Infant Algorithm Documentation
2023-08-18

Definitions

Parameter	Weight (wt), height (ht), and head circumference (hc)
`p'	Used to designate that the code/variable name applies to wt, ht, and hc (or a subset if that is specified). For example, who`p'z refers to whowtz, whohtz, and whohcz
SP	A subject/parameter combination. For example, if subjid A has 5 weights: 3.2, 4.1, 5.4, 6.5, 7.7, we could say 3.2 was the first weight for that SP.
SPA	A subject/parameter/ageday combination.
VOI	Value of interest
POI	Parameter of interest
DOP	Designated Other Parameter – each parameter has a matching parameter that is used for certain evaluations. For wt the DOP is ht. For ht the DOP is wt. For hc the DOP is ht.
EWMA	The exponentially weighted moving average of z-scores of all nonexcluded values for an SP except for the VOI.
DEWMA	The difference between a VOI's z-score and its EWMA.
absDEWMA	The absolute value of DEWMA.
SDE	Same day extraneous. When there is more than one value for a subject and parameter on the same ageday (can refer to the values themselves or the concept).
SDE-SPA	SPA with >1 value
Non-SDE-SPA	SPA with 1 value
SDE group	All value son an SDE-SPA
SDE exclusion	(excluded as SDE) Ultimately, all values or all but one value of an SDE group will be excluded with one of the SDE codes.
Identical SDEs	SDE values on a SPA that are identical to each other (an SDE group can have some identicals and some not and could have multiple groups of identical SDEs, i.e., if the following values occurred on one SPA: 47, 47, 48, 49, 49, there would be two groups of identical SDEs)
Similar SDEs	SDE groups for which all values are within a certain distance from each other (all values in an SDE group have to meet criteria to be similar) -Weight: Highest weight is no more than 3% higher than lowest weight AND highest weight is no more than 2.5kg higher than lowest weight. (For adults, limit is 0.5*wtallow.) -Height: Difference between highest and lowest height is <2.541 cm if lowest height is <127 cm OR difference between highest and lowest height is <5.081 cm if lowest height is >=127 cm. (For adults, limit is 5.081 cm.) -HC: Difference between highest and lowest HC is <1.271 cm.
SDE ratio	The ratio of non-similar-SDE-SPAS to total SPAs for an SP.
Adjacent SDEs	Non-similar-SDE-SPAs for the same subject/parameter with no non-SDE-SPAs in between
VLEZ	Very low early z-score, tbc`p'z < -3 in first 6 months of life

OOB	Out of bounds (step-specific criteria)
Single	An SP with only one ageday with a nonexcluded measurement
Pair	An SP with only two agedays with nonexcluded measurement
NNTE	No need to EWMA; a subject for which all measurements meet certain criteria and can therefore skip steps #5-#15
VCE	Value contributing to the EWMA. These are values that are used to calculate a EWMA for another value for an SP. Unless otherwise specified, all nonexcluded values for an SP that are not the VOI are a VCE. In some steps there are different rules for determining what is a VCE.
PE	Potential exclusion. In many steps, more than one value for an SP might meet potential exclusion criteria, and then further criteria are used to determine which of those values gets excluded.

Everything below refers to values which have not been excluded (permanently or temporarily) unless otherwise specified.

#0: Dataset preparation

Described previously for pediatric algorithm. For the CHOP dataset specifically, HCs at birth have to be corrected with the formula (measurement / 0.0283495) * 2.54 to undo a prior erroneous conversion to ounces.

#1: Set-up

A. Generate a unique observation identifier.
B. Make variables for ageyears=agedays/365.25 and agemonths=agedays/30.4375. These should not be rounded.
C. For any HC measurements >3y, exclude from cleaning and mark with code “Not cleaned”. (Note that in Stata code, these are excluded from dataset for workflow.)
D. For any HT or WT measurements >20y, exclude from cleaning and mark with code “Not cleaned”. (Note that in Stata code, these are excluded from dataset for workflow.)

#2a: Z-scores

As in the original pediatric growthcleanr algorithm, we will make modified z-scores that limit correction for skew in order to maintain differentiation between high weights. The same values were referred to as standard deviation-scores in the original growthcleanr documentation. We are now referring to them as z-scores to avoid confusion with other ways that people use the term/concept of standard deviation in data cleaning. We follow CDC recommendations to use WHO growth curves for those under 2y and CDC growth curves for children 2y and older. However, this creates a jump in z-scores at 2 years that can create bias in cleaning. Therefore, we have averaged the CDC and WHO z-scores for wt and ht from ages 2 to 4 years to smooth this transition. For hc, the CDC curves are only available through 3 years of age, and WHO is available through age 5y. It is useful to have z-scores available for the measurements done between 3y to 5y (although there are not many), therefore we use WHO hc z-scores through age 5y. growthcleanr does not clean hc measurements above 5y.

A. Obtain LMS values for each ageday and sex combination from the CDC and WHO growthdatafile datasets.

B. Make the modified z-scores (called SD scores in last version). Use the same procedure for CDC and WHO.
    a. For all values of `p' that are exactly equal to the corresponding cdc_`p'_m, the who`p'z is 0.
    b. For all values of `p' above the corresponding cdc_`p'_m, the who`p'z is (`p'-cdc_`p'_m)/(cdc_`p'_csd_pos/2)).
    c. For all values of wt below the corresponding cdc_`p'_m, the who`p'z is (`p'-cdc_`p'_m)/(cdc_`p'_csd_neg/2))
C. Create smoothed z-scores.
    a. For each observation, determine the whoweight (4-ageyears) and cdcweight (ageyears-2).
    b. For wt and ht:
        i. For ageyears 2-4, the smoothed z-score (s`p'z) is equal to (who`p'z * whoweight + cdc`p'z * cdcweight)/2.
        ii. For ageyears <=2, s`p'z is equal to who`p'z.
        iii. For ageyears >=4, s`p'z is equal to cdc`p'z.
    c. For hc, shcz=whohcz at all ages.
D. Indicate that merged values (CSD values) are original WHO and LMS values – will need for corrected z-score section

