**Growth Data Cleaning Code Instructions**

This is a description of the Stata code for the purposes of documentation and translating it for other statistical software. The numbered headings correspond to the headings in the Stata code, and the text of these instructions are in the corresponding sections of the R code. The following reference datasets needed to run the algorithm are available in Supplemental File #4: growthfile\_cdc\_ext, tanner\_ht\_vel\_rev, who\_ht\_vel\_3sd, who\_ht\_maxvel\_3sd.

1. **General principles**
   1. Variables are bolded and in red the first time they are mentioned if they should be kept through subsequent steps, otherwise they should be dropped at the end of a step and recalculated if they come up again in a new step.
   2. When a variable is listed as varname\_\* that indicates that there are two separate variables: varname\_wt and varname\_ht, and that the description or instruction applies to both of them.
   3. All steps are done separately for each parameter unless otherwise noted
   4. You will need to set up a method for keeping track of whether a value is missing or excluded (and in what step). I use variables called **exc\_\*** that are =0 if a value is to be included, =1 if missing, and =2 or higher if it is to be excluded, with each number indicating a different step. I also set up parameter-specific **subjid\_\*** variables that are = to subjid for included values and are blank if the value is missing or should be excluded. These subjid\_\* variables need to be updated with each step.
   5. All steps assume that data are sorted by subjid\_\*, parameter, and age (in days) for nonexcluded and nonmissing values only unless otherwise noted. Sorting needs to be redone after any transformations or exclusions to account for excluded and transformed values.
   6. The next value refers to the nonexcluded nonmissing value with the next highest age for the same parameter and the same subject, and the previous value refers to the nonexcluded nonmissing value with the next lowest age for the same parameter and the same subject.
   7. exc\_\* should only be replaced with a higher value if exc\_\*==0 at the time of replacement, unless otherwise specified.
2. Data **set-up**
   1. The following assumes that data are set up in long format with the following variables: **subjid** (subject identifier), **sex**, **agedays** (age in days), **param** (whether measurement is weight or height), and **measurement** (height in cm or weight in kg depending).
   2. Remove rows that are duplicates for subjid, param, and measurement from further analysis
   3. Sex is coded as 0=Male, 1=Female.
   4. Generate separate variables for weight (**wt**) and height (**ht**), as well as exc\_\* and subjid\_\* variables. Set exc\_\*=0 if value is not missing and exc\_\*=1 if value is missing. In all future steps, exc\_\* should only be changed if it is 0. This helps to keep track of which step resulted in the exclusion a value.
   5. Merge dataset with growthdatafile\_cdc\_ext using sex and agedays. The growthdatafile\_cdc\_ext table was made using Centers for Disease Control (CDC) data.18 LMS values for each day were cubically interpolated as is done in the zanthro function in Stata19, and values for the oldest available age were carried forward through age 21 years.
   6. Create wt and ht z-scores for each measurement (**Z: WtZ** and **HtZ**). Z=(((measurement/M)^L)-1)/(L\*S).
   7. Note: For children age 24-35 months, LMS values are available for both length (measured while lying down) and height (measured while standing). The growthdatafile\_cdc\_ext table includes height values only for this age group.
   8. Create SD scores for each measurement. The z-scores created using the LMS method account for skewness in the distribution of the parameters (particularly weight), which can lead to small changes in z-score with large changes in weight in subjects with very high weight, and large changes in z-score for smaller changes in weight in subjects with low weights. We will create SD scores, based on the recommended method for outliers from the CDC,20 that will be less affected by this issue. The variables growthdatafile\_cdc\_ext **cdc\_\*\_csd\_pos** and **cdc\_\*\_csd\_neg** correspond to ½ of the absolute value of the difference between the median and the value with a z-score of +2 (csd\_pos) and -2 (csd\_neg). Create SDorigscores **WtSDorig** and **HtSDorig** by dividing the difference between the value and the median by the SD score. Use cdc\_\*\_csd\_pos if the value is above the median, cdc\_csd\_\*\_neg if the value is below the median (values below the median should have a negative SDorig). These SD-scores, rather than z-scores, will be used for most of the remaining methods.
3. SD-score **recentering:** Because the basis of the method is comparing SD-scores over time, we need to account for the fact that the mean SD-score for the population changes with age.
   1. Determine the median cdc\*sd for each parameter by year of age: median\*sd.
   2. The median\*sd should be considered to apply to midyear-age, defined as the age in days with the same value as the integer portion of (365.25\*year + 365.25/2).
   3. Linearly interpolate median\*sd for each parameter between each midyear-age, naming the interpolated values rc\*sd.
