Sense2Stop Data Curation Documentation

As the intended design and planned analyses of the Sense2Stop study has already been described (Battalio et al., 2021), a major focus of this documentation will be on articulating decisions made when constructing the dataset used to test the primary aim hypotheses. Scientific and practical considerations underpinning these decisions are also discussed in this documentation.

1. Primary aim

The Sense2Stop study (Battalio et al., 2021) sought to investigate the following scientific question through a micro-randomized trial (MRT) among smokers who have expressed a willingness to quit smoking:

Primary Aim Hypothesis

Administration of a prompt to perform a stress regulation exercise, as compared to no prompt, will reduce the likelihood of being stressed in the subsequent two hours, and this effect will be stronger if the prompt is administered when the individual is stressed.

The ***primary proximal outcome*** in the Sense2Stop study is trichotomous: it reflects the probability that each moment of time during the 120-minute period immediately following micro-randomization is classified as 'probably stressed', 'probably not stressed', or 'physically active'.

2. Implemented study protocol during lab visits

To provide context for the decisions described in this documentation, we provide details regarding the 1st and 2nd lab visits not covered in existing work on the Sense2Stop study (Battalio et al., 2021). Participants were asked to complete 3 lab visits during the study.

1st lab visit: A smartphone and wearable devices are loaned to participants. The smartphone contains pre-installed apps, including a Sense2Stop app and three additional apps focused on helping participants perform self-regulation exercises while wearable devices consist of a chest band and straps to be worn on the left and right wrist. Participants were oriented on study procedures, including that:

- Participants may access any of the apps focused on helping participants perform self-regulation exercises as often as they wanted.
- Pressing a 'start of day' button each day within the Sense2Stop app is required before the Sense2Stop app could send any ***Ecological Momentary Assessment (EMA)*** for that day. We note that existing work on the Sense2Stop study (Battalio et al., 2021) describe three possible types of EMAs that could have been triggered by the Sense2Stop app.
- Wearable devices should be worn at all times, except when sleeping. We note that the wearable devices continued to collect data regardless of whether the participant pressed the 'start of day' button, as long as both smartphone and wearable devices were not switched off.
- Wearable devices and smartphone must be returned during the 3rd lab visit.

During the 1st lab visit, study staff also (manually) pre-set 'wake time' and 'sleep time' on the smartphone in consultation with the participant. 'Wake time' refers to the time of day at which

the Sense2Stop app will automatically notify the participant to press the 'start of day' button within the Sense2Stop app. 'Sleep time' refers to the time of day at which the Sense2Stop app will automatically pause all notifications for that day until the next time the 'start of day' button is pressed. Incorporating 'wake time' and 'sleep time' into the study design ensured that the Sense2Stop app did not trigger EMAs when a participant was asleep.

Finally, participants were instructed by study staff to stop smoking the morning of their 2nd lab visit. In effect, participants were asked to view the date of their 2nd lab visit as their ***Quit Day***.

2nd lab visit: Study staff (manually) activated the Sense2Stop app's micro-randomization capabilities. Since this activation cannot be done remotely, only participants who completed their 2nd lab visit could have been micro-randomized.

For those participants who had the Sense2Stop app's micro-randomization capabilities activated, this caused the functionality of the 'start of day' button to change slightly: pressing the 'start of day' button each day within the Sense2Stop app is now required before the Sense2Stop app could send prompts to perform a self-regulation exercise **and** EMAs for that day.

Notably, study staff did not convey this change in functionality to participants. As long as participants did not notice that they can control the dose of the intervention by not pressing the start of day button, this results in a situation akin to 'blinding'.

3. Criteria for including vs. not including participants in all analyses

75 adult smokers between the ages of 18 and 65 years were enrolled in the Sense2Stop study. These participants met eligibility criteria for enrolment (see Battalio et al., 2021), including having met the study's definition of an ***active smoker*** – an individual who reported to have smoked one or more tobacco cigarettes per day for the past year.

Once enrolled into the study, (based on Institutional Review Board-approved study protocol; Spring, 2018), participants were subsequently removed from the study if they:

C1. Informed study staff either (a) before the scheduled date of their 2nd lab visit or (b) during the day they completed their 2nd lab visit that they wish to withdraw from the study.

C2. Did not complete their 2nd lab visit but did not inform study staff that they wish to withdraw.

Note that both C1 & C2 are prior to the start of the micro-randomized portion of the trial. Additionally, study staff conducted a 'pilot run' of study procedures on 5 of the 75 participants to identify potential barriers to fidelity to the intended study design and, when necessary, adjusted study procedures for all subsequent participants. Hence, participants were not included in all analyses if they

C3. Were part of the trial's 'pilot run'.

Finally, participants who were never micro-randomized were not included in all analyses. Specifically, participants were not included in all analyses if they:

C4. Had no micro-randomizations between their 'first day' and their 'last day', inclusive; 'first day' and 'last day' are defined below.

Table 1. Participants excluded entirely from all analyses.

Criterion violated	No. of participants	Participant IDs
C1	9	204, 209, 210, 220, 232, 236, 237, 239, 246
C2	3	201, 257, 263
C3	5	101, 102, 103, 104, 105
C4	9	206, 215, 217, 218, 230, 241, 247, 254, 270
Total	26	

From here onward, we will only focus on the remaining N = 75-(9+3+5+9) = 49 participants.

Definition of 'First Day' of the MRT, 'Last Day' of the MRT, and 'Quit Day'
First Day (***Day 0***): the date when a participant completed their 2nd lab visit
Last Day (***Day 10***): 10 days after the First Day
Quit Day: the date when a participant completed their 2nd lab visit, i.e., Quit Day coincides with First Day

Exception: One participant (Participant 213) informed study staff that they wish to withdraw from the study 5 days after they completed their 2nd lab visit. For this specific participant, the 'Last Day' will be Day 4.

What happens when participants delay in completing their 2nd lab visit
Only one participant delayed completing their 2nd lab visit while the rest completed their 2nd lab visit three days after completing their 1st lab visit. More specifically, Participant 251 completed their 2nd lab visit one day later than scheduled due to bad weather, i.e., four days after completing their 1st lab visit. For this specific participant, the 'First Day' and 'Quit Day' will still be the date when they completed their 2nd lab visit and 'Last Day' will still be 10 days after their First Day.