#2b: Corrected z-scores

A number of infants are born significantly premature or with intrauterine growth retardation and then have significant catch-up growth after birth. Much of this catch-up growth occurs in the hospital and is not captured with outpatient data resulting in large jumps in z-scores. If there are large time lapses between measurements, there can be very large differences in z-scores (>5) between informative measurements that are hard to distinguish algorithmically from errors. In order to address this, growthcleanr for subjects whose first weight is recorded before 10 months of age with an swtz<-2, we identify the gestational age (GA) on the Fenton premature infant growth curves with the median weight that corresponds to the subject's weight. That gestational age becomes the subject's assigned gestational age (AGA) for the purposes of growth cleaning.** We correct their age based on this AGA and use this to calculate z-scores based on the Fenton growth curves (for corrected age <350 post-menstrual days or <50 post-menstrual weeks) or the WHO growth curves using corrected age for those <2yo. Then there is a step to see whether the uncorrected or corrected z-score seems to fit the first few weights better; corrected z-scores are only kept if they fit the weights better than uncorrected. In future steps these z-scores are used as an extra check – if there is a corrected z-score available, then both the uncorrected and the corrected z-score have to be beyond the threshold in order for a value to be excluded.*

**Pregnancy/premature infant dating – ignore if you know this. GA is counted from the first day of a pregnant person's last menstrual period (LMP), which incidentally is about two weeks before conception but is easier to date/identify than conception. After birth, we can evaluate the age of preemies in three ways. 1) Chronological age (equivalent to agedays), which is regular number of days from birth. 2) Post-menstrual age (pmagedays) is the total number of days from the LMP, which equals the GA + the chronological age. 3) Corrected age (cagedays), which is the chronological age "corrected" for the degree of prematurity, compared to a term pregnancy of 40 weeks. Typically, the latest GA we would correct for would be 36 6/7 days (258 days) An infant born at 32 weeks is 8*7=56 days early. So at agedays==200, their cagedays is 144. The corrected age is pmagedays - 280 (the length of a term pregnancy), or agedays + GA - 280 (all in days). The Fenton growth curves are made using post-menstrual age and go up to a pmagedays of 350 (50 weeks, or 10 weeks old for a term infant). WHO curves can be used for corrected ages. If the first weight is larger than the maximum weight on the Fenton growth curves (5.8*

kg for males, 5.4 kg for females), that subject does not get corrected – attempting to correct these using a different method created new problems.

***AGA can also stand for appropriate-for-gestational age, as opposed to small- or large-for gestational age. I'm using it to mean something different here. I'm sure it will confuse someone, so feel free to use a different variable name if you want.*

A. Identify those subjects whose first wt is potentially eligible for correction (potcorr): swtz<-2 and agemonths<10.
B. Convert potcorr wts to intwt (integer grams rounded to nearest 10) to facilitate merging with Fenton curve data.
    a. For the purposes of merging, replace intwt of 250 to 560 with 570 because ht and wt for Fenton are only available down to 23 weeks, for which the median weight is 570g. Intwt <250 should not be merged.
C. Merge with fentlms_foraga on sex and intwt. For subjects with potcorr wts who have a matching intwt in Fenton, the variable fengadays becomes the assigned gestational age (aga) for that subject.
D. Corrected z-scores will be made for those with an aga <259 (<37 weeks). Determine the postmenstrual age (pmagedays=aga + agedays) and corrected age (cagedays=pmagedays-280). There may be negative cagedays; these can be set to missing as they will not be used. Replace fengadays with pmagedays to facilitate merging with Fenton LMS values.
E. Merge with fentlms_forz on sex and fengadays.
F. Make unmodified Fenton z-scores using wt in unrounded grams, the Fenton LMS values (fen_wt_l, fen_wt_m, fen_wt_s), and the formula below.
    a. Z = ((WT / M) ^ L – 1) / (L * S)
G. Make corrected-for-GA z-scores. Have to do for both WHO and CDC to smooth corrected z-scores between 2-4y to facilitate smoothing corrected with uncorrected z-scores between 2-4y.
    a. Merge again with growthfile_who on sex and agedays but using the value for cagedays as the agedays variable. This will provide LMS values for the corrected age.
    b. Compute modified z-scores for WHO using the LMS values. When computing these corrected z-scores, if the agedays is >730 and the cagedays is <=730, adjust the ht when computing WHO z-scores by adding 0.8 to the original measurement (0.7 cm for CDC).
H. First round of smoothing – when chronological age is 2-4y, smooth the WHO and CDC corrected z-scores using same *method* from 2aC. (WT/HT only, not HC.)
I. Second round of smoothing – when chronological age is 2-4y, smooth corrected z-score from round 2bH with the smoothed uncorrected *z-score value* from 2aC.
J. The following rules should be used to assign the sc`p'z for each point. This is the z-score that will get recentered and will go on to be cleaned. (Note this is done at different points not all together.)
    a. If a subject has a potentially correctable first weight and an assigned gestational age from the Fenton curves that is <259 days use the rules below:
        i. pmagedays < 350, the sc`p'z is the Fenton z-score
        ii. pmagedays > 350 and agedays<= 730, sc`p'z is the corrected WHO z-score
        iii. Ageyears >= 2 and ageyears<4, sc`p'z is the smoothed sc`p'z from 2bI
        iv. Ageyears>=4, sc`p'z is the same is s`p'z
    b. If a subject does not meet criteria for a, sc`p'z is s`p'z at all ages.
K. Check that the corrected z-scores are more consistent with the growth than uncorrected (i.e., check that the potentially correctable weight is not a large error).

a. For both s`p'z and sc`p'z, calculate the absolute value of the sum of the difference between the z-score for the first weight and the first three wt z-scores <2 years chronological age. (If fewer than 3 weights are available, use just one or two).
b. If the absolute value of the result of step a is smaller for sc`p'z than for s`p'z, continue to use the corrected z-scores. If the absolute value of the result of step `b is larger for sc`p'z than for s`p'z, replace sc`p'z for all values for that subject with the corresponding value of s`p'z.
c. If there are no z-scores meeting criteria in step a retain the corrected z-scores

#3: Recentering and set up

The mean z-score for the population changes over time. Because growthcleanr is based on identifying differences between measured and expected z-scores, this baseline trend can lead to bias and overidentification of implausible values for certain age combinations. Therefore, we identify an approximate median z-score at each age and subtract it from each z-score to attempt to "recenter" the z-scores near zero at all ages.

A. An external file (rcfile-2023-08-15) contains recentering values for each ageday for wt, ht, and hc (rcwtzrc`p'z). These are not sex-specific. Obtain the recentering value that matches with each observation.
B. Create a recentered, to-be-cleaned z-score, tbc`p'z=s`p'z - rc`p'z.
C. Recenter the corrected z-scores, creating ctbc`p'z=sc`p'z - rc`p'z.
D. (I set up exclusion variables in this step – this is probably not the same as your workflow.