   4. For ages below the first midyear-age, let rc\*sd equal the median\*sd for the earliest year. For ages above the last midyear\_age, let rc\*sd equal the median\*sd for the last year.
   5. Subtract rcsd\_\* from SDorig to create the recentered SD-score. This recentered SD-score, labeled **tbc\*sd** (stands for "to be cleaned") will be used for most of the rest of the analyses.
   6. In future steps I will sometimes refer to measprev and measnext which refer to the previous or next wt or ht measurement for which exc\_\*==0 for the subject and parameter, when the data are sorted by subject, parameter, and agedays. SDprev and SDnext refer to the tbc\*sd of the previous or next measurement.
4. Dataset **split (optional).** In Stata, many of the rest of the steps require you to repeat calculations as many times as there are observations for the subject with the highest number of observations. This step is not necessary but improved efficiency substantially in Stata.
   1. Depending on the statistical software, you can split the dataset at this point into three groups: short (subjects with <=10 total observations), medium (subjects with >10 and <=30 observations) and long (subjects with >30 observations). Perform all the rest of the steps separately for each of the 3 datasets and then re-combine them at the end. Note that all observations for a subject stay together.
5. **Temporary duplicates**: I use duplicates to refer to more than more than one recorded value for a parameter on the same day, and we need to select which one to include in our analysis. The overall strategy will be to select a measurement using a simple strategy that will be used temporarily, and select permanently in a later step after we have a somewhat cleaner dataset that can help us identify the best duplicate.
   1. For subjects/parameters with duplicates: Determine median\_tbc\*sd for both parameters: the median tbc\*sd for each subject and parameter including only non-duplicate values with exc\_\*==0. The median of the same parameter as the duplicate will be referred to as median\_tbc\*sd, the median of the other parameter will be referred to as median\_tbcOsd.
   2. For each subject/parameter with duplicates and at least one value for the subject/parameter on a day with no duplicates, select the value closest to the median\_tbc\*sd for temporary inclusion, and assign all other duplicates exc\_\*=2.
      1. For each subject/parameter with duplicates and no values for the subject/parameter on a day with no duplicates, select the value closest to the median\_tbcOsd for temporary inclusion, and assign all other duplicates exc\_\*=2. If median\_tbcOsd is missing because there are no values for the other parameter, randomly choose one duplicate value for each subject/parameter/age to keep as exc\_\*=0 and replace exc\_\*=2 for all other duplicates for that subject/parameter/age.
6. **EWMA calculation description**: Most of the next steps will involve calculating the exponentially weighted moving average of SDorig for each subject and parameter as described below. The values are also anchored with a minimum of 5. (Note: there is no corresponding section of code for #6; this description will be used in future steps.)
   1. The overall goal of the EWMASD calculation is to identify the difference between the SD-score and what we might predict that SD-score should be, in order to determine whether it should be excluded.
   2. Only nonmissing SD-scores for a parameter that are not designated for exclusion are included in the following calculations. Data must also be sorted by agedays.
   3. For each SD-score SDi and associated agedaysi calculate the following for every other SDorig (SDj…SDn) and associated agedays (agedaysj…agedaysn) for the same subject and parameter
      1. ΔAgej=agedaysj-agedaysi
      2. EWMASD=SDi=[Σj🡪n(SDj\*((5+|agedaysi-agedaysj|)^-1.5))]/[ Σj🡪n((5+|agedaysi-agedaysj|)^-1.5)]
      3. For most steps involving EWMA calculations, there are 3 EWMASDs that need to be calculated. I will note if not all of these need to be done for a given step.
         1. EWMASDall calculated as above
         2. EWMASDbef calculated excluding the SD-score just before the SD-score of interest (sorted by agedays). For the first observation for a parameter for a subject, this should be identical to EWMASDall rather than missing.
         3. EWMASDaft calculated excluding the z-score just after the SD-score of interest (sorted by agedays). For the last observation for a parameter for a subject, this should be identical to EWMASDall rather than missing.
      4. For each of the three EWMASDs, calculate the dewma\_\*=SD-EWMASD
   4. EWMASDs and ΔEWMASDs will change if a value is excluded or manipulated using one of the methods below, therefore EWMASDs and ΔEWMASDs must be recalculated anew for each step where they are needed.
   5. For these calculations, use variables that allow for precise storage of numbers (e.g. double); otherwise, rounding errors can cause problems in a few circumstances
7. Identify **switches** (weight and height each recorded as the other parameter). Strict criteria are required before replacement as a unit error because often values that initially appear to be unit errors have additional problems.
   1. For each day on which a subject had both a weight and a height recorded, calculate tbc\*sd\_sw: SD scores as if the weight had been recorded as the height and the height had been recorded as the weight, recentered using rcsd\_\*.
