillance research.[1] The availability of objectively measured information on large population can allow investigators to conduct clinical research more easily than when using traditional primary data collection methods. Unfortunately, data errors are common and can bias study results.[2] For example, previous studies have identified common errors in anthropometric measures documented in EHRs that lead to implausible values, including unit errors, recording wrong digits (adding, leaving out), and swapping height and weight measurements.[3] Therefore, ensuring data quality of EHRs is a key foundation for research using these secondary data, and identifying and correcting such errors is critical to producing robust and unbiased findings.

Traditional approaches for identifying and removing implausible values in large EHR datasets involve trimming outliers or implausible values based on absolute measurements, ranges observed in the study dataset, or percentiles in reference to a population growth standard for pediatric data.[4,5] While these methods are simple to perform and can remove a large number of errors, they can incorrectly remove extreme values (eg, outliers) that are valid or include erroneous values for an individual that are within the cutoff range.[6] To overcome these limitations, trajectory-based methods, which evaluate a series of height and weight measures per individual over time using statistical models, were developed to identify implausible values or errors based on analysis of longitudinal data[7–9]; these algorithms, however, were designed only for pediatric data. The "*growthcleanr*" tool, an automated algorithm developed by Daymont et al,[7,10,11] uses an exponentially-weighted moving average (EWMA) to identify outliers. It was recently expanded to accommodate adult data.[12] No study has used the adult algorithm for a real-world application, and limited information is available on the comparison of *growthcleanr* to other commonly-used pediatric data cleaning methods.

The objective of this study was to use *growthcleanr* to identify outliers from height and weight trajectories and errors, such as unit errors, wrong digits, duplicates and carried-forward values, in pediatric and adult EHR height and weight measurements and compare *growthcleanr* with other commonly used pediatric data cleaning algorithms for population health research. The data used are from the National Patient-Centered Clinical Research Network (PCORnet) MedWeight Study, a large, multi-institutional study conducted in a national network.

## MATERIALS AND METHODS

### Materials

**PCORnet clinical research network and PCORnet MedWeight study**

PCORnet is a distributed research network with more than 60 participating healthcare systems in the United States; it facilitates interoperability of data across institutions by use of a Common Data Model (CDM), in which healthcare systems all organize their clinical data in the same format.[13,14] In 2020, PCORnet had EHR data from 337 hospitals, 3564 primary care practices, 338 emergency departments, and 1024 community clinics. Between 2009 and 2019, > 60 million individuals contributed data with an average of 2.63 years of follow-up.[13] The PCORnet MedWeight Study is an ongoing observational study which has captured data from 15 prior and currently enrolled healthcare institutions of PCORnet to examine prescription medication-induced weight gain associated with the use of 5 classes of medication: (1) antihypertensives, (2) antiepileptics, (3) diabetes medications, (4) antidepressants, and (5) antipsychotics. The MedWeight Study provided a unique opportunity to access millions of clinical height and weight measurements from diverse health care institutions across the United States. These 5 medications were chosen because some subclasses of the medications had been reported to be associated with weight gain, and medication-induced weight gain is a common cause of medication nonadherence.[15] This analysis included all available height and weight measurements in the EHR from children and adults who used any of the 5 classes of medications from January 1, 2009 to June 30, 2020. The Institutional Review Boards from Harvard Pilgrim Health Care Institute approved the study.

**growthcleanr algorithm**

*growthcleanr* is an automated R package[16] designed for data cleaning of secondary data from EHRs. It incorporates algorithms that analyze longitudinal height and weight data and identify implausible values based on patient-specific trajectory analyses using EWMA and the US CDC growth reference (https://github.com/carriedaymont/growthcleanr).[7,12] Before flagging outliers using the trajectory method, the algorithm first performs removing biologically implausible values and identifies common errors in EHR anthropometric data, such as measurements that have swapped height and weight measurements, wrong units, carried-forwards, and duplicates. Carried-forwards are identical values carried forward from the initial value of another measurement date rather than re-measured. This could occur in EHR when a height or weight measurement was needed but not re-measured. Carried-forward values can introduce bias in estimating the growth velocity among pediatric patients. Duplicates are duplicated values on the same day, with same values. Extraneous-same-day measurements are measurements of the same day that varied by a trivial amount (eg, <1 cm or 1 kg) or measurements that repeatedly occurred (see Daymont et al[7] for more detailed criteria). These measurements could introduce bias by adding noise to measurements or changing the weighting of particular time points in the trajectory modeling using EWMA (see Daymont et al[7] for a more detailed description of cleaning steps and exclusion flags). In a previous validation study,[7] the performance of the pediatric algorithm was compared with physician review, and the accuracy of the cleaning was tested with simulated errors. Briefly, the physician review involved 2 clinical experts who regularly evaluate pediatric growth trajectory; they reviewed and marked implausible values from plotted curves of all height and weight measurements of randomly selected patients. Simulated errors were generated by introducing errors, such as unit errors, switch errors, duplicates, and carried-forward errors, to a clean dataset. The validation study demonstrated that compared with physician judgment as the gold standard, *growthcleanr* had 97% (95% CI, 94–99%) sensitivity and 90% (95% CI, 85–94) specificity for flagging implausible values; it flagged 95% and 98% of the simulated errors for weight and height measurements, respectively. Validation study of the adult algorithm is currently underway and has not been published.