Subsequent implications of 'First Day' and 'Last Day'
Since data collection performed by wearable devices cannot be terminated by the study team remotely and participants may neglect to return wearable devices by their 'Last Day', any data-points providing information on their physiology or behavior during time periods beyond their 'Last Day' will not be included in analyses. Specifically, we will not include any data collected after 11:59PM on the participant's 'Last Day'.

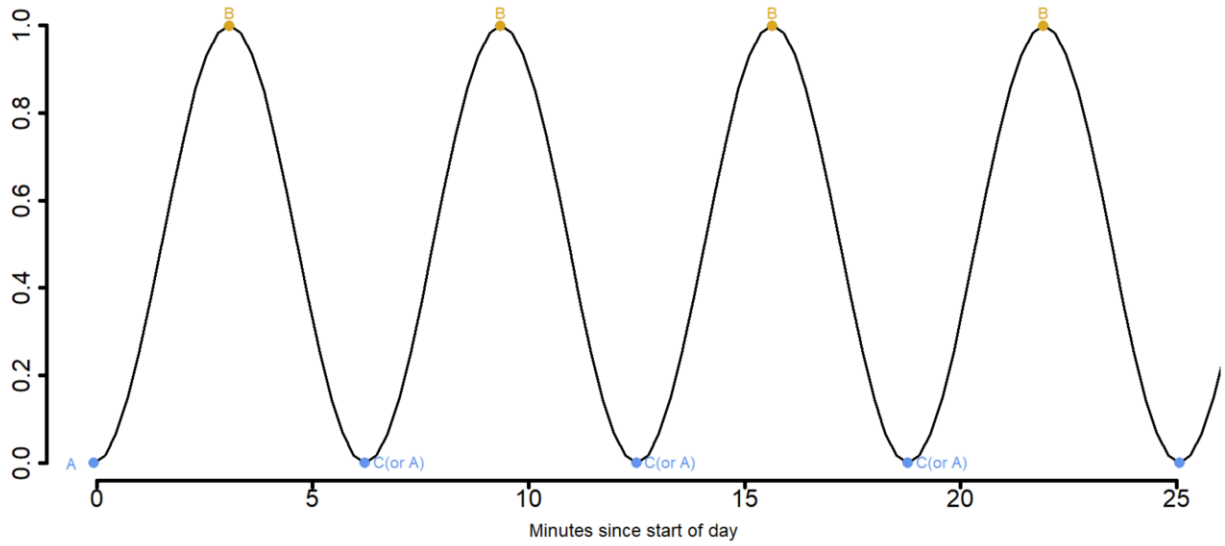
4. Working with sensor-derived assessments to construct the primary proximal outcome

Background

A detection algorithm developed by Sarker and colleagues (Sarker et al., 2016; Sarker et al., 2017) was used to predict when an individual might be experiencing stress, but with some modifications specific to the Sense2Stop study. During the course of the study, data collected by wearable devices were initially transformed into a ***stress likelihood time series***, a continuous-time measure for physiological arousal ranging between 0 to 1. Intuitively, this measure does not directly capture the intensity of physiological arousal, but rather, the likelihood that any physiological arousal occurred at a particular moment in time; a value closer to 1 (closer to 0) signifies that physiological arousal is more likely (less likely) to have occurred at that specific moment.

Subsequently, ***episodes***, defined as time intervals in which there is an increasing trend immediately followed by a decreasing trend in the stress likelihood time series, were constructed from the stress likelihood time series. Within any given episode, the stress likelihood time series thus takes the form of two successive valleys with a peak sandwiched in between (see Figure 1 below). We refer to the beginning, peak, and end of an episode occurred as 'A', 'B', and 'C', respectively.

Figure 1. A conceptual stress likelihood time series. In this figure, the text 'C (or A)' refers to the fact that C also marks the beginning (i.e., A) of a subsequent episode.



Finally, the area under the curve defined by the stress likelihood time series between A and B is then used to label the episode more specifically as one of three types:

- a ***probably stressed*** episode, when the area under the curve exceeds a pre-specified threshold *c* and good quality data is available for more than 50% of the time between A and B; or

- a ***probably not stressed*** episode, when the area under the curve lies below a pre-specified threshold *c* and good quality data is available for more than 50% of the time between A and B; or
- an ***unknown*** episode, either when good quality data is not available for more than 50% of the time between A to B or when a physical activity confound exists between A and B.

The code used for episode-labeling can be found here: https://github.com/MD2Korg/stream-processor/blob/master/src/main/java/md2k/mcerebrum/cstress/features/StressEpisodeClassification.java In the Sense2Stop study, c=0.36 (lines 47-48 in the linked code).

Episode type and specific time-stamps for A, B, C constitute the observed data which we will use as our starting point to construct the primary proximal outcome. For brevity, we will collectively refer to this data as the ***episode classification data stream*** from here onward.

The primary proximal outcome is based on the trichotomous variable $Y_{i,d,t}$, defined to have a value of 'no', 'yes', or 'active', if participant i is within a probably not stressed episode, within a probably stressed episode, or is physically active, respectively, at minute t of day d. Here, since the subscript t indexes each minute within a day, t ranges between 0 and 1439 (=24×60−1); since the subscript d indexes each day between 'First Day' and 'Last Day', d ranges between 0 and 4 (for Participant 213) or between 0 and 10 (for the rest of the 48 participants). In what follows, we will describe a procedure for constructing the trichotomous variable $Y_{i,d,t}$.

Procedure for constructing the trichotomous variable $Y_{i,d,t}$
The procedure was motivated by noticing that:

Observation 1. the episode classification data stream does not differentiate between episodes that are unknown due to the existence of a physical activity confound and those that are unknown due to poor data quality

Hence, we needed to bring in information from other data sources to help us distinguish between these two cases.

Observation 2. the duration of time between B and C exceeds 5 minutes in a sizeable number of probably stressed and probably not stressed episodes

This observation is not consistent with the duration of time between B and C that we expect to observe. We believe that the actual duration of time between B and C may be shorter than what the data (i.e., the episode classification data stream) suggests. Hence, we devised rules (see below) to censor episodes (all types).