#4: No need to clean

Identifying a subset of subjects for which all measurements meet a simple set of criteria that, by definition, mean that they will not trigger exclusion in specific steps will make those steps go faster. Subjects identified as NNTE (no need to EWMA) can skip steps #5-#15. They still need to be cleaned with step #17 (absolute) and subsequent steps.

A. Label subjects meeting all of the following criteria for each parameter as NNTE (no need to EWMA)
    a. No SDEs
    b. No measurements that are identical for a given parameter (no matter when they occur)
    c. All tbc`p'z are > -3 and <3
    d. For each parameter, the difference between the max and min tbc`p'z is <2.5.
    e. The |difference|between the tbc`p'z and the prior tbc`p'z (when one exists) is <1 AND the |difference between the tbc`p'z and the next tbc`p'z (when one exists) is <1.

#5: Temporary SDEs

When there are same day extraneous (SDE) values, growthcleanr identifies the value that best fits with the other measurements for that subject and excludes the others. For a small number of subjects, having implausible values excluded affects which SDE value is selected and, conversely, which SDE value is selected can affect which values subsequently get excluded. The final SDE step is done after the exclusion of more extreme implausible values and before exclusion of less extreme values. But the algorithm requires only one value to be included on each day, so a more crude method of determining which SDE to include is performed temporarily here, and will be repeated at the end of each subsequent step until SDEs are handled permanently.

<u>Step 5 does not apply to NNTEs</u>

A. Calculate the median tbc`p'z for each SP with an SDE. If an SDE-SPA is the only SPA for that SP, calculate the median tbc`p'z for the DOP for that subject, referred to as mediantbc`o'z (o for other). For example, for a subject with 5 tbcwtz values (-0.4, -0.2, -0.2, 0.4, 0.6) and 3 tbchtz values (0.1, 0.3, 0.5), its mediantbc`p'z would be -0.2 and its mediantbc`o'z would be 0.3.
B. For each SDE, determine the absd_median: the absolute value of the difference between each tbc`p'z and its mediantbc`p'z or mediantbc`o'z.
C. On each SDE-SPA, select the value with the lowest absd_median and temporarily exclude the others from subsequent steps.

#6: Carried Forwards

There are frequently values that are identical to the value immediately before. On inspection, the vast majority of these values are consistent with being "carried forward" from prior measurements, rather than being independently recorded. Retaining them can make cleaning more difficult, but more importantly, they can lead to systematic bias. Therefore, the majority of these values are excluded. When there are no SDEs, this is straightforward. When there are SDEs, it is a bit more cumbersome. In order to attempt to distinguish between those measurements that are more likely to reflect independent identical measurements, we have added new steps to the infant algorithm that apply to all pediatric ages. This identifies values with very similar z-scores so that users can choose to reinclude these. Note that whether or not users decide to reinclude these values, they will not be used in the remainder of cleaning steps.

<u>Step 6 does not apply to NNTEs or singles. Perform all steps in order separately for each parameter.</u>

A. If no SDEs, exclude all values that are identical to the prior value for the same parameter with the exclusion code "Carried Forward".
B. If there are SDEs, compare all values for each parameter from each day to all of the values for the same parameter on the prior ageday, and exclude any identical values with the exclusion code "Carried Forward".
C. Identify if weights are whole or half imperial (wholehalfimp), meaning a wt is in whole pounds (absolute value of modulus <0.01 when multiplied by 2.20462262) or a ht/hc are in whole or half inches (absolute value of modulus <0.01 when divided by 1.27).
D. Determine length of each carried forward (CF) string, meaning after the initial value, how many consecutive CFs are there after the initial value. 39.3, 39.3 is a CF string of length 1. 39.2, 39.2, 39.2, 39.2, is a CF string of length 3. (Note, temporary SDEs are excluded in this step, so any CFs that are excluded as SDEs won't be counted.)
E. Determine the |difference| between the initial s`p'z (unrecentered/uncorrected) and each CF s`p'z for that string.
F. If there is only 1 CF in the string AND the |difference| in s`p'z between the initial value AND the CF is <0.05, relabel the CF "1 CF deltaZ <0.05".
G. If there is only 1 CF in the string AND the |difference| in s`p'z between the initial value AND the CF is >=0.05 and <0.1 AND the value is wholehalf imp, relabel the CF "1 CF deltaZ <0.1 wholehalfimp".
H. If there are >=2 CFs in the string AND the CF is >16y for girls or >17y for boys AND the |difference| in s`p'z between the initial value AND the CF is <0.05, relabel the CF "Teen 2 plus CF deltaZ <0.05". It is possible that some CFs in a string will be relabeled and some will not.
I. If there are >=2 CFs in the string AND the CF is >16y for girls or >17y for boys AND the |difference| in s`p'z between the initial value AND the CF is >=0.05 and <0.1 AND the value is wholehalf imp, relabel the CF "Teen 2 plus CF deltaZ <0.1 wholehalfimp". It is possible that some CFs in a string will be relabeled and some will not.

#7: BIVs

Growthcleanr distinguishes between biologically implausible values (BIVs), which refers to values that are exceedingly rare in the population, and implausible values, which are values that are extremely unlikely to represent the true size of a child given their other available measurements. The following limits are used to designate biologically implausible values. These values are based on clinical experience and evaluation of large datasets. These are done after carried forwards so that they do not interfere with determining which values are directly after another value. This step does apply to temp SDEs.

<u>Step 7 does not apply to NNTEs. Step 7 DOES apply to temp SDEs.</u>

A. Exclude values beyond the following limits with the exclusion code "Absolute BIV".
    a. Low wt
        i. wt < 0.2 & agedays==0
        ii. wt < 1 & agedays!=0
    b. High wt
        i. wt > 600
        ii. wt > 35 & ageyears < 2
        iii. wt > 10.5 & agedays==0
    c. Low ht
        i. ht < 18
    d. High ht
        i. ht > 244
        ii. ht > 65 & agedays==0
    e. Low hc
        i. hc < 13
    f. High hc
        i. hc > 75
        ii. hc > 50 & agedays==0
B. Exclude values beyond the following limits with the exclusion code "Standardized BIV". Note these use smoothed z's (s`p'z) without correction or recentering.
    a. wt
        i. swtz < -25 & ageyears < 1
        ii. swtz < -15 & ageyears >= 1
        iii. swtz > 22
    b. ht
        i. shtz < -25 & ageyears < 1
        ii. shtz < -15 & ageyears >= 1
        iii. shtz > 8
    c. hc
        i. shcz < -15
        ii. shcz > 15
C. Redo temporary SDEs (steps 5A-5C).