### Other cleaning algorithms

Currently, there is no other available anthropometric data cleaning algorithm for adult height and weight; therefore, we only examined the results from *growthcleanr* for adults. For pediatric data, we compared the performance of *growthcleanr* to 2 other available pediatric data cleaning algorithms: conditional percentiles (*cp*)[9] and PaEdiatric ANthropometric measurement Outlier Flagging pipeline (*peanof*)[8] (see Table 1 for comparison of the 3 methods). The *cp* algorithm identifies implausible values based on a child's growth trajectory. This algorithm calculates a conditional mean and variance for each weight or height measurement and flags outliers ± 4 standard deviation (SD) from the expected value. The algorithm is flexible and has been used to clean pediatric growth data in many prior studies.[17–19] The *cp* method is available in STATA (see appendix of Yang and Hutcheon[9]).

The *peanof* algorithm is an automated method that uses the WHO growth guideline to identify implausible values and utilizes an algorithm that checks implausible increments and decrements in longitudinal data using a robust linear regression method to flag outliers (typically ± 2 SD); other errors, such as including unit errors, are also flagged (https://github.com/hangphan/peanof/). The method has been validated against manually curated results by clinicians[8]; the sensitivity of correctly identifying outlier weights was 90.9% and the error rate was 2.4% (plausible weight identified by clinicians but flagged as outliers by *peanof*).

## Methods

### Data cleaning steps

From the database of patients in the PCORnet MedWeight Study ($N = 4\ 051\ 139$), we extracted subject ID, study site, sex, age in days, height (cm), and weight (kg) for the data cleaning process. We followed the required data preparation steps by *growthcleanr* (https://carriedaymont.github.io/growthcleanr/articles/quickstart.html) and used the following parameter values: 20 years for "adult cutoff age" (age to switch to adult cleaning algorithm) and 400 lb for "weight cap", which is a *growthcleanr* function that addresses artifacts introduced by EHR vendors that might have a cap (or restriction) on weight data entry; 400 lb is a common data cap used by EHR vendors. If subject has at least one weight equal to the weight cap, *growthcleanr* either excludes that weight or all measurements of the subject depending on whether the weight is in line with other weights of the subject. We also followed the instruction for running large datasets (https://carriedaymont.github.io/growthcleanr/articles/large-data-sets.html) and split the data into smaller datasets to improve computing performance. *Growthcleanr* (version 2.0.0) was then run using "batch mode" on a computing cluster (Harvard Medical School O2 Computing Cluster).

After cleaning the data with *growthcleanr*, we additionally excluded measurements outside of the inclusion range for the MedWeight Study: (1) heights <121.9 cm (4 ft) or >213.4 cm (7 ft) for adults, (2) weights <22.7 kg (50 lb.) or >317.5 kg (700 lb.) for adults, (3) BMI <15 or >90 for adults, and (4) BMI z-score <−4 or >8 for children based on the CDC reference data.[20]

For pediatric data, we also performed cleaning using the *cp* and *peanof* methods and compared the cleaning results from these 3 methods.

### Statistical analysis

We used SAS (SAS Studio) and R (version 4.0.8) for statistical analyses. We described the demographic characteristics of the study population and calculated the median and interquartile range (IQR) of age at first height or weight data, length of follow-up, and counts of height and weight measurements. Next, we determined the number of exclusions by *growthcleanr* for both pediatric and adult data. For pediatric data, we compared the results of *growthcleanr* with the other 2 cleaning methods (*cp* and *peanof*) using a Sankey diagram (R package, *networkD3*[21]), a flow diagram in which the width of the arrows depicts the proportion to the flagged values across each cleaning method. Based on the 2000 CDC Growth Chart,[22] we calculated age- and sex-specific BMI percentile and defined severe obesity as >120% above the 95th percentile,[23] using same-day height and weight measurements below 20 years of age. In the case of missing same-day height, we imputed height using the closest height measurement ±60 days (for measurements <18 years) or mean of all height measurements above age 18 (for measurements ≥18 years).

# RESULTS

## Demographics of the study cohort

The MedWeight study included 4 047 679 patients; 94% had at least one height or weight measurement. We performed height and weight data cleaning for these participants, who contributed 71 246 369 weight and 51 525 487 height measurements (see Supplementary Figures S1 and S2 for a more detailed distribution of the measurements). The cohort was 56% female, 70% white, and 84% non-Hispanic (Table 2). A total of 687 226 participants contributed pediatric data, and 3 267 293 contributed adult data; 146 687 of them contributed both. The median (IQR) age at the first height or weight measurement was 11.0 (5.7–15.2) years for