Preliminary data cleaning steps. The following episodes were removed from the episode classification data stream:

1. Episodes whose time-stamp for A was recorded to have occurred on January 1, 1970[1].
2. Episodes regarded as ***duplicates***. Specifically, there exist episodes within the episode classification data stream which have identical time-stamps for A and B, but differ only in their time-stamp for C. Only the episode with the shortest duration of time between A and C was retained while all other episodes were regarded as duplicates.
3. Episodes for which B occurred either prior to 12am on the First Day or after 11:59pm on the Last Day.

After removing the above episodes from the episode classification data stream, we are left with 18451 episodes. Of these, 11606 were labelled as probably not stressed, 1074 were labelled as probably stressed, and 5771 were labelled as unknown. From here onward, we will only focus on the remaining 18451 episodes.

Working with unknown episodes. We use the rule in Box 1 which leverages physical activity data. In brief, physical activity data is represented as a minute-level indicator for whether physical activity was detected (=1) or not (=0) by an activity detection algorithm.

Box 1. Rule for labeling unknown episodes as physically active episodes

1 IF physical activity was detected in more than 50% of minutes between A to B
2 THEN regard the time between A to C as a physically active episode.
3 ELSE regard the information on the trichotomous variable $Y_{i,d,t}$ between A and C as missing.

After applying the rule in Box 1, we found that 2515 of the 5771 unknown episodes could be regarded as physically active episodes. Table 2 summarizes the number of episodes of each type after applying the rule in Box 1.

Table 2. No. of episodes of each type after applying the rule in Box 1

Episode Type	No. of Episodes	Percent of Total
'yes'	1074	5.8
'no'	11606	62.9
'unknown' episodes now regarded as physically active episodes (i.e., 'active')	2515	13.6
Subtotal	15195	
'unknown' episodes for which the information on the trichotomous variable $Y_{i,d,t}$ between A and C is now regarded as missing	3256	17.6
Total	18451	100

From here onward, we will focus on the 1074+11606+2515=15195 episodes.

[1] Specifically, these were episodes whose time-stamp for A was recorded to have occurred on 12:00am, January 1, 1970 in Universal Coordinated Time (UTC).

Censoring episodes. Initial inspection of the 1074 and 11606 probably stressed and probably not stressed episodes, respectively, show that the duration of time between B and C exceed 5 minutes in a sizeable number. More specifically, Table 3 shows that 678 probably stressed episodes exceed 5 minutes, with 10% exceeding 11.8 minutes (out of the 1074 in Table 2); Table 3 also shows that 5006 probably not stressed episodes exceed 5 minutes, with 10% exceeding 9.02 minutes (out of the 11606 in Table 2). The maximum recorded time between B and C is 664 minutes (about 11 hours) and 904 minutes (about 17 hours) for probably stressed and probably not stressed episodes, respectively.

For completeness, Table 3 also displays median, 90th percentile, and max duration of time between B and C for physically active episodes. Inspection of the 2515 physically active episodes show that 1386 episodes exceed 5 minutes; the maximum duration of time between B and C was 3308 minutes (about 55 hours).

Since these summary statistics are not consistent with the duration of time between B and C that we expect to observe for probably stressed, probably not stressed, or even physically active episodes, we believe that the actual duration of time between B and C may be shorter than what the data (i.e., the episode classification data stream) suggests. Information on data quality after B (in all episode types) was not used when determining the specific time-stamp for C (in all episode types); hence, the long duration of time we sometimes observe between B and C.

Table 3. Of the 15195 episodes in Table 2, we display the no. of episodes for which time between B and C exceed minutes (column 2), median time in minutes between B and C (column 3), 90th percentile time in minutes between B and C (column 4), and max time in minutes between B and C (column 5)

Episode Type	No. of episodes	Median	90th percentile	Max
‘yes’	678	6.01	11.80	664.99
‘no’	5006	4.01	9.02	904.00
‘active’	1386	5.99	12.99	3308.00
Total	7070			

We use the rule in Box 2 for assigning a new value for C in episodes (any type) which exceed 5 minutes. The rule in Box 2 censors episodes (any type) at C*, the specific time-point between B and C, inclusive, that we view as the cut-point beyond which we are uncertain whether the individual continued to be stressed, not stressed, or physically active.

The rule in Box 2 capitalizes on heart rate data, represented as the average number of beats over a 1-minute-long window; heart rate data was one of several sensor-derived assessments used towards constructing the stress likelihood time series (Figure 1) for this study.

Employing the rule in Box 2 requires constructing a minute-by-minute binary indicator (i.e., for each minute *t* of day *d*) for whether heart rate data was observed. When no heart rate data was observed at all between B and C of a particular episode, the episode was censored at B (i.e., C*=B); this scenario occurred in 1932 (=446+1486) of the 7070 episodes in Table (see Table 4)

and could have resulted from not all of the heart rate data generated during the conduct of the trial having been 'saved' for future data analysis.

Table 4. Of the 7070 episodes in Table 3, in how many have we been able to observe heart rate data between A and B **and/or** between B and C?

Scenario	No. of episodes (all types)	Percent of Total (all types)
Heart rate data was observed between A and B **and** also observed between B and C	4774	67.52
Heart rate data was not observed between A and B **but** observed between B and C	364	5.15
Heart rate data was observed between A and B **but** not observed between B and C	446	6.31
Heart rate data was not observed between A and B **and** also not observed between B and C	1486	21.02
Total	7070	100

Box 2. Rule for censoring episodes (any type) when time between B and C exceed 5 minutes

1 IF no heart rate data was observed between B and C
2 THEN censor the episode at C*=B AND regard the information on the trichotomous variable $Y_{i,d,t}$ between C* and C as missing.
3 ELSE IF some heart rate data was observed between B and C, BUT there is a period of at least five consecutive minutes within B and C having no observed heart rate data
4 THEN censor the episode at the beginning of this no heart rate period (which we denote by C*) AND regard the information on the trichotomous variable $Y_{i,d,t}$ between C* and C as missing.
3 ELSE do not censor the episode

A notable number of episodes for which time between B and C exceeds 5 minutes (i.e., the 7070 episodes in Table 3) were censored after applying the rule in Box 2. Specifically,

- 69% of the 678 probably stressed episodes whose time between B and C exceeds 5 minutes were censored
- 46% of the 5006 probably not stressed episodes whose time between B and C exceeds 5 minutes were censored
- 36% of the 1386 physically active episodes whose time between B and C exceeds 5 minutes were censored

Summary statistics are also displayed in Figure 2 and Table 3.