#9: Evil Twins

An important weakness in the original pediatric growthcleanr algorithm was that it often failed to identify two or more implausible measurements that occurred next to each other, even if they were extremely deviant from a child's other measurements. This step is now added to identify these multiple

extreme values, although it also identifies some single values. It works quite differently from other growthcleanr steps. It identifies out-of-bounds (OOB) measurements, with a z-score that differs from an adjacent z-score by more than 5 and excluding the OOB measurement that is furthest from the median for that SP. As with other growthcleanr steps, only one measurement per SP gets excluded per round; the step gets repeated until only two OOB measurements are identified (there can never be fewer than two, because differences between adjacent z-scores will be reciprocal). Using the median to determine which measurement gets excluded worked better than multiple other more sophisticated methods. The corrected z-score is used as an additional check when available.

Step 9 does not apply to NNTEs, singles, or pairs. Perform all steps in order separately for each parameter.

A. Determine the difference between each tbc`p'z and the tbc`p'z from both adjacent measurements
B. For values with a ctbc`p'z, determine the difference between each ctbc`p'z and the ctbc`p'z for both adjacent measurements.
C. Determine if a measurement is "out of bounds" (OOB)
    a. defined as having a tbc`p'z that is <-5 or >5 from either adjacent z-score.
    b. If a corrected z-score is available, the ctbc`p'z must also be <-5 or >5 from either adjacent ctbc`p'z.
D. For any SP with at least one OOB tbc`p'z, exclude the OOB measurement that is farthest away from the mediantbc`p'z for that parameter with the code "Evil Twin".
E. Repeat steps 9A and 9B for each SP until there are no more than 2 OOB measurements identified in step 9A.
F. Redo temporary SDEs (steps 5A-5C).

***EWMA calculations**

Below are the basic steps in calculating an exponentially weighted moving average (EWMA) and related values, along with some background and explanation. EWMA calculations will be used in most of the subsequent steps. In an update from the original pediatric growthcleanr, the exponent is now adjusted the minimum age difference between the value of interest and the next closest value.

A. Background:
    a. For each VOI (value of interest), all other nonexcluded measurements for that SP are the VCEs (values contributing to the EWMA). This is modified in some steps.
    b. The weights in the EWMA are the absolute value of the distance, in days, of each VCE from the VOI plus 5, raised to the a negative power.
        i. Adding 5 to the distance in days prevents outsize influence of extremely close measurements.
        ii. Raising to a negative power weights closer measurements more heavily. The exponent was determined iteratively.
B. Determination of exponent. This varies by the maximum difference in age between the VOI and each adjacent value. The goal is that if one or more of the values is very far away in age we want to discount those values even further, otherwise they can have an outside influence on the EWMA.
    a. Calculate the difference in days between the ages of the VOI and each adjacent VCE.
    b. If the larger of the values in Ba is <=365, the exponent is -1.5. If the larger of the values in Ba is >365.25 and <=730, the exponent is -2.5, and if the larger of the values in Ba is >730.5, the exponent is -3.5.

C. The EWMA for each VOI is the sum of the tbc`p'z for all the VCEs multiplied by their respective weights, divided by the sum of the weights.
    a. The DEWMA is the tbc`p'z for the VOI minus it's DEWMA. The absDEWMA is the absolute value of the DEWMA.
D. The EWMA-bef (EWMA before) is the EWMA calculated without the z-score and corresponding weight for the nonexcluded measurement just before the VOI. For the first measurement for an SP, the EWMA-bef is the same as the EWMA.
E. The EWMA-aft (EWMA after) is the EWMA calculated without the z-score and corresponding weight for the nonexcluded measurement just after the VOI. For the last measurement for an SP, the EWMA-aft is the same as the EWMA.
    a. The DEWMA-bef, absDEWMA-bef, DEWMA-aft, and absDEWMA-aft are calculated as above for the DEWMA.
    b. Calculating the EWMA-bef and EWMA-aft helps to determine whether value with a high absDEWMA is erroneous or is near an erroneous value that is distorting the EWMA. An erroneous value will generally have a high DEWMA, DEWMA-bef, and DEWMA-aft, while a nonerroneous value will only have two of the three be high. The exception to this is strings of similar erroneous measurements. More extreme strings of errors are addressed in the evil twins step; less extreme strings of errors are often not detectable.
F. In many steps, calculation of the cDEWMA will be required – this is the corrected DEWMA, which is calculated using the ctbc`p'zs when they are available for a subject. When they are not available, you do not need to calculate a cDEWMA. Currently, there are no steps that require a cDEWMA-bef or cDEWMA-aft, but they would be calculated in the same way.

#11: Extreme EWMA

This is the first of the primarily EWMA-based steps, which identifies values that deviate from their expected value based on an exponentially weighted moving average. Values on the first or last day for a subject are not excluded in this step because they require a more detailed approach that will be done after SDEs are handled definitively. Temp SDEs are included as VOIs (values of interest) because we want to exclude all values that meet extreme EWMA criteria in this step. However, they are not included as VCEs (in other words, they do not contribute to the EWMAs calculated for other values) because we have already selected a different value as most likely to be representative of the child's size on that day.

<u>Step 11 does not apply to NNTEs, singles or pairs. Perform all steps in order separately for each parameter.</u>

A. Calculate EWMAs, EWMA-befs, and EWMA-afts for all nonexcluded values AND for temporarily excluded SDEs. In other words, nonexcluded values and temporarily excluded SDEs are VOIs in this step (which is usually not the case).
    a. EWMAs are calculated for temporarily excluded SDEs so that we can identify and exclude extreme values at this stage. Temporarily excluded SDEs should NOT be included as VCEs in EWMA calculations – they get a EWMA calculated for them, but do not contribute to the EWMA of other values.
B. When corrected z-scores are available, calculate cDEWMAs for all nonexcluded values AND for temporarily excluded SDEs.
C. For each VOI, calculate the abssum: |tbc`p'z + DEWMA|, making sure to take the absolute value of the sum, not the sum of the absolute values).
D. Identify potential exclusions (PEs). If cDEWMA is not available, ignore that criterion.

   a. DEWMA > 3.5, DEWMA_bef > 3, DEWMA_aft > 3, cDEWMA>3.5, and tbc`p'z > 3.5 AND not first or last value
   b. DEWMA < -3.5, DEWMA_bef < -3, DEWMA_aft <-3, cDEWMA<-3.5, and tbc`p'z < -3.5 AND not first or last value

E. For each SP with PEs, exclude the PE with the highest abssum with the exclusion code "EWMA1-Extreme".
   a. Note that one of the excluded values should occur on the first or last ageday for an SP.
F. Redo temporary SDEs.
G. If any SP has remaining PEs, repeat steps 11A-11F.