**Table 1.** Comparison of different pediatric anthropometric data cleaning methods

| Methods | *growthcleanr* | Conditional percentile (*cp*) | PaEdiatric ANthropometric measurement Outlier Flagging pipeline *(peanof)* |
|---|---|---|---|
| Automated/package | R package available | STATA codes available | Python package available |
| Age range | 2–65 years | N/A (but designed for pediatric growth data) | 2–20 years |
| Outlier detection | • Biological implausible values based on CDC growth chart<br>• Implausible increments/decrements<br>• Outliers flagged by exponentially moving weight average method | • Conditional growth percentiles of growth trajectory (±4SD) | • Biological implausible values based on WHO growth guideline<br>• Implausible increments/decrements<br>• Outliers flagging by robust linear regression method (± 2SD) |
| Error detection | • Duplicates<br>• Carry-forward<br>• Height/weight switches<br>• unit errors<br>• Transposition (10s and 1s digit transposed, eg,: 95, 96, 59, 95, 96)<br>• Weight cap (electronic data entry)<br>• Too many errors in the series of longitudinal data | • No | • Unit errors<br>• Rounding effect bias (Myer's index)<br>• Too many errors in the series of longitudinal data |
| Error correction | Yes (for height/weight switch, transposition, and unit error) | No | No |
| Simultaneous cleaning height and weight | Yes | No | Yes |
| Applicable for adult data? | Yes | No | No |
| Variables required | ID, sex, age, anthropometric measurement | ID, age, anthropometric measurement | ID, sex, age, anthropometric measurement |
| Flexibility of data structure | More flexible to data errors (eg, will run if age is a negative value) | Simple data structure requirement: height and weight cleaned separately | Very specific with the required data structure, must structure the data to fit the program specification (eg, will not run if age is a negative value) |
| Time[a] | Parallel processing ∼3 h (1.5 h data splitting, 1 h data processing in parallel, 0.5 h data merging) | ∼6 h | ∼ 54 h |
| Codes for batch/parallel processing for large data | Yes | No | No |
| Other functionality | Computing BMI (carried forward height); summary and visualization program using *growthviz* python program | Summary (total number of measurement and total number of children with outliers) | Summary (first, last, total *N*, and total length of measurements) |

[a] For processing pediatric data: 687 226 children 24 211 409 height and weight measurements.

the pediatric cohort and 54.3 (39.3–66.2) years for the adult cohort; the median (IQR) length of follow-up data, defined as the time between first and last measures, was 4.3 (1.1–8.6) years and 2.5 (0.2–6.3) years for pediatric and adult cohorts, respectively. Before cleaning, the mean (SD) was 143.5 (43.0) cm for pediatric height, 46.8 (155.1) kg for pediatric weight, 169.0 (25.9) cm for adult height, and 84.4 (33.9) kg for adult weight, and obvious outliers were evident based on the minimum and maximum values (Table 3).

## Cleaning using growthcleanr

Of 24 211 409 pediatric measurements, 17.5% were flagged for exclusion, as were 11.8% of 98 560 447 adult measurements (Table 4) by *growthcleanr*. Most of the exclusions for pediatric data were attributed to carried-forward (8.0%) and extraneous-same-day (7.6%) measurements, while most of the exclusions for adult data were due to identical-same-day (4.9%) and extraneous-same-day (4.4%) measures, which are measures that are different values but on the same day. After exclusions, the mean height and weight remained nearly the same compared to the means before cleaning; the SDs were reduced from 43.0 to 28.5 cm for pediatric height, 155.1 to 27.8 kg for pediatric weight, 25.9 to 10.3 kg for adult height, and 33.9 to 23.1 kg for adult weight (Table 3). The maximum values were also substantially lower with a stable median (IQR) (Table 3), confirming the removal of extreme outliers. After cleaning, 0.5% and 0.6% of the "included/cleaned" pediatric height

**Table 2.** Demographic of PCORnet MedWeight Study Cohort with height and weight data

| | Children<br>*n* = 687 226 | Adults<br>*n* = 3 267 293 |
|---|---|---|
| **Person-level characteristic** | | |
| Sex | | |
| Male | 319 154 (46) | 1 420 817 (43) |
| Female | 368 043 (54) | 1 846 369 (57) |
| Unknown | 29 (0) | 107 (0) |
| Race | | |
| American Indian or Alaska Native | 2233 (0) | 11 862 (0) |
| Asian | 11 208 (2) | 47 517 (1) |
| Black or African American | 106 707 (16) | 550 645 (17) |
| Native Hawaiian or Other Pacific Islander | 999 (0) | 3394 (0) |
| White | 471 198 (69) | 2 288 566 (70) |
| Multiple race | 18 272 (3) | 34 452 (1) |
| Refuse | 6237 (1) | 35 634 (1) |
| Other | 35 596 (5) | 174 327 (5) |
| No information | 34 776 (5) | 120 896 (4) |
| Unknown | 2233 (0) | 11 862 (0) |
| Ethnicity | | |
| Non-Hispanic | 583 987 (85) | 2 739 609 (84) |
| Hispanic | 79 344 (12) | 394 589 (12) |
| Unknown | 23 895 (3) | 133 095 (4) |
| | **Median (IQR)** | |
| Age at first height or weight measurements (year) | 11.0 (5.7–15.2) | 54.3 (39.3–66.2) |
| Duration of available data (year) | 4.3 (1.1–8.6) | 2.5 (0.2–6.3) |
| Number of height measurements per person | 8 (3–18) | 6 (2–16) |
| Number of weight measurements per person | 11 (4–25) | 8 (3–21) |

and weight measurements, respectively, and 0.3% and 1.4% of the "included/cleaned" adult height and weight measurements were still considered biological implausible based on the inclusion range for the MedWeight Study (Supplementary Table S1); the values that were considered biologically implausible for this study were different than what *growthcleanr* uses. *growthcleanr* was designed for a wide range of purposes and, thus, used a very broad limit for implausible biological values.