   2. Perform a EWMA calculation.
      1. Use tbc\*sd (NOT tbc\*sd\_sw) to calculate the EWMA)
      2. In addition to the standard all/bef/aft dewma\_\* variables calculate all/bef/aft dewma\_\*\_sw variables by subtracting EWMASD from tbc\*sd\_sw
   3. Label pairs of height/weight measurements recorded on the same day as a switch if ­all of the following criteria are met for BOTH parameters (Note: values that are the first or last measurement for that subject/parameter cannot be replaced as switches.)
      1. For the weight value: exc\_wt==0 & tbcwtsd>4 & |tbcwtsd\_sw|<3 & dewma\_wt>3 & dewma\_wt\_bef>2 & dewma\_wt\_aft>2 & |dewma\_wt\_sw|<0.3 & |dewma\_wt\_sw\_bef|<0.5 & |dewma\_wt\_sw\_aft|<0.5
      2. For the height value: exc\_ht==0 & tbchtsd<-7 & |tbchtsd\_sw|<3 & dewma\_ht<-6 & dewma\_ht\_bef<-5 & dewma\_ht\_aft<-5 & |dewma\_ht\_sw|<0.3 & |dewma\_ht\_sw\_bef|<0.5 & |dewma\_ht\_sw\_aft|<0.5
   4. For pairs of measurements that meet criteria for a switch, do the following (Table 1):
      1. Replace wt with the value that was originally recorded as the ht, and replace ht with the value that was originally recorded as the wt. Leave measurement as its original value.
      2. Replace tbc\*sd with the values for tbc\*sd\_sw
      3. Switch exc\_wt and exc\_ht
8. Identify **unit errors** (weight/height recorded in wrong units). As with switches, strict criteria are required before replacement as a unit error because often values that initially appear to be unit errors have additional problems.
   1. Generate variables transformed by conversion factors for kg/lbs and inches/cm:
      1. wt\_d\_2=wt/2.204622
      2. wt\_t\_2=wt\*2.204622
      3. ht\_d\_2=ht/2.54
      4. ht\_t\_2=ht\*2.54
   2. Calculate SD scores for each transformed variable and recenter these (tbc\_\*\_sd\_d\_2 and tbc\_\*\_sd\_t\_2)
   3. Perform a EWMA calculation
      1. In addition to the standard all/bef/aft dewma\_\* variables calculate all/bef/aft variables for dewma\_\*\_d\_2 and dewma\_\*\_t\_2 by subtracting EWMASD from tbc\*sd\_d\_2 and tbc\_\*\_sd\_t\_2
   4. Also calculate d\_prevsd\_\*=tbc\*sd-tbc\*sdprev and d\_nextsd\_\*=tbc\*sdi-tbc\*sdnext. d\_prevsd\_\* will be missing for the first value and d\_nextsd\_\* should be missing for the last value. (Note: values that are the first or last measurement for that subject/parameter cannot be replaced as unit errors.)
   5. Identify a value as a unit error if any of the following sets of criteria are met:
      1. For wt\_d\_2: dewma\_wt>3 & dewma\_wt\_bef>2 & dewma\_wt\_aft>2 & tbcwtsd>3 & d\_nextsd\_wt>2 & d\_nextsd\_wt is not missing & d\_prevsd\_wt>2 & d\_prevsd\_wt is not missing & abs(dewma\_wt\_d\_2)<0.3 & abs(dewma\_wt\_d\_2\_bef)<0.5 & abs(dewma\_wt\_d\_2\_aft)<0.5 & abs(tbcwtsd\_d\_2)<3 & exc\_wt==0
      2. For wt\_t\_2: dewma\_wt<-3 & dewma\_wt\_bef<-2 & dewma\_wt\_aft<-2 & tbcwtsd<-3 & d\_nextsd\_wt<-2 & d\_nextsd\_wt is not missing & d\_prevsd\_wt<-2 & d\_prevsd\_wt is not missing & abs(dewma\_wt\_t\_2)<0.3 & abs(dewma\_wt\_t\_2\_bef)<0.5 & abs(dewma\_wt\_t\_2\_aft)<0.5 & abs(tbcwtsd\_t\_2)<3 & exc\_wt==0
      3. For ht\_d\_2: dewma\_ht>5 & dewma\_ht\_bef>4 & dewma\_ht\_aft>4 & tbchtsd>7 & d\_nextsd\_ht>4 & d\_nextsd\_ht is not missing & d\_prevsd\_ht>4 & d\_prevsd\_ht is not missing & (dewma\_ht\_d\_2)<0.3 & abs(dewma\_ht\_d\_2\_bef)<0.5 & abs(dewma\_ht\_d\_2\_aft)<0.5 & abs(tbchtsd\_d\_2)<3 & exc\_ht==0
      4. For ht\_t\_2: dewma\_ht<-5 & dewma\_ht\_bef<-4 & dewma\_ht\_aft<-4 & tbchtsd<-7 & d\_nextsd\_ht<-4 & d\_nextsd\_ht is not missing & d\_prevsd\_ht<-4 & d\_prevsd\_ht is not missing & abs(dewma\_ht\_t\_2)<0.3 & abs(dewma\_ht\_t\_2\_bef)<0.5 & abs(dewma\_ht\_t\_2\_aft)<0.5 & abs(tbchtsd\_t\_2)<3 & exc\_ht==0
   6. For values that are identified as unit errors by one of the sets of criteria above, replace wt or ht with the corresponding transformed value and replace tbc\*sd with the recentered sd-score for the transformed value
9. Exclude values that are **carried forward**. For the purposes of this analysis, any value that is identical to the preceding value for the same parameter and subject is considered carried forward. Because of variations in measurement, he chances of having identical measurements, even at an age/interval when little or no growth would be expected, is small.
   1. Calculate d\_prev\_wt=wt-wtprev and d\_prev\_ht=ht-htprev. Use original measurements rather than transformed measurements (unit errors and switches).
   2. Unlike most steps, do this step for temporarily excluded duplicate values (exc\_\*==2) in addition to included values (exc\_\*==0). Comparing all values for one day to all values from the prior day. If there are any values with a d\_prev\*==0, the value on the latter day should be excluded.
   3. Replace exc\_\*=3 for all values with d\_prev\*==0 & either exc\_\*==0 OR exc\_\*==2)
   4. Replace exc\_\*=0 if exc\_\*==2 & redo step 5 (temporary duplicates)
10. Exclude **extreme errors** **with SD cutoffs.** For this, a cutoff of |SD|>25 is used. Because of differences in SD and z score, there are some very extreme values with a |z|>25 that are implausible with an |SD|<25, so both are used to exclude extreme errors. This works better than using a lower value for the limit for |SD|.
    1. Generally we only evaluate measurements where exc\_\*==0, but for this step we also need to evaluate measurements with exc\_\*==2.