Figure 2. Of the episodes whose time between B and C exceeds 5 minutes (reported in Table 3), the percentage which were censored after applying the rule in Box 2.

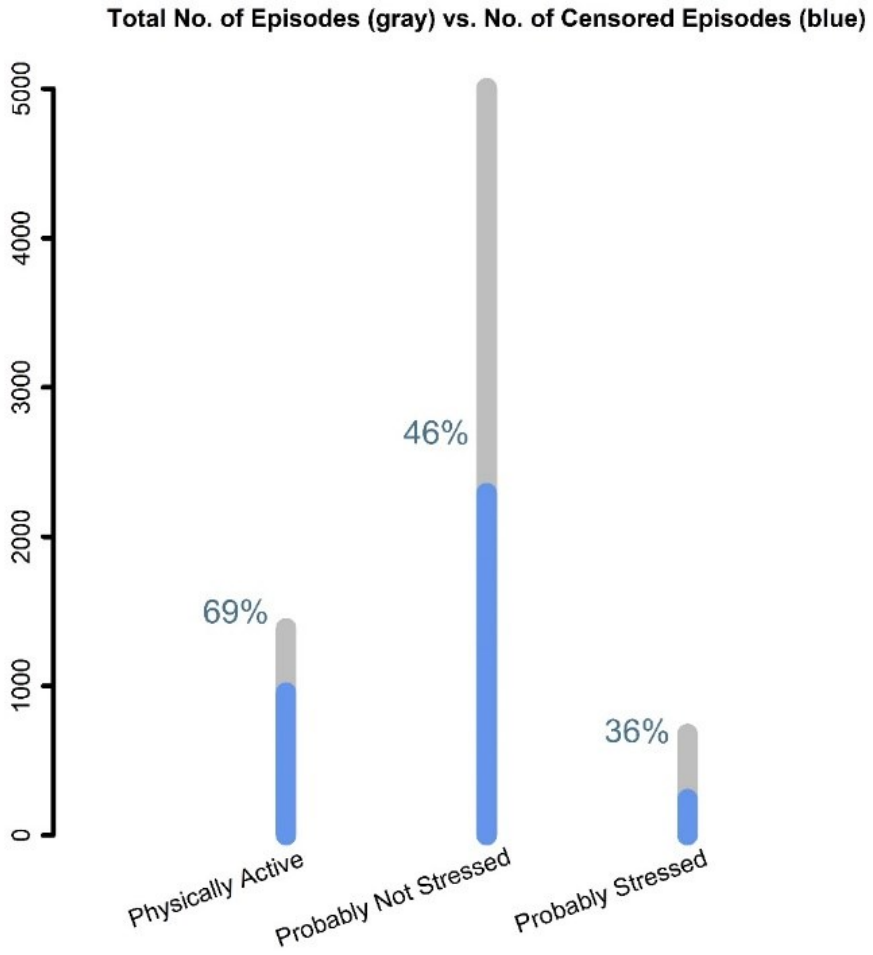


Table 3. Of the episodes reported in Table 3, we display the no. of episodes for which the duration of time between B and C exceed 5 minutes (column 2). After applying the rule in Box 2, we display the median time in minutes between B and C* (column 3), 90th percentile time in minutes between B and C* (column 4), and max time in minutes between B and C* (column 5).

Episode Type	No. of episodes	Median	90th percentile	Max
'yes'	678	6.01	11.00	18.02
'no'	5006	5.01	9.99	18.01
'active'	1386	0.00	10.00	20.04
Total	7070			

Thus far, we have focused on the time between B and C. Turning our attention to the total duration of time of episodes, we find that even after applying the rule in Box 2, that the actual duration of time between A and C may be shorter than what the data suggests. Below, we discuss

summary statistics supporting this view. From here onward, by convention, we use the notation C* to denote:

- the end of episodes whose time between B and C exceeded 5 minutes, but were not censored after applying the rule in Box 2; and
- the end of episodes whose time between B and C exceeded 5 minutes, but were censored after applying the rule in Box 2; and
- the end of episodes whose time between B and C did not exceed 5 minutes (and hence, Box 2 did not apply to these episodes)

We calculated the median, 90th percentile, and maximum time between A and C* by episode type. Table 4 displays the result of this calculation; we observe that the max time between A and C* is 4116 minutes (about 69 hours) among probably stress episodes, 809 minutes (about 13 hours) among probably not stressed episodes, and 75 minutes (about 1.25 hours) among physically active episodes.

Table 4. Of the 15195 episodes reported in Table 2, we display the no. of episodes by type (column 2). After applying the rule in Box 2, we display the median time in minutes between A and C* (column 3), 90th percentile time in minutes between A and C* (column 4), and max time in minutes between A and C* (column 5).

Episode Type	No. of episodes	Median	90th percentile	Max
'yes'	1074	8.01	15.99	4116.01
'no'	11606	7.99	17.02	809.00
'active'	2515	9.00	18.86	75.00
Total	15195			

Hence, we further censor episodes based on a specified threshold. More specifically, if the time between A and C* of an episode (i.e., any of the 15195 in Table 4) exceeds 17 minutes, we censor the episode exactly 17 minutes after A; in other words, a new value for C* is selected and this new value is 17 minutes after the beginning of an episode. The threshold we selected (i.e., the 17 minutes) was the 90th percentile of the time between A and C*, calculated using all episode types (i.e., 'marginal' over episode type). Table 5 displays the median, 90th percentile and max time between A and C* (in minutes) after applying further censoring (i.e., the 17-minute threshold).

Table 5. Of the 15195 episodes reported in Table 2, we display the no. of episodes by type (column 2). After applying further censoring, we display the median time in minutes between A and C* (column 3), 90th percentile time in minutes between A and C* (column 4), and max time in minutes between A and C* (column 5).

Episode Type	No. of episodes	Median	90th percentile	Max
'yes'	1074	8.01	15.99	17.00
'no'	11606	7.99	17.00	17.00
'active'	2515	9.00	17.00	17.00
Total	15195			

Figure 3. Of the 15195 episodes reported in Table 2, we display the percentage of episodes which were censored, either via the rule in Box 2 and/or using the 17-minute threshold.