## Comparison with different cleaning methods

After cleaning with *cp* and *peanof*, the mean (SD) and median (IQR) for pediatric height and weight measurements were comparable to that of *growthcleanr*, but *cp* failed to remove obvious implausible outliers (see maximum value in Table 3 and Supplementary Figure S3a and b for examples). *growthcleanr* did not exclude more extreme measurements than *peanof*; $N = 62$ had SD $< -10$ or $>10$ after cleaning by *growthcleanr* versus $N = 2$ after cleaning by *peanof*. The prevalence of severe obesity was 9.0, 9.2, and 8.0% after cleaning by *growthcleanr*, *cp*, and *peanof*, respectively. In the Supplementary Materials, we showed examples of outcomes by the 3 cleaning methods under different scenarios of outliers and errors. For example, Supplementary Figure S3c shows an example of extraneous-same-day measurements excluded by *growthcleanr* and replaced by mean values or the most fitted value. *cp* only flagged about 0.5% and *peanof* flagged about 12.5% of these extraneous-same-day values as outliers. Supplementary Figure S3d shows an example of carried-forward measurements. Among the 8% of pediatric data flagged as carried-forwards by *growthcleanr*, *cp* only flagged 0.5% of them as outliers, while *peanof* flagged about 14.6% of them as outliers. *growthcleanr* identified errors of swapped measurement where height and weights are swapped. Even though *peanof* did not have this function, all the swapped measurements were flagged as outliers by *peanof*.

Comparing the 3 pediatric data cleaning methods, *growthcleanr* flagged the most measurements as outliers or errors, ie, 19.5% of the pediatric heights and 17.5% of the pediatric weights. *cp* only flagged 0.03% and 0.1% of the pediatric heights and weights, respectively, and *peanof* flagged more heights (22.1%) and fewer weights (11.0%) (Table 1, Figure 1). In addition to carried-forwards and same-day duplicates, *growthcleanr* also set an error load (proportion of measurements of an individual that were flagged) and flagged all measurements for exclusion once the error load was above a certain threshold (see Supplementary Figure S3e for an example).

While *growthcleanr* flagged the most measurements for exclusion, 0.2% and 9.4% of measurements included by *growthcleanr* were flagged as outliers or errors by *cp* and *peanof*, respectively. A detailed distribution of measurements included by *growthcleanr* but flagged as outliers by *peanof* is shown in the Supplementary Table S2. *peanof* flagged most of these measurements based on the ordinary least square regression method and WHO cutoff. While some of these values were true outliers, many likely resulted from having few measurements or higher weighting at other time frames, which led to easier deviation from the regression; additionally, many of these measurements were the last measurements in the series of measurements for that individual (see Supplementary Figure S3f–h for examples). *peanof* did outperform *growthcleanr* in the scenario where an individual had few plausible values and many implausible values that were outside of the WHO reference range. *peanof* correctly flagged the implausible values as outliers based on WHO cutoff while the trajectory-based *growthcleanr* flagged the plausible value, which was outside of the trajectory modeled by EWMA, as outliers (see Supplementary Figure S3i for an example).

# DISCUSSION

We demonstrate the utility of using an automated algorithm, *growthcleanr*, to evaluate potential outliers and errors in pediatric and adult height and weight data from multicenter EHR height and weight data. The proportions of outliers and errors in the data were 18% for children and 12% for adults, and most of the excluded data were carried forwards for pediatric data and same-day extraneous values for adult data. After cleaning with *growthcleanr*, we further removed 0.5% of pediatric heights, 0.6% of pediatric weights, 0.3% of adult heights, and 1.4% of adult weights as they fell outside the inclusion range for the MedWeight Study.

To our knowledge, there are currently 5 automated data cleaning pipelines for EHR data,[7,8,24–26] with 2 focusing on height and weight data cleaning, but only on pediatric data, ie, the *cp* and *peanof* methods.[8,12] *growthcleanr* is the only available automated method with published code with an adult algorithm. A previous study comparing *growthcleanr* with other pediatric growth data cleaning approaches found superior performance (higher specificity and sensitivity) over the CDC growth reference method and a regression-based method.[27] *growthcleanr* also overcomes several limitations of other cleaning methods. For example, the *cp* algorithm requires a prior measurement to "condition on", to evaluate the plausibility of the subsequent measurement. Thus, it naturally