    2. Replace exc\_\*=4 if |tbc\*sd|>25 & either exc\_\*==0 OR exc\_\*==2.
    3. Replace exc\_\*=4 if |\*z|>25 & either exc\_\*==0 OR exc\_\*==2 & the value is not switched or transformed.
    4. Redo temporary duplicates as in step 5.
11. Exclude **extreme errors with EWMA**
    1. Erroneous measurements can distort the EWMA for measurements around them. Therefore, if the EWMA method identifies more than one value for a subject and parameter that meets criteria for exclusion, we will only exclude the value that deviates the most from expected in any given step. Then we will repeat the entire process until no new measurements are identified that meet criteria for exclusion.
    2. Perform a EWMA calculation
       1. Only use values where exc\_\*==0 to determine the EWMAs. However, calculate dewma\_\* variables for values where exc\_\*==0 or exc\_\*==2
    3. Identify all values that meet all of the following criteria as potential exclusions:
       1. There are 3 or more measurements for that subject and parameter
       2. (dewma\_\*>3.5 & dewma\_\*\_bef>3 & dewma\_\*\_aft>3 & tbc\*sd>3.5) **OR** (dewma\_\*<-3.5 & dewma\_\*\_bef<-3 & d & dewma\_\*\_aft<-3 & tbc\*sd<-3.5)
       3. exc\_\*==0
    4. If there is only one potential exclusion identified in step 11c for a subject and parameter, replace exc\_\*=5 for that value
    5. If there is more than one potential exclusion identified in step 11c for a subject and parameter, calculate abssum\_\*=|tbc\*sd+dewma\_\*| for each exclusion and replace exc\_\*=5 for the value with the highest abssum\_\*
    6. For subjects/parameters with only 2 values
       1. Calculate abstbc\*sd=|tbc\*sd|
       2. Replace exc\_\*=6 for values that meet all of the following criteria
          1. There are 2 measurements for that subject and parameter
          2. (dewma\_\*>3.5 & tbc\*sd>3.5) OR (dewma\_\*<-3.5 & tbc\*sd<-3.5)
          3. If there are 2 measurements for a subject/parameter that meet criteria ii, only replace exc\_\*=6 for the value with the larger abstbc\*sd.
    7. Recalculate temporary duplicates as in step 5
    8. If there was at least one subject who had a potential exclusion identified in step 11c, repeat steps 11b-11g. If there were no subjects with potential exclusions identified in step 11c, move on to step 12.
12. Redo **duplicates** **with EWMA**. This will be the final time duplicates are done. For some duplicates it is very difficult to tell which one is likely representative. If the duplicates are very similar to each other, we will select one. If it is very difficult to tell which one is correct and the duplicates are not very similar, we will exclude all duplicates for that subject/parameter on that day
    1. Replace exc\_\*=0 for all temporarily excluded duplicates (exc\_\*==2)
    2. Select which duplicate to include in EWMA calculations using the same criteria as in step 5. However, do not include values in these medians that were excluded in steps 9-11 (exc\_\*=3, 4, 5 or 6)
       1. Determine median\_tbc\*sd and median\_tbcOsd as in step 5.
       2. For each subject/parameter with duplicates and at least one non-duplicate value, select the value closest to the median\_tbc\*sd for inclusion in EWMA calculations.
    3. Calculate a EWMA step for all subjects/parameters with duplicates and at least one non-duplicate value with the following modifications
       1. For calculating the EWMA, include only the duplicate selected in 12c
       2. Calculate dewma\_\* for all values of duplicates
       3. You do not need to calculate EWMAbef or EWMAaft for this step
       4. Calculate abssum2\_\*=|2\*dewma\_\*|+|tbc\*sd| (note that this is different from how abssum\_\* was calculated in step 11).
    4. For each subject/parameter/age with duplicates and at least one non-duplicate value:
       1. Replace exc\_\*=7 for all values except the value that has the smallest abssum2\_\*.
       2. Determine dup\_tot\_\* (# of days with duplicates for that subject/parameter) and nodup\_tot\_\* (# of days with nonexcluded non-duplicates for that subject/parameter).
       3. If dup\_tot\_\*/(dup\_tot\_\*+nodup\_tot\_\*) is greater than ½, replace exc\_\*=7 for all duplicates for that subject/parameter for each age where the largest measurement minus the smallest measurement for that subject/parameter/age is larger than the maximum difference (ht 3cm; wt 0-9.999 kg 0.25kg; wt 10-29.9999 kg 0.5 kg; wt 30kg and higher 1 kg).
    5. For each subject/parameter/age with duplicates and no nonduplicate values:
       1. Replace exc\_\*=7 for all values except the value with the smallest |tbc\*sd-median\_tbcOsd|. If median\_tbcOsd is missing because there are no values for the other parameter, set median\_tbc\*sd=0, which will result in the value with the less extreme tbc\*sd being retained.
       2. If the largest measurement minus the smallest measurement for that subject/parameter/age is larger than the maximum difference (ht 3cm; wt based on the smallest weight for that subject 0-9.999 kg 0.25kg; wt 10-29.9999 kg 0.5 kg; wt 30kg and higher 1 kg), replace exc\_\*=7 for all duplicates for that subject/parameter/age.
    6. For any values that were excluded with exc\_\*=4, 5, or 6 that are also duplicates, replace exc\_\*=7 so that these are considered excluded as duplicates.
13. Calculate **plus/minus measurements** with allowable errors and corresponding recentered SD scores