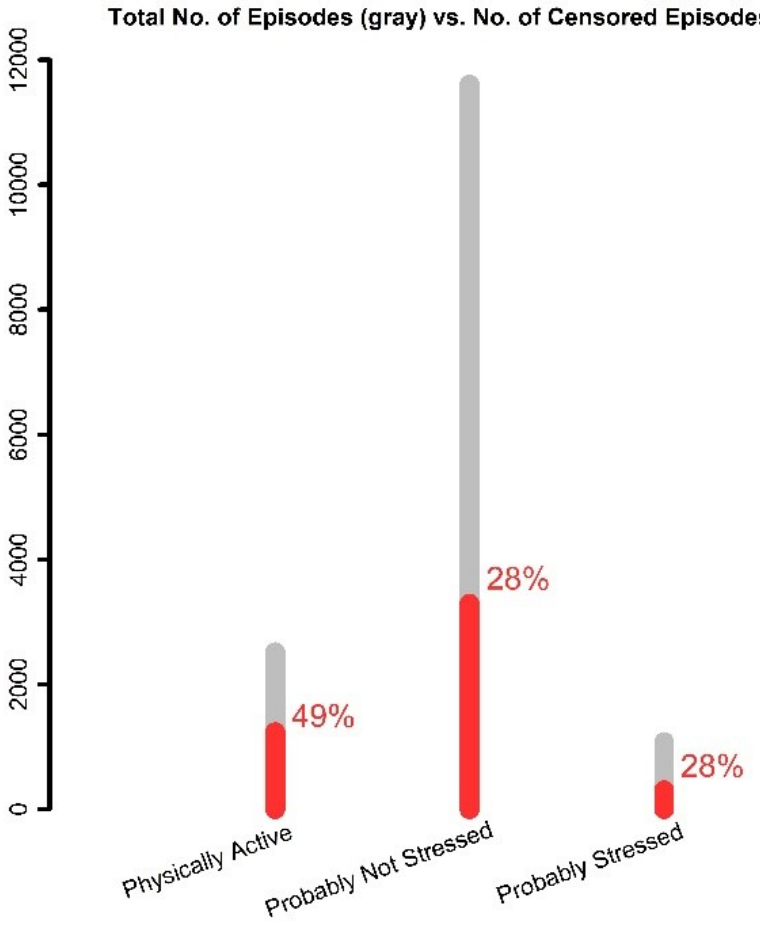
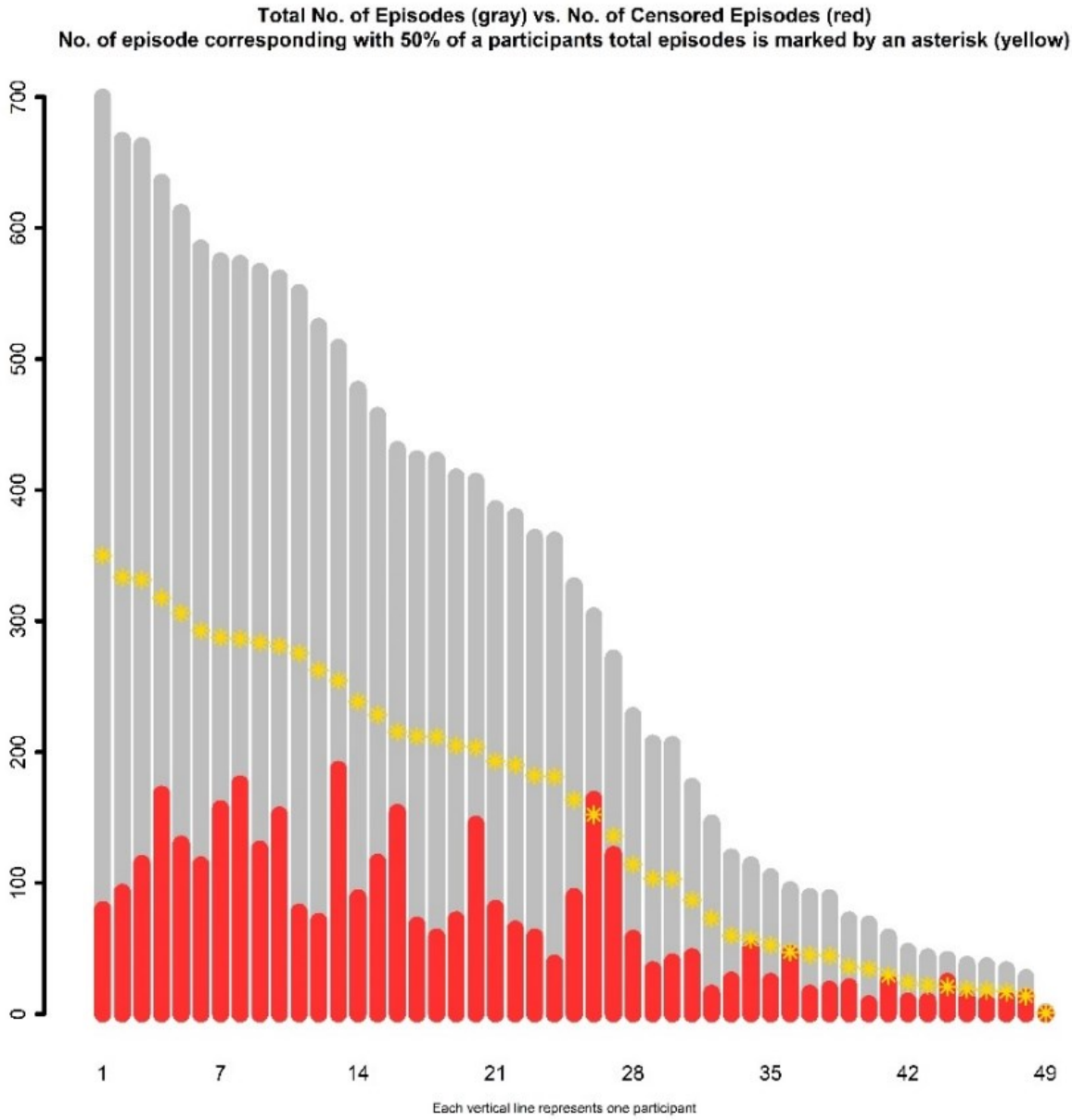


Figure 4. Of the 15195 episodes reported in Table 2, we display the no. of episodes which were censored, by participant. The figure shows that for the vast majority of participants, less than 50% of their episodes were censored (depicted by red bars lying below the yellow asterisks).