**Table 3.** Distribution of height and weight measurements before and after the data cleanings

| | Before cleaning | After cleaning | | |
|---|---|---|---|---|
| | | *growthcleanr* | *cp* | *peanof* |
| **Pediatric height (cm)** | | | | |
| *N* | 9 561 133 | 7 700 129 | 9 558 098 | 7 450 605 |
| Mean (SD) | 143.5 (43.0) | 143.1 (28.5) | 143.3 (29.8) | 142.9 (28.3) |
| Median (IQR) | 152.0 (126.8–164.6) | 151.1 (126.5–164.0) | 152.0 (126.7–164.5) | 150.1 (126.0–164.0) |
| Min | 0 | 14.5 | 0 | 37.2 |
| Max | 20 568.9 | 315.5 | 16 505.0 | 219.0 |
| **Pediatric weight (kg)** | | | | |
| *N* | 14 650 276 | 12 273 895 | 14 620 237 | 14 105 853 |
| Mean (SD) | 46.8 (155.1) | 47.4 (27.8) | 46.8 (29.0) | 46.1 (26.6) |
| Median (IQR) | 44.2 (24.4–62.9) | 45.2 (25.0–63.5) | 44.3 (24.5–63.0) | 44.0 (24.4–62.5) |
| Min | 0 | 0.35 | 0 | 0.51 |
| Max | 579 998.3 | 337.9 | 22 125.0 | 262.2 |
| **Pediatric BMI percentile** (%) | | | | |
| <5th | | 5.5 | 6.2 | 6.2 |
| 5–85th | | 57.3 | 56.8 | 58.5 |
| 85-95th | | 15.5 | 15.3 | 15.1 |
| >95th | | 21.7 | 21.7 | 20.2 |
| **Pediatric severe obesity**[a] (%) | | 9.0 | 9.2 | 8.0 |
| **Adult height (cm)** | | | | |
| *N* | 41 964 354 | 36 923 662 | | |
| Mean (SD) | 169.0 (25.9) | 168.5 (10.3) | | |
| Median (IQR) | 167.6 (160.0–175.3) | 167.6 (160.0–175.3) | | |
| Min | –144.78 | 50.0 | | |
| Max | 18 136.01 | 233.7 | | |
| **Adult weight (kg)** | | | | |
| *N* | 56 596 093 | 49 960 931 | | |
| Mean (SD) | 84.4 (33.9) | 84.8 (23.1) | | |
| Median (IQR) | 81.6 (68.0–97.7) | 81.7 (68.1–97.7) | | |
| Min | 0 | 20.0 | | |
| Max | 128 949.1 | 429.8 | | |

*cp*: conditional percentile; *peanof*: PaEdiatric ANthropometric measurement Outlier Flagging pipeline.

[a] %BMI$_{p95}$ $\geq$ 120.

cannot be applied to an individual's first measurement. Additionally, like any model-based approach, the *cp* algorithm largely depends on the accuracy of the fitted growth model. If this model does not adequately describe the underlying growth pattern, implausible values identified by this approach would be less meaningful. Since *growthcleanr* uses Z- (or SD) scores[7] and uses an exponentially weighted moving average to flag errors in growth data, it is not affected by this issue of model validity.

Some other advantages of *growthcleanr* include an easy-to-use open-source package. The algorithm incorporates additional functionalities designed explicitly for EHR data, such as checking carried-forward, extraneous-same-day-measures, swapped height and weight measurements, and unit errors. It can accommodate large datasets, with the ability to parallel process to reduce computation time. Evaluation of same-day-extraneous results is particularly useful for an automated data cleaning algorithm because analyses generally cannot use multiple nonidentical values for the same subject for the same day, and they are quite common in EHR data. If the cleaning algorithm does not address same-day extraneous values, the researcher will have to remove these extraneous values at a later time before analyses. Results from *growthcleanr* can also be easily incorporated for visualization on a separate Python Jupyter notebook (https://github.com/mitre/growthviz). This algorithm has also been validated for pediatric data, where results were compared with judgments of physician reviewers and datasets with simulated errors to check for accuracy.[7]

Given these strengths, there are still some limitations and challenges to using *growthcleanr*. For example, users must be familiar with R to smoothly execute the algorithm. There are some subtle details in the software version, packages, and variable names that might render some technical challenges for first-time users. Implausible values in individuals with fewer measurements are less likely to be identified given the trajectory-based method used by the algorithm. Specifically, at least 2 measurements are needed for the EWMA methods to work and too few measurements would make the EWMA trajectory unstable. If a person only had one measurement, *growthcleanr* still checked for its plausibility via the "biological implausible value" and "error cleaning" steps. However, *growthcleanr* was designed to accommodate wide uses and used a very wide range of biological implausible values (for adult weight $\leq$ 20 kg and $>$ 500 kg; height $\leq$ 50 cm and >244 cm). Because the range of included values might differ by study, users are expected to perform additional inspection or cleaning because the *growthcleanr* algorithm innately accommodates some outliers (or extreme values that could still be real). Users are required to set specific cutoff values based on the inclusion range of the study. A small number of biologically implausible measurements are retained after using *growthcleanr*, primarily consecutive nonidentical extreme outliers.