#13: SDEs

In this step, for each SDE group either one value is retained for evaluation with the remainder of the algorithm, or all values for the SDE group are excluded because there is too little information to guide choice of the SDE to retain. The choice of SDE to retain is determined by EWMA for most values, but other strategies are required when there are no non-SDE days for an SP.

Step 13 does not apply to NNTEs. Perform all steps in order separately for each parameter.

A. For each SDE group, remove all but one of each group of identical SDEs with exclusion code "SDE-Identical".
B. Identify similar SDE groups. Similar SDE groups are defined as groups for which all values are within a certain distance from each other. In other words, all values in an SDE group have to meet the criteria to be similar. Criteria are parameter-specific
   a. Weight: Highest weight is no more than 3% higher than lowest weight AND highest weight is no more than 2.5kg higher than lowest weight. (For adults, limit is 0.5*wtallow.)
   b. Height: Difference between highest and lowest height is <2.541 cm if lowest height is <127 cm OR difference between highest and lowest height is <5.081 cm if lowest height is >=127 cm. (For adults, limit is 5.081 cm.)
   c. HC: Difference between highest and lowest HC is <1.271 cm.
C. Exclude all values in non-similar-SDE groups meeting the following criteria with code "SDE-All-Exclude".
   a. In an SP with an SDE ratio >1/3
   b. If the first and second SPA are adjacent non-similar-SDE-SPAs, exclude all values from the first and second SPA
   c. If the last and next-to-last SPA are adjacent non-similar-SDE-SPAs, exclude all values from the last and next-to-last SPA
   d. Adjacent in an SP with an SDE ratio >1/4
   e. Adjacent to at least 2 other SDE groups
D. Calculate EWMAs for all SDE groups with at least one other SPA for that SP (even if all of those SPAs are SDE-SPAs). Similar-SDE-group nonexcluded values (but not non-similar SDE values) that are not on the SPA of interest should be included in the EWMA calculation. If there is only one SPA for a subject, that SP does not have a EWMA.
E. For all non-similar SDE groups with a EWMA, if the lowest absDEWMA is >1, exclude all values for that SDE group with code "SDE-All-Extreme". For all remaining SDE groups with a EWMA, retain the value with the lowest absDEWMA and exclude all others with code "SDE-EWMA".

   Note 1: All remaining SDE-SPAs should be similar-SDE-SPAs, or else they would have been excluded in step 13Ci. They should also be the only SPA for their SP, or else they would have been addressed in steps 13E.

F. For remaining SDE-SPAs for which there is at least one value for the DOP, determine the absolute value of the difference between the z-score for the VOI and the median z-score for the DOP. If the lowest absolute difference is >2, exclude all values of the SDE group with the code "SDE-All-Extreme." Otherwise, retain the value in the SDE group with the lowest absolute difference and exclude all others with the code "SDE-One-Day".
G. For remaining SDE-SPAs for which there are no values for the DOP and 3 or more members in the SDE group, determine the absolute value of the difference between the z-score for the VOI and the median z-score for the POI. Retain the ctbc`p'z in the SDE group with the lowest absolute difference and exclude all others with the code "SDE-One-Day". If there are an even number of values in the SDE-group and the middle two values are different, they will be equidistant from the median. In this case, determine whether the ageday value is even or odd. If it is even, include the lower/exclude the higher of the two middle SDE values with code "SDE-One-Day". If it is odd, include the higher/exclude the lower of the two middle values with code "SDE-One-Day".
H. For remaining SDE-SPAs (which will have no values for the DOP and only 2 members in the SDE group), determine whether the ageday value is even or odd. If it is even, include the lower/exclude the higher of the two SDE values with code "SDE-One-Day". If it is odd, include the higher/exclude the lower of the two values with code "SDE-One-Day".

Note 2: Unlike in the original-peds algorithm, do not replace prior exclusion codes with SDE codes.

*For adult heights, do not exclude for reasons Eb-Ee.
Limit for defining similars for adult weight is 0.5*wtallow, for height should be 5.081
No step H for adults (because there are no z-scores)

#15: Moderate EWMA

This is the second primarily EWMA-based step. In this step, birth HT and birth HC are ignored because they often contain very large amounts of error. Growthcleanr is primarily aimed at evaluating outpatient data. Children may be born in a different health system than the one where they are followed for primary care, and birth measurements are often abstracted from birth records into outpatient records. This creates the opportunity for a lot of error. HT and HC appear to be even more prone to measurement error than WT because a) birth weight is important to a lot of people, so measuring and recording WT accurately and precisely is much more of priority and b) HT and HC are harder to measure accurately. Birth HT and HC will be evaluated in the next step. As in the extreme EWMA step, corrected z's are evaluated when they are available.

This step involves additional criteria compared to the extreme EWMA step to help avoid excluding plausible values. The first are the differences between the tbc`p'z and the next and prior tbc`p'zs. The second are the differences between the tbc`p'z and the next and prior tbc`p'zs modified by adding/subtracting 5% to WTs or 1cm to HT/HC (denoted as plus or minus). The exclusion limits depend on whether the VOI is a first, last, or middle value for that SP and for first and last values, how far it is from the adjacent value – looser limits are required when less information is available regarding a measurement. Also, there is huge variability in growth trajectory after birth, so these z-score differences are allowed to be bigger.