    1. In order to help determine if a deviation from EWMA could be explained by an allowable degree of error, calculate the following:
       1. **wt\_plus**=wt + 0.05\*wt (5% of weight)
       2. **wt\_minus**=wt - 0.05\*wt (5% of weight)
       3. **ht\_plus**=ht + 1 (1cm)
       4. **ht\_minus**=ht - 1 (1cm)
    2. Foreach of the above calculate and then recenter the SD score for the new value, generating **tbc\*sd\_plus** and **tbc\*sd\_minus**
14. Exclude **moderate errors based on EWMA.** This step is similar to step 11, with repeated exclusions of 1 value at a time, but with different criteria than step 11. There are several criteria used as checks to make sure that values with a large dewma\_\* are not truly likely to be representative.
    1. Perform a EWMA calculation. In addition to standard dewma\_\* variables, calculate dewma\_\*\_plus and dewma\_\*\_minus using the tbc\*sd scores generated in step 13.
    2. Calculate:
       1. d\_prevsd=tbc\*sd-tbc\*sdprev
       2. d\_prevsd\_minus=tbc\*sd\_minus-tbc\*sdprev
       3. d\_prevsd\_plus=tbc\*sd\_plus-tbc\*sdprev
       4. d\_nextsd=tbc\*sd-tbc\*sdnext
       5. d\_nextsd\_minus=tbc\*sd\_minus-tbc\*sdnext
       6. d\_nextsd\_plus=tbc\*sd\_plus-tbc\*sdnext
       7. d\_agedays\_prev=agedays-agedaysprev
       8. d\_agedays\_next=agedaysnext-agedays
    3. Generate abs\_2ndlast\_sd=|tbc\*sd| for the second-to-last measurement for a subject/parameter
    4. Calculate tbcOsd which is the tbc\*sd for the OTHER parameter for the same subject and ageday with exc\_\*==0 (this may be missing).
    5. Calculate median\_tbcOsd which is the median\_tbc\*sd for the OTHER parameter for the same subject with exc\_\*==0 (this may be missing)
    6. Identify values for possible exclusion if they meet one of the following sets of criteria. You can generate a temporary exclusion variable temp\_exc\_\* equal to the number indicated to keep track of which set of criteria were met
       1. Replace temp\_exc\_\*=8 if the value is one of 3 or more measurements for a subject/parameter AND the value is neither the first nor the last measurement AND one of the following sets of criteria are met
          1. dewma\_\*>1 & dewma\_\*\_bef>1 & dewma\_\*\_aft>1 & d\_nextsd\_\*>1 & d\_prevsd\_\*>1 & d\_prevsd\_plus\_\*>1 & d\_prevsd\_minus\_\*>1 & d\_nextsd\_plus\_\*>1 & d\_nextsd\_minus\_\*>1
          2. dewma\_\*<-1 & dewma\_\*\_bef<-1 & dewma\_\*\_aft<-1 & d\_nextsd\_\*<-1 & d\_prevsd\_\*<-1 & d\_prevsd\_plus\_\*<-1 & d\_prevsd\_minus\_\*<-1 & d\_nextsd\_plus\_\*<-1 & d\_nextsd\_minus\_\*<-1
       2. Replace temp\_exc\_\*=9 if the value is the first of 3 or more measurements for a subject/parameter AND d\_agedays\_next<365.25 AND one of the following sets of criteria are met
          1. dewma\_\*>2 & dewma\_\*\_aft>1 & d\_nextsd\_\*>1 & d\_nextsd\_plus\_\*>1 & d\_nextsd\_minus\_\*>1
          2. dewma\_\*<-2 & dewma\_\*\_aft<-1 & d\_nextsd\_\*<-1 & d\_nextsd\_plus\_\*<-1 & d\_nextsd\_minus\_\*<-1
       3. Replace temp\_exc\_\*=10 if the value is the first of 3 or more measurements for a subject/parameter AND d\_agedays\_next>365.25 AND one of the following sets of criteria are met
          1. dewma\_\*>3 & dewma\_\*\_aft>1 & d\_nextsd\_\*>1 & d\_nextsd\_plus\_\*>1 & d\_nextsd\_minus\_\*>1
          2. dewma\_\*<-3 & dewma\_\*\_aft<-1 & d\_nextsd\_\*<-1 & d\_nextsd\_plus\_\*<-1 & d\_nextsd\_minus\_\*<-1
       4. Replace temp\_exc\_\*=11 if the value is the last of 3 or more measurements for a subject/parameter AND d\_agedays\_prev<730.5 AND abs\_2ndlast\_sd <2 AND one of the following sets of criteria are met
          1. dewma\_\*>2 & dewma\_\*\_bef>1 & d\_prevsd\_\*>1 & d\_prevsd\_plus\_\*>1 & d\_prevsd\_minus\_\*>1
          2. dewma\_\*<-2 & dewma\_\*\_bef<-1 & d\_prevsd\_\*<-1 & d\_prevsd\_plus\_\*<-1 & d\_prevsd\_minus\_\*<-1
       5. Replace temp\_exc\_\*=12 if the value is the last of 3 or more measurements for a subject/parameter AND d\_agedays\_prev<730.5 AND abs\_2ndlast\_sd >=2 AND one of the following sets of criteria are met
          1. dewma\_\*>abs\_2ndlast\_sd & dewma\_\*\_bef>1 & d\_prevsd\_\*>1 & d\_prevsd\_plus\_\*>1 & d\_prevsd\_minus\_\*>1
          2. dewma\_\*<(-1\*abs\_2ndlast\_sd) & dewma\_\*\_bef<-1 & d\_prevsd\_\*<-1 & d\_prevsd\_plus\_\*<-1 & d\_prevsd\_minus\_\*<-1
       6. Replace temp\_exc\_\*=13 if the value is the last of 3 or more measurements for a subject/parameter AND d\_agedays\_prev>730.5 AND abs\_2ndlast\_sd <2 AND one of the following sets of criteria are met
          1. dewma\_\*>3 & dewma\_\*\_bef>1 & d\_prevsd\_\*>1 & d\_prevsd\_plus\_\*>1 & d\_prevsd\_minus\_\*>1 & ((tbc\*sd-tbcOsd)>4 OR ((tbc\*sd-median\_tbcOsd)>4 & tbcOsd is missing) OR median\_tbcOsd is missing)
          2. dewma\_\*<-3 & dewma\_\*\_bef<-1 & d\_prevsd\_\*<-1 & d\_prevsd\_plus\_\*<-1 & d\_prevsd\_minus\_\*<-1 & ((tbc\*sd-tbcOsd)<-4 OR ((tbc\*sd-median\_tbcOsd)<-4 & tbcOsd is missing) OR median\_tbcOsd is missing)
       7. Replace temp\_exc\_\*=14 if the value is the last of 3 or more measurements for a subject/parameter AND d\_agedays\_prev>730.5 AND abs\_2ndlast\_sd>=2 AND one of the following sets of criteria are met
          1. dewma\_\*>(1+abs\_2ndlast\_sd) & dewma\_\*\_bef>1 & d\_prevsd\_\*>1 & d\_prevsd\_plus\_\*>1 & d\_prevsd\_minus\_\*>1 & ((tbc\*sd-tbcOsd)>4 OR ((tbc\*sd-median\_tbcOsd)>4 & tbcOsd is missing) OR median\_tbcOsd is missing)
