## SLEEP_DATA_README

***WE KNOW THERE ARE SOME ERRORS IN THE SLEEP DATA BUT WE HAVE DECIDED IT WILL WORK BEST TO LET PRIMARY AUTHORS DECIDE HOW THEY WANT TO MANAGE THEM***

Background: The difficulty with the sleep data is that it requires free response, and the only way that REDCap can verify a standard time entry is in 24-hour format, and some participants struggled with that. Further, there is a variety of ways the participants could make errors and we have yet to find a way to safely and confidently catch all of them without making assumptions or potentially losing good data along with bad. As such, **the goal of our initial data cleaning was to take a conservative approach**: instead of trying to correct problems by making additional assumptions, we took responses at face value as much as possible, and then use a missing value when this produces impossible or ambiguous results (but not results that are simply unlikely), rather than trying to guess at the correct response. This still leaves a number of known issues in a small percentage of entries we will discuss now:

<u>The following rules have been applied in the cleaned version of the data:</u>

1. **Calculation of Time In Bed (TIB)**. Number of hours between out of bed time in morning (*sleepdiary_outofbed*) and bedtime in evening (*sleepdiary_bedtime*).
    a) **If this is negative**, add 24 hours to adjust for date change. For example, if bedtime is 23:00 and outofbed is 07:00, the first step gives TIB=-16. Adding 24 gives TIB=8.
    b) **Identify entries in which <u>12-hour clock (am/pm)</u> was used instead of 24-hour clock**. To do this, we identified rows where both of the following conditions are met:
        i) (*sleepdiary_outofbed*) < (*sleepdiary_bedtime*)
        ii) Both (*sleepdiary_outofbed*) and (*sleepdiary_bedtime*) are between 1:00 and 12:59
    c) **For cases identified in step 1b, subtract 12.** For example, if bedtime was 11:00 and outofbed was 07:00, the first two steps give TIB=20. Subtracting 12 gives TIB=8.

2. **Calculation of "Sleep Attempt" (sleepattempt)**. Same rules as described above for TIB but calculating the difference between wake time in morning (*sleepdiary_waketime*) and fall asleep time in evening (*sleepdiary_fallasleep*). The same corrections were made if negative or appeared to have issues with 12-hour clock.

3. **Calculation of Total Sleep Time (TST)**. Sleep attempt (sleepattempt) minus sleep latency (*sleepdiary_sleeplatency*) and wake after sleep onset (*night_awakening_time*)

4. **Calculation of Sleep Efficiency**. TST/TIB

5. **If a row used all 12-hour clock times and Time In Bed TIB) <3, replace with missing value**. This is because we can't distinguish 0-3 hours in bed from 12-15 hours in bed, so the true TIB for these responses is unknown without making assumptions. An example case is where someone says they went to bed at 12:00 and got out of bed at 12:30.

6. **Same as (5) to correct for 12-hr clock in Total Sleep Time (TST)**, except we used 2 hours as the cutoff instead of 3.

> -NOTE: We selected 3hrs for TIB and 2 hrs for TST fairly arbitrarily. Once we have finished all data cleaning, we will make the code available and primary authors can change these metrics if preferred

7. If the sum of sleep latency (*sleepdiary_sleeplatency*) and wake after sleep onset (*night_awakening_time*) is greater than the time spent attempting to sleep (*sleepdiary_waketime* - *sleepdiary_fallaslseeep*)—i.e., **if TST is negative—replace TST with missing value**.

8. **If sleep efficiency (SE) is greater than 1, replace SE, TST, and TIB all with missing value**. Clearly something went wrong with either TST or TIB, so we won't consider either reliable.

9. **If TIB=0, replace SE, TST, and TIB with missing value**, as there are either errors or we are misinterpreting something.

10. There are also two new columns: **TIB_12** and **TST_12** are “1” if a 12 hour clock was assumed for TIB and TST calculations, respectively, otherwise “0”

Impact on data after following these rules:
For TST, this results in 1.9% missing values. 96% of non-missing TST values are between 4 and 12 hours.

**KNOWN ERRORS:**

1. **The corrections listed above were ONLY made to the higher-level sleep calculations (e.g. TIB, TST). Even when we corrected for 12-hour clock, we DID NOT change the time entry in the rows for bedtime *(sleepdiary_bedtime)*, fall asleep time *(sleepdiary_fallaslseeep)*, wake time *(sleepdiary_waketime)*, and out of bed time *(sleepdiary_outofbed)*.** We could not be 100% certain in how primary authors would want to handle this, particularly circadian researchers, and as such left the time entries reported by the participants untouched.