<u>Step 15 does not apply to NNTEs, singles or pairs. HT and HC when agedays==0 are excluded for most of this step, but keep them in at the beginning. Perform all steps in order separately for each parameter.</u>

A. Calculate plus/minus values.
    a. wt_plus is 1.05*wt, wt_minus is 0.95*wt
    b. ht_plus is ht+1, ht_minus is ht-1
    c. hc_plus is hc+1, hc_minus is hc-1
B. Make z-scores for the plus/minus values above – smooth and recenter them as was done for other z-scores to make tbc`p'z_plus and tbc`p'z_minus.
C. Exclude HT and HC when agedays==0 from the remainder of this step
D. Calculate DEWMAs, DEWMA-befs, DEWMA-afts, and cDEWMAs using the usual method.
    a. Unlike in the extreme EWMA step, SDEs are NOT included at all here.
E. Calculate the difference between the tbc`p'z and the tbc`p'z for the prior and next value.
    a. Calculate the difference between the tbc`p'z_plus and the tbc`p'z for the prior and next value.
    b. Calculate the difference between the tbc`p'z_minus and the tbc`p'z for the prior and next value.
F. Identify criteria required for all exclusions: one of the following, depending on whether the dEWMA is pos or neg. If there is no next or prior value, those criteria are considered met.
    a. DEWMA-bef>1 & DEWMA-aft>1 & Difference next z > 1 & difference next z plus > 1 & difference next z minus >1 & difference prior z >1 & difference prior z plus > 1 & difference prior z minus >1
    b. DEWMA-bef<-1 & DEWMA-aft<-1 & Difference next z < -1 & difference next z plus < -1 & difference next z minus < -1 & difference prior z < -1 & difference prior z plus < -1 & difference prior z minus < -1
G. Determine the tbc`p'z for the DOP on the same day for each VOI (tbc`o'z) AND the median of the DOP for each VOI when there is no tbc`p'z for the DOP on the same day (median_tbc`o'z).
H. Identify potential exclusions. If there is no cDEWMA, that criterion is considered automatically met. All potential exclusions require Fa or Fb (the sign of the z-score difference should match the dEWMA).
    a. For middle values "EWMA2 middle"
        i. DEWMA>1 & cDEWMA>1
        ii. DEWMA<-1 & cDEWMA<-1
    b. For birth values with next value <1 year later "EWMA2 birth WT"
        i. DEWMA>3 & cDEWMA>3
        ii. DEWMA<-3 & cDEWMA<-3
    c. For birth values with next value >=1 year later "EWMA 2 birth WT ext"
        i. DEWMA>4 & cDEWMA>4
        ii. DEWMA<-4 & cDEWMA<-4
    d. For first values (not at birth) with next value <1 year later "EWMA 2 first"
        i. DEWMA>2 & cDEWMA>2
        ii. DEWMA<-2 & cDEWMA<-2
    e. For first values (not at birth) with next value >=1 later "EWMA 2 first ext"
        i. DEWMA>3 & cDEWMA>3
        ii. DEWMA<-3 & cDEWMA<-3
    f. For last values with prior value <2 year before and prior |tbc`p'z |<2 "EWMA2 last"
        i. DEWMA>2 & cDEWMA>2
        ii. DEWMA<-2 & cDEWMA<-2
    g. For last values with prior value <2 year before and prior |tbc`p'z |>=2 "EWMA2 last high"
        i. DEWMA>prior tbc`p'z & cDEWMA>3

ii. DEWMA<prior tbc`p'z & cDEWMA<-3

h. For last values with prior value >=2 years before and prior |tbc`p'z|<2 "EWMA2 last ext"
   i. DEWMA > 3 & cDEWMA > 3 & either
      1. tbc`p'z-tbc`o'z > 4
      2. tbc`p'z – median_tbc`o'z > 4 & tbc`o'z is missing
      3. tbc`o'z and median_tbc`o'z are both missing
   ii. DEWMA < -3 & cDEWMA 3 & either
      1. tbc`p'z-tbc`o'z <- 4
      2. tbc`p'z – median_tbc`o'z <- 4 & tbc`o'z is missing
      3. tbc`o'z and median_tbc`o'z are both missing

i. For last values with prior value >=2 years before and prior |tbc`p'z|>=2 "EWMA2 last ext high"
   i. DEWMA > (1 + |prior tbc`p'z|) & cDEWMA>3 & either
      1. tbc`p'z-tbc`o'z > 4
      2. tbc`p'z – median_tbc`o'z > 4 & tbc`o'z is missing
      3. tbc`o'z and median_tbc`o'z are both missing
   ii. DEWMA < (-1 - |prior tbc`p'z|) & cDEWMA>3 & either
      1. tbc`p'z-tbc`o'z <- 4
      2. tbc`p'z – median_tbc`o'z <- 4 & tbc`o'z is missing
      3. tbc`o'z and median_tbc`o'z are both missing

I. For each potential exclusion, calculate the abssum – the absolute value of the sum of the tbc`p'z and the dewma_`p'.

J. Exclude the potential exclusion with the highest abssum for each SP and repeat D-I if any SPs had >1 potential exclusion.

#16: Moderate EWMA for birth HT and HC

This is the last primarily EWMA-based step. It is similar to step 15, but only evaluates birth HT and HC measurements. Unlike 15, it can apply to pairs. Only birth measurements are excluded in this step. Unlike many other steps, it does not need to be repeated.

<u>Step 16 does not apply to NNTEs or singles. It does not apply to WT measurements. Perform all steps in order separately for each parameter.</u>

A. Calculate DEWMAs, DEWMA-befs, DEWMA-afts, and cDEWMAs using the usual method.
   a. Unlike in the extreme EWMA step, SDEs are NOT included at all here.

B. Calculate the difference between the tbc`p'z and the tbc`p'z for the prior and next value.
   a. Calculate the difference between the tbc`p'z_plus and the tbc`p'z for the prior and next value.
   b. Calculate the difference between the tbc`p'z_minus and the tbc`p'z for the prior and next value.