Observed values in the trichotomous variable $Y_{i,d,t}$. Finally, we may now regard participant *i* at minute *t* of day *d* as:

- ***probably stressed*** (i.e., assign a value of ***'yes'*** to the trichotomous variable $Y_{i,d,t}$) if participant *i* was within a physically active episode at minute *t* of day *d*
- ***probably not stressed*** (i.e., assign a value of ***'no'*** to the trichotomous variable $Y_{i,d,t}$) if participant *i* was within a physically active episode at minute *t* of day *d*
- ***physically active*** (i.e., assign a value of ***'active'*** to the trichotomous variable $Y_{i,d,t}$) if participant *i* was within a physically active episode at minute *t* of day *d*

Missing values in the trichotomous variable $Y_{i,d,t}$. The trichotomous variable $Y_{i,d,t}$ is said to have a missing value if we are not able to assign a value of 'yes', 'no', or 'active' to $Y_{i,d,t}$ using the rules described above.

5. Micro-randomizations

Background
A participant was regarded as ***eligible for micro-randomization*** at minute *t* of day *d* if they met a pre-specified set of criteria, which were described by Battalio and colleagues (Battalio, et al., 2021). In addition to these pre-specified criteria, micro-randomization did not occur at minute *t* of day *d* if (a) the participant did not press the 'start of day' button on that day; and/or (b) *t* is after 'sleep time' on that day; and/or (c) *t* is before activation of the Sense2Stop app's micro-randomization capabilities on the participant's 'First Day', e.g., if activation occurred at 3pm, all minutes prior to 3pm on the participant's 'First Day' were regarded as ineligible for micro-randomization.

Micro-randomization at each eligible minute *t* of day *d* was ***stratified*** – that is, the probability of an eligible minute *t* of day *d* being micro-randomized to a prompt versus no prompt was balanced according to several factors (Battalio, et al., 2021), which included, importantly, whether an eligible minute *t* of day *d* was within a probably stressed episode or within a probably not stressed episode. Randomization probabilities within each stratum were calibrated to lead to higher average number of prompts delivered across all eligible minutes within probably stressed episodes than across all eligible minutes within probably not stressed episodes.

The code for calculating the randomization probabilities can be found here https://github.com/MD2Korg/mCerebrum-EMAScheduler/blob/master/ema_scheduler/src/main/java/org/md2k/ema_scheduler/scheduler/emi/ProbabilityEMI.java

Let *t** denote a particular eligible minute of participant *i*'s day *d* as determined either by pre-specified criteria (Battalio, et al., 2021) or conditions (a) – (c) described above. Data collection in the study yielded information associated with micro-randomization at *t**. These included:

- **Data Source #1:** the specific timestamp[2] at which minute *t** of participant *i*'s day *d* was deemed eligible for micro-randomization

[2] This timestamp was originally represented as milliseconds elapsed since 12am of January 1, 1970 in Universal Coordinated Time (UTC)

- **Data Source #2:** the probability associated with t^*, i.e., the probability that a prompt would be delivered
- **Data Source #3:** the stratum associated with t^*, i.e., whether the probability associated with t^* (in Data Source #2) was calculated based on a formula that was calibrated for probably stressed episodes or a formula that was calibrated for probably not stressed episodes
- **Data Source #4:** the decision associated with t^*, i.e., whether randomization resulted in a decision to deliver any prompt or no prompt

We will use $\Omega_{i,d}$ to denote the set of all eligible minutes of participant i's day d. Further, we use Ω to denote the collection of all sets $\Omega_{i,d}$ across all participants and across all days, i.e., $\Omega = \{\Omega_{i,d} :$ all participants i and days $d\}$. Hence, if $I_{i,d,t}$ denotes whether ($I_{i,d,t}=1$) or not ($I_{i,d,t}=0$) participant i was eligible at minute t on day d for micro-randomization, $I_{i,d,t}=1$ if t **is in** the set $\Omega_{i,d}$ and $I_{i,d,t}=0$ if t **is not in** the set $\Omega_{i,d}$.

For the rest of Section 5, we will focus on Data Sources #1 and #2, and ignore Data Sources #3 and #4. We will return to Data Source #3 and #4 in Sections 6 and 7.

Constructing the set Ω

Data Source #1 contains 5506 timestamps corresponding to each t^* between 'First Day' and 'Last Day' across the 49 participants (refer back to Section 1). In other words, there were 5506 micro-randomizations between 'First Day' and 'Last Day' among the 49 participants.

Although, at first sight, it may appear reasonable to solely use the information in Data Source #1 to construct the set Ω, considering our constructed trichotomous variable $Y_{i,d,t}$ and the intended study design in combination with Data Source #1 suggests otherwise.

Our construction of the set Ω was motivated by observing that there were minutes t^* for which:

Observation 1. the time (denoted by s) of the most recent minute prior to t^* having any observed classification may have been such that time between s and t^* exceeds more than a few minutes

This observation is inconsistent with the fact that the classification derived from the stress likelihood time series and the subsequent micro-randomization based on this classification were intended by study designers to be almost instantaneous of each other.

Observation 2. the associated micro-randomization may have occurred even if the individual was neither within a probably stressed episode nor within a probably not stressed episode at s, and further, a classification prior to t^* may not always have been observed even if micro-randomization at t^* occurred

This observation is inconsistent with the fact that micro-randomizations were intended by study designers to only occur when there was data of sufficient quality to inform the calculation of the stress likelihood time series (i.e., a

classification must exist prior to t) and only when an individual was within a probably stressed episode or probably not stressed episode (i.e., never when an individual was in an unknown episode).*

Observation 3. the associated randomization probabilities may either have been exactly zero or exactly one in a substantial number of cases; this observation occurred only for minutes t^* associated with probably not stressed episodes

Even though randomization probabilities (specifically, permitted within probably not stressed episodes, but not permitted within probably stressed episodes) were permitted by study designers to be exactly zero or exactly one, this observation is inconsistent with the fact that this situation was expected by study designers to be extremely rare during the conduct of the trial. Further, randomization probabilities of exactly zero or exactly one violate the so-called 'positivity assumption' inherent in many intent-to-treat estimators, including the estimator to be utilized in estimating the treatment effect for this trial's primary aim.

We will regard the above-described values of t^* as minutes when an individual should also be ineligible for micro-randomization.