          2. dewma\_\*<(-1-abs\_2ndlast\_sd) & dewma\_\*\_bef<-1 & d\_prevsd\_\*<-1 & d\_prevsd\_plus\_\*<-1 & d\_prevsd\_minus\_\*<-1 & ((tbc\*sd-tbcOsd)<-4 OR ((tbc\*sd-median\_tbcOsd)<-4 & tbcOsd is missing) OR median\_tbcOsd is missing)
    7. If there is only one potential exclusion identified in step 14f for a subject and parameter, replace exc\_\*=temp\_exc\_\* for that value
    8. If there is more than one potential exclusion identified in step 14f for a subject and parameter, calculate abssum\_\*=|tbc\*sd+dewma\_\*| for each exclusion and replace exc\_\*=temp\_exc\_\* for the value with the highest abssum\_\*
    9. If there was at least one potential exclusion identified in step 14f, repeat steps 14b-14h. If there were no potential exclusions identified in step 14f, move on to step 15.
15. Exclude heights based on **absolute differences** in measurement. Once pairs of measurements with implausible amounts of absolute difference between them are identified, the EWMA will be used to determine which value is less likely to be representative and should be excluded. For subjects/parameters with 3 or more measurements, this is done by looking at the dewma\_\* of each of the 2 values in a pair using a EWMA that excludes the other value in the pair. For subjects/parameters with 2 measurements, this is done by looking at the absolute value of the tbc\*sd. The values at +3Z and -3Z from the Tanner and WHO height velocity references are used as a starting point to identify bounds of plausible growth.21,22 Because growth velocity varies more widely with shorter time intervals, evaluation of growth velocity requires comparison based on age, sex, and the interval between measurements. Three additional files are needed: tanner\_ht\_vel\_rev, who\_ht\_vel\_3sd and who\_ht\_maxvel\_3sd. The Tanner height velocity reference includes ages 2 to 18 years. The WHO height velocity reference includes ages birth to 5 years. Tanner values are all based on growth velocity over one year; multiple intervals are available for WHO. For pairs of measurements where either could be used, WHO will be used if difference between ages is < 9 months and Tanner used otherwise.
    1. As with steps 11 and 14, only one value will be excluded per round, and the step will be repeated until there are no more values to exclude.
    2. For each height, calculate d\_age=agedays of next value-agedays of current value.
    3. For each height, calculate mid\_agedays=0.5\*(agedays of next value + agedays of current value)
    4. Generate variable tanner\_months= 6+12\*(round(mid\_agedays/365.25)).
    5. Merge dataset with tanner\_ht\_vel\_rev using sex and tanner\_months – this will give you min\_ht\_vel and max\_ht\_vel.
    6. Calculate the following:
       1. mindiff\_ht=0.5\*min\_ht\_vel\*((d\_agedays/365.25)^2)-3 if d\_agedays<365.25
       2. replace mindiff\_ht=0.5\*min\_ht\_vel-3 if d\_agedays>365.25
       3. maxdiff\_ht=2\*max\_ht\_vel\*((d\_agedays/365.25)^1.5)+5.5 if d\_agedays>365.25
       4. replace maxdiff\_ht=2\*max\_ht\_vel\*((d\_agedays/365.25)^0.33)+5.5 if d\_agedays<365.25
    7. Generate whoagegrp\_ht=agedays/30.4375 rounded to the nearest integer.
    8. Generate whoinc\_age\_ht based on values of d\_agedays\_ht, according to Table 2.
    9. Merge dataset with who\_ht\_vel\_3sd and who\_ht\_maxvel\_3sd using sex and whoagegrp\_ht. This will give you variables whoinc\_i\_ht and maxwhoinc\_i\_ht for various intervals where i is 1,2, 3,4, or 6 and corresponds to whoinc\_age\_ht.
    10. Generate who\_mindiff\_ht=whoinc\_i\_ht (based on the corresponding value of whoinc\_age\_ht); make who\_mindiff\_ht missing if whoinc\_i\_ht or whoinc\_age\_ht is missing.
    11. Generate who\_maxdiff\_ht=max\_whoinc\_i\_ht (based on the corresponding value of whoinc\_age\_ht); make who\_maxdiff\_ht missing if max\_whoinc\_i\_ht or whoinc\_age\_ht is missing.
    12. Scale allowed value based on d\_agedays\_ht:
        1. replace who\_mindiff\_ht=who\_mindiff\_ht\*d\_agedays\_ht/(whoinc\_age\_ht\*30.4375) if d\_agedays\_ht<(whoinc\_age\_ht\*30.4375)
        2. replace who\_maxdiff\_ht=who\_maxdiff\_ht\*d\_agedays\_ht/(whoinc\_age\_ht\*30.4375) if d\_agedays\_ht>(whoinc\_age\_ht\*30.4375)
    13. Replace mindiff\_ht/maxdiff\_ht with adjusted WHO value if Tanner value is missing or if both Tanner and WHO values are available and age difference is < 9 months:
        1. replace mindiff\_ht=0.5\*who\_mindiff\_ht-3 if who\_mindiff\_ht is not missing & d\_agedays\_ht<(9\*30.4375)
        2. replace maxdiff\_ht=2\*who\_maxdiff\_ht+3 if who\_maxdiff\_ht is not missing & d\_agedays\_ht<(9\*30.4375)
        3. replace mindiff\_`ht=0.5\*who\_mindiff\_ht-3 if mindiff\_ht is missing & who\_mindiff\_ht is not missing
        4. replace maxdiff\_ht=2\*who\_maxdiff\_ht+3 if maxdiff\_ht is missing & who\_maxdiff\_ht is not missing
    14. Replace mindiff\_ht=-3 if mindiff\_ht is missing
    15. Determine the min/maxdiffs for the previous age: mindiff\_prev\_ht, maxdiff\_prev\_ht,
    16. Determine d\_prev\_ht=ht-htprev (set to missing for the first value for a subject) and d\_ht=htnext-ht (set to missing for the last value for a subject)
    17. Perform a EWMA calculation with the following modifications:
        1. Generate pair=1 if (d\_prev\_ht<mindiff\_prev\_ht OR d\_ht<mindiff\_ht OR d\_prev\_ht>maxdiff\_prev\_ht OR d\_ht>maxdiff\_ht) AND exc\_ht==0
        2. Generate bef\_g\_aftm1=1 if all of the following are met:
           1. pair==1
           2. pair==1 for the previous value
           3. |dewma\_htbef| for the value of interest is greater than |dewma\_htaft| for the previous value
           4. the value of interest is not the first ht for that subject
        3. Generate aft\_g\_befp1=1 if all of the following are met:
           1. pair==1
           2. pair==1 for the next value
           3. |Δewma\_htaft| for the value of interest is greater than |Δewma\_htbef| for the next value
           4. the value of interest is not the last ht for that subject
        4. Determine prev\_tbchtsd (tbchstsd of the prior value) and next\_tbchtsd (tbchtsd of the next value)
        5. Determine the total number of ht values for each subject (tot\_ht)
    18. Identify a value for possible exclusion if any of the following sets of criteria are met. For values identified by each set of criteria determine the value of temp\_diff using the formula given
        1. d\_prev\_ht<mindiff\_prev\_ht & bef\_g\_aftm1\_ht==1 & exc\_ht==0 & mindiff\_prev\_ht is not missing (temp\_diff=|dewma\_ht\_bef|)
        2. d\_ht<mindiff\_ht & aft\_g\_befp1\_ht==1 & exc\_ht==0 & mindiff\_ht is not missing (temp\_diff=|dewma\_ht\_aft|)
        3. d\_ht>maxdiff\_ht & aft\_g\_befp1\_ht==1 & exc\_ht==0 & mindiff\_ht is not missing (temp\_diff=|dewma\_ht\_aft|)
        4. d\_prev\_ht<mindiff\_prev\_ht & tot\_ht==2 & |tbchtsd|>|prev\_tbchtsd|
        5. d\_ht<mindiff\_ht & tot\_ht==2 & |tbchtsd|>|next\_tbchtsd| (temp\_diff is missing)
        6. d\_prev\_ht>maxdiff\_prev\_ht & tot\_ht==2 & |tbchtsd|>|prev\_tbchtsd| (temp\_diff is missing)
        7. d\_ht>maxdiff\_ht & tot\_ht==2 & |tbchtsd|>|next\_tbchtsd| (temp\_diff is missing)
    19. If there is only one potential exclusion identified in step 15r for a subject and parameter, replace exc\_ht=15 for that value if it met criteria i, ii, v, or vi and exc\_ht=16 if it met criteria iii, iv, vii, or viii.
    20. If there is more than one potential exclusion identified in step 15r for a subject and parameter, determine which value has the largest temp\_diff and replace exc\_ht=15 for that value if it met criteria i or ii, and exc\_ht=16 for that value if it met criteria iii or iv.
    21. If there was at least one potential exclusion identified in step 15r, repeat steps 15b-15t. If there were no subjects with potential exclusions identified in step 15r, move on to step 16.
16. Exclude measurements for subjects/parameters with only **1 or 2 measurements** with exc\_\*=0. This step uses a variety of criteria, including the tbc\*sd of the other parameter.
    1. Identify subjects/parameters with only 2 values with exc\_\*==0, and determine the following:
       1. absd\_tbc\*sd=the absolute value of the difference between tbc\*sd for that value and the other value for that subject/parameter
       2. absd\_agedays\_\*=the absolute value of the difference in agedays for that value and the other value for that subject/parameter
       3. tbcOsd=the tbc\*sd for the other parameter on the same ageday
       4. median\_tbcOsd= the median tbc\*sd for the other parameter
       5. abs\_d\_\*\_O is the absolute value of the difference between tbc\*sd and tbcOsd if tbcOsd is not missing; it is the absolute value of the difference between tbc\*sd and median\_tbcOsd if tbcOsd is missing but median\_tbcOsd is not missing; and is missing if tbcOsd and median\_tbcOsd are both missing
    2. For subjects/parameters with 2 values with exc\_\*==0:
       1. If absd\_agedays\_\*>365.25 and absd\_tbc\*sd>3; replace exc\_\*=17 for the value of the pair that has the largest abs\_d\_\*\_O.
          1. If abs\_d\_\*\_O is missing, replace exc\_\*=17 for the value of the pair with the higher |tbc\*sd|
       2. If absd\_agedays\_\*<365.25 and absd\_tbc\*sd>2; replace exc\_\*=18 for the value of the pair that has the largest abs\_d\_\*\_O.
          1. If abs\_d\_\*\_O is missing, replace exc\_\*=18 for the value of the pair with the higher |tbc\*sd|
    3. Identify subjects/parameters with exactly 1 value for which exc\_\*=0. This will include subjects/parameters for which a value was excluded in step 16b. Determine tbcOsd and median\_tbcOsd as described in step 16aiii and 16aiv above.
    4. For subjects/parameters with 1 value for which exc\_\*=0, replace exc\_\*=19 if one of the following sets of criteria are met:
       1. |tbc\*sd|>3 & |tbc\*sd-tbcOsd|>5 & tbcOsd is not missing
       2. |tbc\*sd|>3 & |tbc\*sd-median\_tbcOsd|>5 & tbcOsd is missing & median\_tbcOsd is not missing
       3. |tbc\*sd|>5 & tbcOsd is missing & median\_tbcOsd is missing
17. Exclude measurements based on **error load** for the subject. If too high of a proportion of a subject’s measurements are excluded it is a) difficult to tell which measurements are representative and b) seems likely that there are other additional errors.
    1. For each subject/parameter determine the following:
       * 1. tot\_exc\_\*=the total number of values for which exc\_\* is equal to 4, 5, 6, or 8-19
         2. tot\_inc\_\*=the total number of values for which exc\_\*=0
       1. For subjects/parameters where tot\_exc\_\* > 0.5 x tot\_inc\_\* and tot\_exc\_\*>=2; replace all values where exc\_\*=0 to exc\_\*=20
       2. For subjects/parameters where tot\_exc\_\* > tot\_inc\_\* and tot\_exc\_\*>=2, replace all values for the OTHER parameter where exc\_\*=0 to exc\_\*=21