C. Identify criteria required for all exclusions: one of the following, depending on whether the dEWMA is pos or neg. If there is no next or prior value, those criteria are considered met.
   a. DEWMA-bef>1 & DEWMA-aft>1 & Difference next z > 1 & difference next z plus > 1 & difference next z minus >1 & difference prior z >1 & difference prior z plus > 1 & difference prior z minus >1

   b. DEWMA-bef<-1 & DEWMA-aft<-1 & Difference next z < -1 & difference next z plus < -1 & difference next z minus < -1 & difference prior z < -1 & difference prior z plus < -1 & difference prior z minus < -1

D. Identify potential exclusions. If there is no cDEWMA, that criterion is considered automatically met. All potential exclusions require Ca or Cb (the sign of the z-score difference should match the dEWMA). Must meet one line – for example, either ai or aii, not both.
   a. For birth values with next value within one year "EWMA2 birth HT HC"
      i. DEWMA>3 & cDEWMA>3
      ii. DEWMA<-3 & cDEWMA<-3
   b. For birth values with next value >=1 year later "EWMA2 birth HT HC ext"
      i. DEWMA>4 & cDEWMA>4
      ii. DEWMA<-4 & cDEWMA<-4

17. Raw Differences

The raw differences step compares raw as in unstandardized values rather than Z scores. This allows identification of implausible values with less deviant Z scores. It cannot be used for weight because weight is too variable. One thing this step does is identify pairs of values in which it appears that the height or head circumference is shrinking after allowing for a degree of measurement error. However, it also identifies pairs of values in which the next value either shows an inadequate rise in height or head circumference given the expected growth velocity for that age and age interval, or too large of a rise given the expected growth velocity for that age and age interval. Once a measurement pair with an implausible growth pattern is identified, the EWMA is used to determine which value of the pair should be excluded. If there are only two measurements, the one with the more extreme Z score is excluded. The measurement error in height and head circumference measurements makes it more difficult to distinguish between implausible measurement pairs and imprecise measurements. The most complicated part of this step is determining the growth velocity, which must be scaled for each age and each age interval. For height, WHO growth velocity is used for younger children and Tanner for older children. For HC, only WHO is used both because Tanner is not available and WHO is available for all ages needed.

<u>Step 17 does not apply to NNTEs or singles. It does not apply to WT measurements.</u>

A. Calculate d_agedays, which is the age interval (difference in agedays) between the VOI and the next value for the SP. This should always be positive.
B. To merge with the Tanner height velocity table, you need the tanner_months variable, which is 6+0.5*(agemonths of the VOI + agemonths of the next value) for the SP. Set tanner_months to missing if agemonths<30. (The 6+ accounts for the fact that the Tanner height velocity values are provided for the midpoint of each year.)
C. Merge with tanner_ht_vel_rev using sex and tanner_months. This will give you min_ht_vel and max_ht_vel variables. Set a minimum max_ht_vel as below. Scale the min and max allowed Tanner height velocity based on the age interval between the VOI and the next measurement for the SP. Additional allowances are made for measurement error and unusually fast growth.
   a. Set minimum max_ht_vel to 2.54 with the following higher minimums by d_agedays_ht
      i. d_agedays_ht>2*30.4375, min max_ht_vel=2*2.54
      ii. d_agedays_ht>0.5*365.25, min max_ht_vel=4*2.54
      iii. d_agedays_ht>365.25, min max_ht_vel=8*2.54
   b. Scale min/max and add allowances
      i. If d_agedays<365.25, mindiff_ht=0.5*min_ht_vel*(d_agedays/365.25)-3

ii. If d_agedays>365.25, mindiff_ht=0.5*mindiff_ht_vel-3
iii. If d_agedays<365.25, maxdiff_ht=2*max_ht_vel*(d_agedays/365.25)^1.5 + 5.5
iv. If d_agedays>365.25, maxdiff_ht=2*max_ht_vel*(d_agedays/365.25)^0.33 + 5.5

D. Generate the WHO age group variable
   a. whoagegroup_`p' is the rounded agemonths of the agemonths.
   b. Set whoagegrp_`p' to missing if the agemonths OR the next agemonths are >24.
E. Generate WHO age group variables for HT (these are different for HC because no 1-month age interval is available for HC).
   a. whoinc_age_ht = 1 if d_agedays_ht >= 20 & d_agedays_ht < 46
   b. whoinc_age_ht = 2 if d_agedays_ht >= 46 & d_agedays_ht < 76
   c. whoinc_age_ht = 3 if d_agedays_ht >= 76 & d_agedays_ht < 107
   d. whoinc_age_ht = 4 if d_agedays_ht >= 107 & d_agedays_ht < 153
   e. whoinc_age_ht = 6 if d_agedays_ht >= 153 & d_agedays_ht < 199
   f. If agedays_ht<20, replace who_inc_age_`p' = 1
   g. If agedays_ht>=200, replace d_agedays_`p' = 200 AND replace who_inc_age_`p' = 6
F. Merge using sex and whoagegrp_ht with who_ht_vel-3sd and who_ht_maxvel_3sd
   a. This will provide who_inc_`i'_ht and max_who_inc_`i'_ht with `i' representing 1, 2, 3, 4, and 6. Convert these into variables who_mindiff_ht and who_maxdiff_ht
G. Scale WHO velocity values and add tolerance for measurement error
   a. If d_agedays is < than the age group integer * 30.4375, scale the who_mindiff variable by d_agedays/(whoinc_age_`p'*30.4375). For example, if d_agedays==49, the age value integer is 2. If who_mindiff is 5, the scaled who_mindiff would be 5 * 49 / (2 * 30.4375).
   b. If d_agedays is > than the age group integer * 30.4375, scale the who_maxdiff variables by d_agedays/(whoinc_age_`p'*30.4375).
   c. Adjust the who_mindiff_ht/who_maxdiff_ht (scaled if appropriate) as below
      i. If d_agedays < 9 months, replace who_mindiff with who_mindiff * 0.5 – 3
      ii. If d_agedays < 9 months, replace who_maxdiff with who_maxdiff * 2 + 3
H. Determine whether Tanner or WHO (or nothing) will be used for mindiff_ht and maxdiff_ht
   a. If WHO and Tanner are available, and d_agedays is <9 months, use WHO
   b. If only one of WHO and Tanner are available, use that one
   c. If neither is available (because lower than age limit for Tanner or the age interval does not match any of the WHO intervals), mindiff_ht=-3 and maxdiff_ht =missing (there is no upper limit to the difference in HT for these measurement pairs).
   d. After doing above, for birth measurements, add an extra 1.5 cm to all allowances for mindiff_ht and maxdiff_ht (i.e., replace mindiff_ht with mindiff_ht-1.5 and replace maxdiff_ht with maxdiff_ht+1.5).
I. Generate variable containing the mindiff_ht/maxdiff_ht of the previous measurement
J. Generate WHO age group variables for HC (these are different for HC because no 1-month age interval is available for HC). NOTE: unlike for HT in 17E, age group variables are not assigned if the age interval is outside of the ranges below.
   a. whoinc_age_hc = 2 if d_agedays_hc >= 46 & d_agedays_hc < 76
   b. whoinc_age_hc = 3 if d_agedays_hc >= 76 & d_agedays_hc < 107
   c. whoinc_age_hc = 4 if d_agedays_hc >= 107 & d_agedays_hc < 153
   d. whoinc_age_hc = 6 if d_agedays_hc >= 153 & d_agedays_hc < 199
K. Merge using sex and whoinc_age_hc with who_hc_vel-3sd and who_hc_maxvel_3sd
   a. This will provide who_inc_`i'_hc and max_who_inc_`i'_htcwith `i' representing 2, 3, 4, and 6. Convert these into variables who_mindiff_hc and who_maxdiff_hc
L. Scale WHO velocity values and add tolerance for measurement error