**Table 4.** Results of height and weight cleaning using the *growthcleanr* algorithm

| *growthcleanr* results | *N* | % |
|---|---|---|
| *Pediatric data (N = 24 211 409)* | | |
| Include | 19 974 024 | 82.5% |
| Exclude-Carried-Forward | 1 946 014 | 8.0% |
| Exclude-Extraneous-Same-Day | 1 834 778 | 7.6% |
| Exclude-Min-Height-Change | 190 070 | 0.8% |
| Exclude-Too-Many-Errors-Other-Parameter | 56 105 | 0.2% |
| Exclude-EWMA-8 | 53 451 | 0.2% |
| Missing | 49 066 | 0.2% |
| Exclude-Too-Many-Errors | 37 901 | 0.2% |
| Exclude-EWMA-Extreme | 35 987 | 0.1% |
| Exclude-SD-Cutoff | 13 395 | 0.1% |
| Exclude-EWMA-9 | 8897 | <0.1% |
| Exclude-Max-Height-Change | 7607 | <0.1% |
| Exclude-EWMA-11 | 1143 | <0.1% |
| Exclude-Single-Outlier | 932 | <0.1% |
| Exclude-EWMA-Extreme-Pair | 864 | <0.1% |
| Swapped-Measurements | 587 | <0.1% |
| Exclude-EWMA-12 | 302 | <0.1% |
| Exclude-Pair-Delta-18 | 138 | <0.1% |
| Exclude-Pair-Delta-17 | 125 | <0.1% |
| Exclude-EWMA-13 | 18 | <0.1% |
| Exclude-EWMA-14 | 5 | <0.1% |
| *Adult data (N = 98 560 447)* | | |
| Include | 86 884 593 | 88.2% |
| Exclude-Adult-Identical-Same-Day | 4 830 293 | 4.9% |
| Exclude-Adult-Extraneous-Same-Day | 4 379 926 | 4.4% |
| Exclude-Adult-Distinct-3-Or-More | 1 270 012 | 1.3% |
| Exclude-Adult-BIV | 518 613 | 0.5% |
| Exclude-Adult-EWMA-Moderate | 359 704 | 0.4% |
| Exclude-Adult-Possibly-Impacted-By-Weight-Cap | 98 087 | 0.1% |
| Exclude-Adult-Too-Many-Errors | 85 842 | 0.1% |
| Exclude-Adult-Distinct-Pairs | 69 675 | 0.1% |
| Exclude-Adult-Distinct-Ordered-Pairs | 26 165 | <0.1% |
| Exclude-Adult-EWMA-Extreme | 19 898 | <0.1% |
| Exclude-Adult-Hundreds | 7660 | <0.1% |
| Exclude-Adult-Transpositions | 2556 | <0.1% |
| Exclude-Adult-Distinct-Single | 2385 | <0.1% |
| Exclude-Adult-EWMA-Extreme-RV | 2132 | <0.1% |
| Exclude-Adult-Unit-Errors | 1991 | <0.1% |
| Exclude-Adult-Hundreds-RV | 413 | <0.1% |
| Exclude-Adult-Transpositions-RV | 317 | <0.1% |
| Exclude-Adult-Swapped-Measurements | 132 | <0.1% |
| Exclude-Adult-Weight-Cap-Identical | 53 | <0.1% |

*Note*: EWMA indicated outliers based on exponentially weight moving average (EWMA); SD-cutoff indicated outliers based on evaluation of standard deviation (SD); pair-delta compared heights based on absolute difference of 2 measurements using the WHO height velocity as reference. See https://carriedaymont.github.io/growthcleanr/articles/adult-algorithm.html for more detailed description on reason for exclusion.

Modifications to *growthcleanr* are underway to address this issue. Currently, *growthcleanr* is only designed for cleaning secondary data that have already been collected or recorded and cannot be incorporated into the EHR system for prospective error identification of height and weight entries. The algorithm was specifically designed to include all plausible values when remeasurement is not possible rather than to indicate the need for remeasurement. Finally, PCORnet CDM performs regular data quality checks as part of its data curation program[28]; therefore, PCORnet data might contain fewer errors compared to other EHR data sources. Since *growthcleanr* excludes all measurements if the error load exceeded a certain level, findings from our study may not be generalizable for EHR data that are less curated. Further investigation is warranted to fully evaluate the performance of *growthcleanr* across different datasets of varying quality.

## CONCLUSION

Automated data cleaning is needed for research using large EHR databases. *growthcleanr* is easy to implement for large-scale pediatric and adult height and weight cleaning, and has several advantages compared to other methods: the inclusion of an adult algorithm, efficiency for big data, and handling of extraneous and carried-forward measurements. However, as the algorithm is designed to accommodate a broad range of use, additional cleaning should be implemented for the specific needs of each study. Modifications are being made to further improve *growthcleanr* performance.

## FUNDING

The study was funded by the National Institute of Diabetes and Digestive and Kidney Diseases, R01DK120598 (PI: Block).

## NOTICE

This study was produced for the US Government under Contract Number 75FCMC18D0047, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data-General. No other use other than that granted to the US Government, or to those acting on behalf of the US Government under that Clause is authorized without the express written permission of The MITRE Corporation. For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000. © 2022 The MITRE Corporation.