- The new variables **TIB_12** and **TST_12** will identify entries in which we detected a 12-hour clock issue

2. **A number of participants reported fall asleep time *(sleepdiary_fallaslseeep)* EARLIER than bedtime *(sleepdiary_bedtime)*.** Our cleaning identified a few hundred of these (out of 32,000+) ranging from 1-minute difference to several hours. These will cause issues because as long as the math works out so that SE is less than 1, they are harder to identify and will incorrectly lead to better sleep metrics (e.g., if they report falling asleep at 23:00 and getting in bed at 23:30, this will essentially result in a 1 hour swing of sleep time calculated in the Sleep Efficiency (SE) Score)

- A much smaller number of entries reported **wake times LATER than out of bed times in the morning** in which similar issues apply. This occurred on a much smaller scale though

**<u>RECOMMENDATIONS:</u>**

1) Primary authors should determine the outlier approach (e.g. TST > 20) that they would like to apply to the data to deal with extremes.
2) Primary authors should plan how they want to deal with the frequent use of 12-hour clock if they are going to use any of the time entered variables: bedtime *(sleepdiary_bedtime)*, fall asleep time *(sleepdiary_fallaslseeep)*, wake time *(sleepdiary_waketime)*, and out of bed time *(sleepdiary_outofbed)*
3) Primary authors should identify and determine how they want to deal with the entries in which reported fall asleep time (*sleepdiary_fallaslseeep*) is EARLIER than bedtime (*sleepdiary_bedtime*) or wake time *(sleepdiary_waketime)is* LATER than out of bed time *(sleepdiary_outofbed)* in the morning.
    a) One potential idea is that up to a certain difference (e.g. 30 minutes) the times could be set at the median, and above that number be eliminated.

*If you have suggestions for systematically dealing with the known errors, please reach out to us and let us know and we can attempt to enter it into our data cleaning code.* Fortunately, a vast majority of the entries are good, but 5-6% of the sleep data will need some attention.

**SLEEP VARIABLES**:

*Note: Variables in RED TEXT are not included in the de-identified versions of the data. To get access to this data, you must contact Tony Cunningham (acunnin4@bidmc.harvard.edu) and be brought onto the IRB (which is definitely doable)

| | |
|---|---|
| TIB | Time in bed. Number of hours between out of bed time in morning (*sleepdiary_outofbed*) and bedtime in evening (*sleepdiary_bedtime*). Corrections made to address issues with use of 12 hour clock (see SLEEP_DATA_README) |
| TIB_12 | 1 = 12 hour clock corrections were made in calculation of TIB; 0 = 12 hour clock ccorrections were not made |
| sleepattempt | Sleep attempt - amount of time that participant was attempting sleep. Number of hours between wake time in morning (*sleepdiary_waketime*) and fall asleep time in evening (*sleepdiary_fallasleep*). Corrections made to address issues with use of 12 hour clock (see SLEEP_DATA_README) |
| TST_12 | 1 = 12 hour clock corrections were made in calculation of TST; 0 = 12 hour clock ccorrections were not made |
| TST | Total sleep time. Sleep attempt (sleepattempt) minus sleep latency (*sleepdiary_sleeplatency*) and wake after sleep onset (*night_awakening_time*) |
| SE | Sleep efficiency. TST/TIB |
| sleepdiary_bedtime | This is the time the participants **got into bed**. Required to be in military format (eg. 20:00). Most likely between 19:00 and 03:00, but I'm sure I have some unusual schedules and I know for a fact some night shift workers. Also many mistakes here with people not using 24 hour format correctly |
| sleepdiary_fallasleep | This is the time the participants estimated that they **tried to fall asleep**. Not necessarily when they did fall asleep, but at least turned off the lights and attempted. Required to be in 24-hour format (eg. 20:00). Many mistakes here with people not using 24 hour format correctly |
| sleepdiary_sleeplatency | This is how many minutes it took the participants to fall asleep from when they started trying. Required number entry |
| sleepdiary_wakes | This is the number of times that the participants estimated they woke up last night. It was multiple choice ranging from 0 up to 5+, 0=0, 1=1, 2=2, 3=3, 4=4, 5+ = 5 |
| night_awakening_time | This is the number of minutes that participants spent awake throughout the night. Required number entry |
| sleepdiary_waketime | This is the time the participants estimated that they **woke up in the morning**. Not necessarily when they did fall asleep, but at least turned off the lights and attempted. Required to be in military format (eg. 08:00). Most likely between 05:00 and 12:00, but I'm sure I have some unusual schedules and I know for a fact some night shift workers. Many mistakes here with people not using 24 hour format correctly |
| sleepdiary_outofbed | This is the time the participants estimated that they **physically got out of bed in the morning**. Required to be in military format (eg. 08:00). Most likely between 05:00 and 12:00, but I'm sure I have some unusual schedules and I know for a fact some night shift workers. Many mistakes here with people not using 24 hour format correctly |

| | |
|---|---|
| sleepdiary_fellasleep | This is a question on their perception of difficulty falling asleep the night before. 1 = It was easy, 2 = It took some time, 3 = It was difficult. **Only asked in full version.** |
| sleepdiary_dreams | This is whether or not they remembered dreaming the night before, 1 = Yes, 2 = No, 3 = I don't recall. If they say yes to this it breaks out to the dream content question (potentially identifiable info that has been excluded) |
| sleepdiary_dreamcontent | Free response to “Please describe in as much detail as you'd like the content of your dreams last night.” |
| sleepdiary_nap | This is a question of if they napped the previous day, 1 = Yes, 0 = No |
| sleepdiary_naptime | This is the number of minutes that the participant napped. Required number entry |
| cst | This is whether or not they used a sleep tracker or not to help record their responses (useful to some sleep researchers), 1 = Yes, 0 = No |
| sleepdiary_info | Free response to “Feel free to include any other relevant information about your sleep here, including any disturbances that contributed to you waking up during the night.” |