   a. If d_agedays is less than the age group integer * 30.4375, scale the who_mindiff/who_maxdiff variables by d_agedays/(age interval*30.4375). For example, if d_agedays==49, the age value integer is 2. If who_mindiff is 5, the scaled who_mindiff would be 5 * 49 / (2 * 30.4375).
   b. If d_agedays is more than the age group integer * 30.4375, scale the who_mindiff/who_maxdiff variables by d_agedays/(age interval*30.4375).
   c. Adjust the who_mindiff_hc/who_maxdiff_hc (scaled if appropriate) as below
      i. If d_agedays < 9 months, replace who_mindiff with who_mindiff * 0.5 – 1.5
      ii. If d_agedays > 9 months, replace who_maxdiff with who_maxdiff * 2 + 1.5

M. Determine whether WHO (or nothing) will be used for mindiff_ht and maxdiff_ht
   a. If WHO is available, use it
   b. If WHO is not available is available (the age/age interval does not match any of the WHO intervals), mindiff_hc=-1.5 and maxdiff_hc =missing (there is no upper limit to the difference in HC for these measurement pairs).
   c. After doing above, for birth measurements, add an extra 0.5 cm to all allowances for mindiff_ht and maxdiff_ht (i.e., replace mindiff_ht with mindiff_ht-0.5 and replace maxdiff_ht with maxdiff_ht+0.5).
N. Generate variable containing the mindiff_hc/maxdiff_hc of the previous measurement
O. Generate DEWMA_bef and DEWMA_aft values (do not need cDEWMA values here)
P. For each member of pair you are evaluating, use the version of DEWMA that does not include the other member of the pair. For example, for the second member of the pair, use DEWMA_bef which excludes the value right before the VOI.
Q. Identify value with larger |tbc`p'z| if there are only 2 measurements for the SP – DEWMA will be the same for these
R. Identify VOIs for which
   a. the raw difference between the VOI and the next value is lower than the mindiff
   b. the raw difference between the prior value and the VOI is lower than mindiff_prev
   c. the raw difference between the VOI and the next value is higher than the maxdiff
   d. the raw difference between the prior value and the VOI is higher than the maxdiff_prev
S. Separately for HT and HC, of VOIs that meet the criteria in R, identify the one with the largest |DEWMA| (if available) or the largest |tbc`p'z| if it is a pair, and exclude that value with the code "Min diff" or "Max diff" as appropriate.
T. If there is more than one value in a parameter identified in step R, repeat step 17.

19. 1 or 2 measurements

This step addresses SPs for which there are only 2 remaining nonexcluded measurements (pairs) or one remaining nonexcluded measurement (singles). Exclusion decisions for these are not as robust as they are for values for which there is more information. These exclusions are based on the absolute value of the z-score, the difference between measurement z-scores (for pairs) including corrected z-scores when available, and the difference between a measurement's z-score and the z-score for the DOP. When available that is the z-score for the DOP value on the same day, when no measurement on the same day is available it's the median z-score for the DOP.

<u>Step 17 does not apply to NNTEs. It only applies to singles and pairs, although requires info from non-singles/pairs. Perform all steps separately for each parameter.</u>

A. Identify singles and pairs. Identify first and second values of pairs.
B. For pairs, determine
    a. tbc`p'z of other value
    b. ageday of other value
    c. difference in tbc`p'z and agedays between value
    d. the absolute value of the tbc`p'z of the VOI
    e. the median tbc`p'z of the DOP
C. Repeat step B for ctbc`p'z (do not need to do agedays again in Bb and Bc)
D. Compare the VOI and the DOP
    a. When a value for the DOP exists on the same day as the VOI, compare the |difference| for the tbc`p'z for the DOP and VOI on the same day
    b. When there is no value for the DOP on the same day as the VOI, compare the |difference| for the tbc`p'z for the VOI with the median tbc`p'z for the DOP
E. For pairs, identify the value with the larger absolute difference from step D. If the differences from step Da or Db are the same, or if they are missing (i.e., there are no values for the DOP), then identify the value with the larger |tbc`p'z|.
F. If the |difference| between the two tbc`p'z is >4, and the difference between the two ctbc`p'z (if available) is >4, and the difference in agedays is >=365.25, exclude the more discrepant value with code "2 meas >1 year"
G. If the |difference| between the two tbc`p'z is >2.5, and the difference between the two ctbc`p'z (if available) is >4, and the difference in agedays is <365.25, exclude the more discrepant value with code "2 meas <1 year"
H. Re-evaluate how many values there are for an SP. If one measurement was excluded from a pair in steps F or G, it gets re-evaluated as a single in the next step.
I. For singles, exclude if one of the following criteria is met with code "1 meas"
    a. the |tbc`p'z|>3 and the |difference| from step D is >5
    b. If there are no DOP measurements, the |tbc`p'z| is >5

21. Error load

Sometimes there are so many apparently erroneous values for an SP it is difficult to tell what is real. Some of these look like EHR test subjects. Others just have incredibly messy data. If the ratio of excluded to total values is too high, we recommend exclusion of all values for the SP for that subject. These criteria have changed somewhat from the original pediatric growthcleanr algorithm, and now match those for the current adult algorithm.

<u>Step 21 does not apply to NNTEs. Perform all steps separately for each parameter.</u>

A. Identify values excluded for error. These are all excluded values EXCEPT those excluded as CFs or SDEs.
B. Determine the ratio between values excluded for error / (values excluded for error + included values)
C. If the ratio in B is >0.4, exclude all values for that SP.
D. Error load does not overwrite prior codes.

22. Clean-up (not a step, needed for Stata workflow)

23. Append NNTE values (may not be relevant depending on R workflow, but a reminder just in case)

No-need-to-evaluate values have to be added back in.

A. Add NNTE values back in to dataset.

25. Clean-up (not a step, needed for Stata workflow)