For each participant i's day d, the rule in Box 3 summarizes the rules utilized to construct the set $\Omega_{i,d}$. **Only the minutes t^* which remain after employing the rule in Box 3 (i.e., which remain in the set $\Omega_{i,d}$ by line 19 of Box 3) will be regarded as eligible for micro-randomization when estimating the treatment effect for the primary aim.**

The rule in Box 3 includes an additional criterion on the minutes t^*: if the probability associated with t^* recorded in Data Source #2, denoted by p_{i,d,t^*}, was either below or above thresholds k_1 and k_2, respectively, then t^* will be excluded from the set $\Omega_{i,d}$. The rationale for this additional criterion and selection of thresholds k_1 and k_2 will be described in Section 8.

In Box 3, the notation $\aleph_{i,d}$ was used to distinguish from the minutes t^* which we additionally exclude based on thresholds k_1 and k_2 from those minutes t^* which were not. In the following sections, we will let $\aleph$ denote the collection of all sets $\aleph_{i,d}$ across all participants and across all days, i.e., $\aleph = \{ \aleph_{i,d}$: all participants i and days $d\}$.

This distinction in notation will permit us to assess the incremental impact of employing the additional criterion on the number of minutes we ultimately will regard as eligible for micro-randomization when estimating the treatment effect for the primary aim. In presenting summary statistics in the following sections, we will be clear as to whether the summary statistics apply to minutes t^* within $\aleph_{i,d}$ or within $\Omega_{i,d}$.

Box 3. For each participant i's day d, a rule for constructing $\aleph_{i,d}$ and $\Omega_{i,d}$

Initialize an empty set $\aleph_{i,d} = \{\}$
For each t^* in Data Source #1 belonging to participant i on day d:
    Determine the closest minute s prior to t^* for which $Y_{i,d,t}$ was observed.
    IF s exists
        IF $t^* - s \leq$ 5 minutes
            IF $Y_{i,d,t}$ was either 'yes' or 'no'
                THEN include t^* in $\aleph_{i,d}$
            ELSE do not include t^* in $\aleph_{i,d}$
        ELSE do not include t^* in $\aleph_{i,d}$

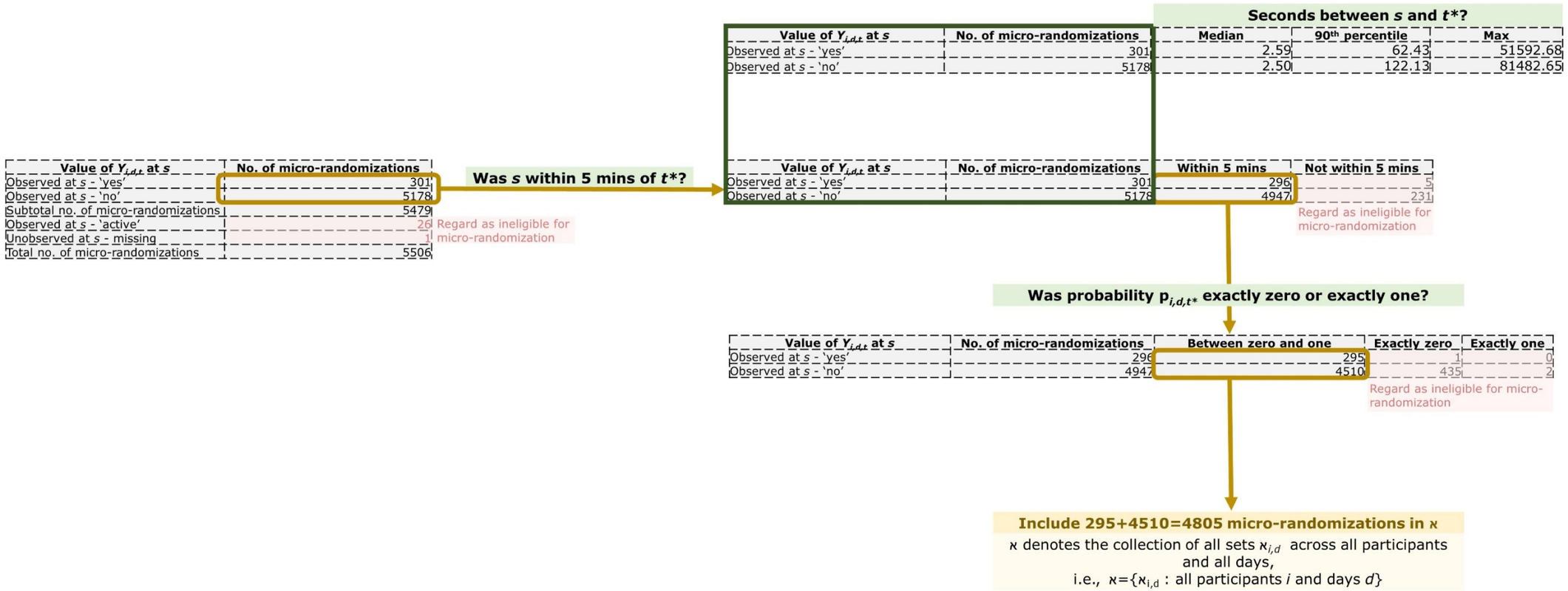
For each t^* in $\aleph_{i,d}$ belonging to participant i on day d:
    IF p_{i,d,t^*} is either exactly zero or exactly one
        THEN remove t^* from $\aleph_{i,d}$
    ELSE retain t^* in $\aleph_{i,d}$

Initialize the set $\Omega_{i,d}$ as $\Omega_{i,d} = \aleph_{i,d}$
Set the value of thresholds k_1 and k_2
For each t^* in $\Omega_{i,d}$ belonging to participant i on day d:
    IF $p_{i,d,t^*} < k_1$ OR $p_{i,d,t^*} > k_2$:
        THEN remove t^* from $\Omega_{i,d}$
    ELSE retain t^* in $\Omega_{i,d}$

END rule with $\aleph_{i,d}$ and $\Omega_{i,d}$ as outputs.

After employing the rule in Box 3, we found that 701 out of the 5506 minutes (about 13%) t^* were deemed ineligible, leaving 4805 minutes (about 87%) t^* included in $\aleph$. In other words, the 4805 minutes pertain to those minutes t^* deemed ineligible for micro-randomization due to either Observation 1, 2 or 3; summary statistics are displayed in Figure 5.