## AUTHOR CONTRIBUTIONS

PDL, IMA, and JPB were in charge of the design of the research, including project conception, development of overall research plan, and study oversight; PDL and SLR-S performed statistical analyses; DLC assisted with implementation of the computer code and supporting algorithms; DSF contributed the analysis method and critical evaluation of the results; PDL, IMA, and JPB drafted the manuscript; JPB had oversight and leadership responsibility for the research activity planning and execution and obtained the funding to lead this project to publication; and all authors (PDL, SLR-S, IMA, MFD, DMJ, WJH, DLC, DSF, and JPB) provided significant contributions to critically review, commentary and revision of the manuscript.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

## ACKNOWLEDGMENTS

The authors wish to thank the research administration and data fulfillment team at Harvard Pilgrim Health Care Institute for their

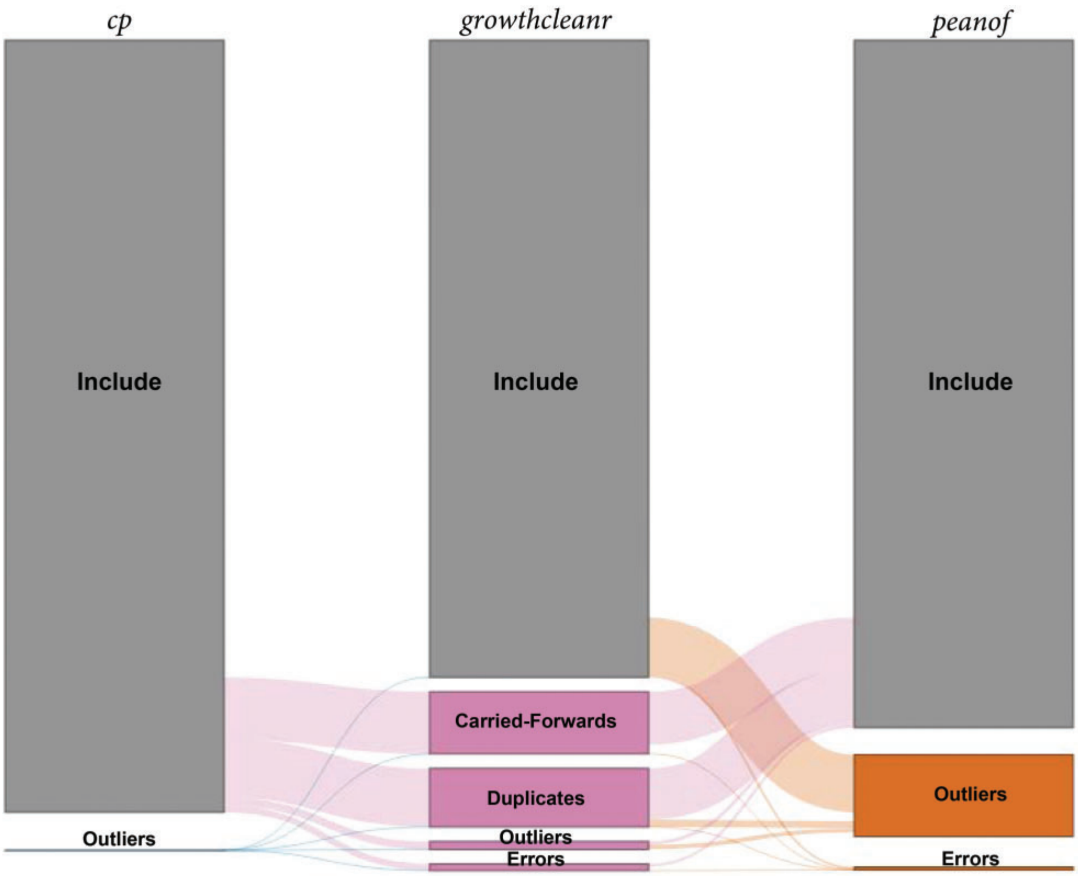

**Figure 1.** Comparison of cleaning results from different automated pediatric anthropometric data cleaning methods. *Note*: The width of the flow diagram shows the proportion of results across methods. The length of the column under each cleaning methods (*cp, growthcleanr, peanof*) indicated the proportion of measurements flagged as include or otherwise, and the width of the lines depicted the proportion to the flagged values across each cleaning method.

assistance in this study, namely, Bridget Nolan, Madison White, Kimberly Barrett, Kshema Nagavedu, Christine Draper, and Megan Wiley. We also wish to thank the developer of *growthcleanr*, Dr Carrie Daymont, for her assistance in navigating the *growthcleanr* package. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention. Approved for Public Release; Distribution Unlimited. Public Release Case Number 22-0483.

## CONFLICT OF INTEREST STATEMENT

None declared.

## DATA AVAILABILITY

*growthcleanr* is an R package that is available freely on Github (https://github.com/carriedaymont/growthcleanr) and the Comprehensive R Archive Network (CRAN, https://cran.r-project.org/package=growthcleanr). The STATA code for calcuating *cp* is available in the Apendix of the article published by Yang et al[9] (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4732581/). *peanof* is an R package that is available freely on Github (https://github.com/hangphan/peanof/).The data underlying this article cannot be shared publicly as they are governed by PCORnet to protect the privacy of individuals that participated in the study. PCORnet data are available to investigators for research purposes through the formal policies and procedures established by PCORnet.