**Table 1**

Before

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| subjid | agedays | param | measurement | wt | ht | tbc\_wt\_sd | tbc\_ht\_sd | exc\_wt | exc\_ht | tbc\_wt\_sd\_sw | tbc\_ht\_sd\_sw |
| j5123 | 2500 | weight | 120 | 120 | . | 19.85 | . | 0 | 1 | 0.06 | . |
| j5123 | 2500 | height | 23 | . | 23 | . | -18.43 | 1 | 0 | . | -0.15 |

After

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| subjid | agedays | param | measurement | wt | ht | tbc\_wt\_sd | tbc\_ht\_sd | exc\_wt | exc\_ht | tbc\_wt\_sd\_sw | tbc\_ht\_sd\_sw |
| j5123 | 2500 | weight | 120 | . | 120 | . | 0.06 | 1 | 0 | 0.06 | . |
| j5123 | 2500 | height | 23 | 23 | . | -0.15 | . | 0 | 1 | . | -0.15 |

Missing indicated by “.”

**Table 2**

|  |  |
| --- | --- |
| d\_agedays\_ht | whoinc\_age\_ht |
| 20-45 | 1 |
| 46-75 | 2 |
| 76-106 | 3 |
| 107-152 | 4 |
| 153-198 | 6 |
| All others | missing |

**References**

18. CDC Percentile Data Files with LMS Values. August 2009. http://www.cdc.gov/growthcharts/percentile\_data\_files.htm. Accessed November 14, 2014.

19. Vidmar S, Carlin J, Hesketh K. Standardizing anthropometric measures in children and adolescents with new functions for egen. *The Stata Journal*. 2004;4(1):50-55.

20. *Cut-Offs to Define Outliers in the 2000 CDC Growth Charts.* Atlanta, GA: National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention; 2014. http://www.cdc.gov/nccdphp/dnpa/growthcharts/resources/BIV-cutoffs.pdf. Accessed July 28, 2014.

21. Tanner JM, Davies PS. Clinical longitudinal standards for height and height velocity for North American children. *J Pediatr*. 1985;107(3):317-329.

22. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards: Growth velocity based on weight, length and head circumference: Methods and development. 2009.