Figure 5. Number of minutes t^* deemed ineligible due to either Observation 1, 2 or 3



6. Intervention decision for each minute between 'First Day' and 'Last Day'

Let $A_{i,d,t}$ be a dichotomous variable denoting whether the decision for participant *i* at minute *t* on day *d* was to deliver any prompt ($A_{i,d,t}$ =1) or no prompt ($A_{i,d,t}$ =0). Box 4 displays the rule used to construct $A_{i,d,t}$, applied to each minute *t* between participant *i*'s 'First Day' and 'Last Day'.

Box 4. Rule for constructing $A_{i,d,t}$

```
For each participant i:
    For each minute t between 'First Day' and 'Last Day'
        IF no decision associated with t was recorded in Data Source #4 (see Section 5)
            THEN set A_{i,d,t} =0
        ELSE set the value of A_{i,d,t} according to the recorded decision (i.e., A_{i,d,t} =1 if the recorded decision was to send any prompt; A_{i,d,t} =0 if the recorded decision was to send no prompt)
```

We note that whenever pre-specified criteria (Battalio, et al., 2021) or conditions (a) – (c) for eligibility for micro-randomization described in Section 5 were met, software was designed to deliver no prompt. Such cases were represented by the absence of recorded information in Data Source #4 regarding such minutes *t*. In these cases, the value of $A_{i,d,t}$ was set to 0.

We note that it is possible for the software to decide to deliver a prompt at minutes *t* which we ultimately regarded as ineligible for micro-randomization by employing the rule in Box 3. Hence, it is possible to observe $A_{i,d,t}$ =1 even if $I_{i,d,t}$ =0.

7. Stratification of micro-randomizations

Background:
The testing the primary aim hypothesis involves estimating the following two quantities:

1. the effect of a prompt to perform a stress regulation exercise, as compared to no prompt, if prompts were administered during a probably stressed episode
2. the effect of a prompt to perform a stress regulation exercise, as compared to no prompt, if prompts were administered during a probably not stressed episode

Hence, to facilitate analysis, we need to construct an indicator $X_{i,d,t}$ for whether micro-randomization (assumed to have occurred instantaneously) at *t* was within a probably stressed episode or a probably not stressed episode.

Constructing $X_{i,d,t}$:
The data collected from the study permits the following two approaches to constructing $X_{i,d,t}$. The two approaches should not necessarily be expected to always yield identical values $X_{i,d,t}$, due to the possibility of deviations in the software implementation from the intended study design.

Approach 1. When participant *i* was eligible for micro-randomization at a minute *t** of day *d*, **we define the variable $X_{i,d,t}$ to be the classification of our constructed**

trichotomous variable $Y_{i,d,t}$ at time *s*, namely, the closest minute prior to *t for which $Y_{i,d,t}$ was observed.**

Approach 2. When participant *i* was eligible for micro-randomization at a minute *t** of day *d*, **we define the variable $X_{i,d,t}$ to be the stratum associated with *t** as recorded in Data Source #3.**

Using the 4805 micro-randomizations in the set א, we compared both approaches to constructing $X_{i,d,t}$. We found that the resulting value of $X_{i,d,t}$ only differ in 14 (about 0.3%) out of the 4805 micro-randomizations included in א. In other words, in the vast majority of cases (4791 out of the 4805, or about 99.7%, micro-randomizations included in א), both approaches yielded identical values for $X_{i,d,t}$. **We utilized Approach 1 in constructing $X_{i,d,t}$.**

8. Excluding minutes *t** from $\Omega_{i,d}$ based on thresholds on p_{i,d,t^*}

Recall that the rule in Box 3 includes an additional criterion on the minutes *t**: if the probability associated with *t** recorded in Data Source #2, denoted by p_{i,d,t^*}, was either below or above thresholds k_1 and k_2, respectively, then *t** will be excluded from the set $\Omega_{i,d}$. Regarding minutes *t** which do not meet this criterion as ineligible for micro-randomization has the effect of omitting them from the calculation of the weights in the estimator to be utilized in estimating the treatment effect for this trial's primary aim.

If we select the thresholds $k_1 = 0.05$ and $k_2 = 0.95$, identical to thresholds already implemented by the software for minutes *t** associated with probably stressed episodes, we find that there are 668 minutes *t** (or about 14% of 4805) for which $p_{i,d,t^*} < k_1$ OR $p_{i,d,t^*} > k_2$ in א; hence, 4137 minutes *t** (or about 86% of 4805) will be included in Ω.

References

[1] Battalio, S. L., Conroy, D. E., Dempsey, W., Liao, P., Menictas, M., Murphy, S., …, and Spring, B. (2021). Sense2Stop: A micro-randomized trial using wearable sensors to optimize a just-in-time adaptive stress management intervention for smoking relapse prevention. Contemporary Clinical Trials, 109, 106534

[2] Spring, B. (2018). Protocol Title: Sense2Stop: Mobile Sensor Data to Knowledge (Version Number 24). Internal report (Northwestern University Institutional Review Board-approved study protocol).

[3] Sarker, H., Tyburski, M., Rahman, M. M., Hovsepian, K., Sharmin, M., Epstein, D. H., ... and Kumar, S. (2016, May). Finding significant stress episodes in a discontinuous time series of rapidly varying mobile sensor data. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 4489-4501).

[4] Sarker, H., Hovsepian, K., Chatterjee, S., Nahum-Shani, I., Murphy, S. A., Spring, B., ..., and Kumar, S. (2017). From markers to interventions: The case of just-in-time stress intervention. In Mobile health (pp. 411-433). Springer, Cham.

[5] Hovsepian K, al'Absi M, Ertin E, Kamarck T, Nakajima M, and Kumar S. cstress: towards a gold standard for continuous stress assessment in the mobile environment. ACM UbiComp. 2015:493–504.
[6] Hossain, S. M., Hnat, T., Saleheen, N., Nasrin, N. J., Noor, J., Ho, B. J., ..., and Kumar, S. (2017, November). mCerebrum: a mobile sensing software platform for development and validation of digital biomarkers and interventions. In proceedings of the 15th ACM Conference on Embedded Network Sensor Systems (pp. 1-14).
[7] Boruvka, A., Almirall, D., Witkiewitz, K., & Murphy, S. A. (2018). Assessing time-varying causal effect moderation in mobile health. Journal of the American Statistical Association, 113(523), 1112-1121.