## REFERENCES

1. Safran C, Bloomrosen M, Hammond WE, *et al.* Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc* 2007; 14 (1): 1–9.
2. Smith N, Coleman KJ, Lawrence JM, *et al.* Body weight and height data in electronic medical records of children. *Int J Pediatr Obes* 2010; 5 (3): 237–42.
3. Roche A. *Growth, Maturation, and Body Composition: The Fels Longitudinal Study 1929–1991*. New York: Cambridge University Press; 1992.
4. Lawman HG, Ogden CL, Hassink S, Mallya G, Vander Veur S, Foster GD. Comparing methods for identifying biologically implausible values in height, weight, and body mass index among youth. *Am J Epidemiol* 2015; 182 (4): 359–65.
5. Evans R, Burns J, Damschroder L, *et al.* Deriving weight from big data: comparison of body weight measurement-cleaning algorithms. *JMIR Med Inform* 2022; 10 (3): e30328.
6. Winkler W. *Problems with Inliers*. Suitland, MD: United States Census Bureau; 1998. https://www.census.gov/content/dam/Census/library/working-papers/1998/adrm/rr9805.pdf
7. Daymont C, Ross ME, Russell Localio A, Fiks AG, Wasserman RC, Grundmeier RW. Automated identification of implausible values in growth data from pediatric electronic health records. *J Am Med Inform Assoc* 2017; 24 (6): 1080–7.
8. Phan HTT, Borca F, Cable D, Batchelor J, Davies JH, Ennis S. Automated data cleaning of paediatric anthropometric data from longitudinal elec-

tronic health records: protocol and application to a large patient cohort. *Sci Rep* 2020; 10 (1): 10164.

9. Yang S, Hutcheon JA. Identifying outliers and implausible values in growth trajectory data. *Ann Epidemiol* 2016; 26 (1): 77–80.e1–2.
10. Daymont C, Neal A, Prosnitz A, Cohen MS. Growth in children with congenital heart disease. *Pediatrics* 2013; 131 (1): e236–42.
11. Gerber JS, Bryan M, Ross RK, *et al.* Antibiotic exposure during the first 6 months of life and weight gain during childhood. *JAMA* 2016; 315 (12): 1258–65.
12. Daymont C. growthcleanr: Adult Algorithm. Secondary Growthcleanr: Adult Algorithm 2021. https://carriedaymont.github.io/growthcleanr/articles/adult-algorithm.html. Accessed November 9, 2021.
13. Forrest CB, McTigue KM, Hernandez AF, *et al.* PCORnet(R) 2020: current state, accomplishments, and future directions. *J Clin Epidemiol* 2021; 129: 60–7.
14. PCORnet. Common Data Model (CDM) Specification, Version 6.0. 2021. https://pcornet.org/wp-content/uploads/2020/12/PCORnet-Common-Data-Model-v60-2020_10_221.pdf. Accessed November 9, 2021.
15. Domecq JP, Prutsky G, Leppin A, *et al.* Clinical review: drugs commonly associated with weight change: a systematic review and meta-analysis. *J Clin Endocrinol Metab* 2015; 100 (2): 363–70.
16. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing [Program]; 2022.
17. Aris IM, Lin PD, Rifas-Shiman SL, *et al.* Association of early antibiotic exposure with childhood body mass index trajectory milestones. *JAMA Netw Open* 2021; 4 (7): e2116581.
18. Papadopoulou E, Botton J, Caspersen IH, *et al.* Maternal seafood intake during pregnancy, prenatal mercury exposure and child body mass index trajectories up to 8 years. *Int J Epidemiol* 2021; 50 (4): 1134–46.
19. Sørensen LMN, Aamodt G, Brantsæter AL, Meltzer HM, Papadopoulou E. Diet quality of Norwegian children at 3 and 7 years: changes, predictors and longitudinal association with weight. *Int J Obes* 2022; 46 (1): 10–20.
20. CDC. Z-score Data Files. Secondary Z-score Data Files. 2009. https://www.cdc.gov/growthcharts/zscore.htm. Accessed August 4, 2009.
21. networkD3: D3 JavaScript Network Graphs from R [program]. R package version 0.4 version; 2017.
22. Kuczmarski RJ, Ogden CL, Grummer-Strawn LM, *et al.* CDC growth charts: United States. *Adv Data* 2000; (314): 1–27.
23. Freedman DS, Butte NF, Taveras EM, *et al.* BMI z-scores are a poor indicator of adiposity among 2- to 19-year-olds with very high BMIs, NHANES 1999-2000 to 2013-2014. *Obesity (Silver Spring)* 2017; 25 (4): 739–46.
24. Shi X, Prins C, Van Pottelbergh G, Mamouris P, Vaes B, De Moor B. An automated data cleaning method for Electronic Health Records by incorporating clinical knowledge. *BMC Med Inform Decis Mak* 2021; 21 (1): 267.
25. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. *J Am Med Inform Assoc* 2020; 27 (12): 1921–34.
26. Liu L, Wu DTY, Spooner SA, Ni Y. Development and evaluation of an automated approach to detect weight abnormalities in pediatric weight charts. *AMIA Annu Symp Proc* 2021; 2021: 783–92.
27. Wu DTY, Meganathan K, Newcomb M, *et al.* A Comparison of Existing Methods to Detect Weight Data Errors in a Pediatric Academic Medical Center. *AMIA Annu Symp Proc* 2018; 2018: 1103–9.
28. Qualls LG, Phillips TA, Hammill BG, *et al.* Evaluating foundational data quality in the National Patient-Centered Clinical Research Network (PCORnet(R)). *EGEMS (Wash DC)* 2018; 6 (1): 3.