# **Primary Analysis: Sense2Stop Micro-randomized Trial**

**Last updated: 20 April 2020**

**<Current draft uses smoking classifications produced by online model(s)>**

## **Introduction**

### In this document, we present results from a 10-day MRT conducted to optimize a novel JITAI developed to help individuals manage their stress by engaging in stress-regulation exercises at times of heightened stress.

## **Methods**

### ***2.1 The Sense2Stop Study***

* 1. ***Decision Points***

There are 7200 (60 minutes x 12 hours x 10 days) *candidate decision points* per participant during the micro-randomized trial of the Sense2Stop study. A *generated* *decision point* is generated out of a candidate decision point, if at the candidate decision point enough good quality data[[1]](#footnote-1) was available in the moment to detect that the participant was at the peak of either an episode detected as stressed or an episode not able to be detected as stressed.

### ***2.3 Availability***

A participant was randomized to receive a reminder to practice stress management exercises if, at a generated decision point, they were considered to be *available* for treatment and none of the following holds: they were currently driving, physically active, their phone battery was less than 10%, they received an ecological momentary assessment (EMA) in the last 10 minutes or an intervention reminder in the last hour[[2]](#footnote-2). If the participant is available at a generated decision point, it is referred to as an *available decision point* (see Table 4 in Section S.1 of the Supplement (at end of this draft) for the distribution of all generated decision points and available decision points during the MRT phase of the Sense2Stop study).

### ***2.4 Participants***

We recruited individuals between the ages of 18 and 65 years who reside within the 25-mile radius of the Chicagoland area and were active smokers of 1 or more tobacco cigarettes a day for the past year. Participants expressed willingness to quit smoking for at least 48 hours during the 15-day quit trial phase of the Sense2Stop study.

### ***2.5 Study Design***

This study is registered on clinicaltrials.gov as NCT03184389.

### ***2.6 Randomization***

At each available decision point, the randomization was independent of prior randomizations and the participants’ responses to previously delivered stress-management exercise reminders. Given the micro-randomized study duration of 10 days, stress-management exercise reminders could be randomized up to 7200 (60 minutes x 12 hours x 10 days) times for each participant. As we wanted reminders each day with equal representation of delivery at times when the participant is detected to be stressed and not able to be detected as being stressed, we used the seqRTS algorithm (cite just-in-time paper).

Randomization probabilities were such that participants should receive an average of 1.5 EMIs per day within pre-lapse days during both stress and not-able-to-classify-as-stress episodes, 1 EMI per day within post-lapse days for stress episodes and 1.5 EMIs per day within post-lapse days for not-able-to-classify-as-stress episodes.

### ***2.7 Proximal Outcome***

As reminders to practice stress-management exercises were intended to help participants reduce their physiological stress close to the time that they were delivered, we operationalized their intended proximal effect as follows.

Define

S = Number of minutes detected as stressed in 120 minutes following an available decision

point

NS = Number of minutes not able to detect as stressed in 120 minutes following an available

decision point

A = Number of minutes detected as physically active in 120 minutes following an available

decision point

The proximal outcome variable takes on three values. Specifically,

Y = {“detected as stressed” if max(S, NS, A) = S,

“not able to detect as stressed” if max(S, NS, A) = NS,

“physically active” if max(S, NS, A) = A}.

### ***2.8 Statistical Analysis***

All analyses were performed using a centered and weighted least squares method (Boruvka et.al., 2018). This method estimates parameters in models for the treatment effects and allows robust (without incurring bias in the treatment effect model) inclusion of covariates to reduce noise. The method is similar to generalized estimating equation and multi-level models in that it accommodates the nested nature of the data (decision points nested within participants) and associated within-participant correlation across time in the outcome. Furthermore, the method takes advantage of the sequential randomization to estimate causal treatment effects. See the Supplementary Material for technical details. We prespecified three primary analyses *(1)*: test for an overall (average, across all time points) effect of delivering a reminder versus providing no reminder on the subsequent 2-hour detected likelihood of stress; *(2)* test for decreasing effect with day in the study; and *(3)* test whether the reminder effect is larger when provided to those who are detected as being stressed vs those who are not able to be detected as stressed. To reduce noise, all analyses were adjusted for a pretreatment detection of stress, operationalized as the detected likelihood of stress in the 2-hours prior to the decision point. See Supplementary Material at the end of this document for full analysis details.

**Need to confirm this with team**: Power analysis was not conducted and instead the pre-determined number of participants in this study was based on the number of devices and budget.

Analyses were conducted using R version 3.6.1 (R Core Team, 2013).

## **Results**

### ***3.1 Participant Sample***

We recruited 70 participants in the Chicagoland, IL, area (see Table 1 for sample characteristics). Of these, we excluded all data from 21 who did not generate any decision points, i.e., there was no point during the 10 days in which the participants were detected to be at the peak of either a detected stressed episode or an episode that was not detected to be a stressed episode. Of these 21 participants, 10 had dropped out early in the study (after 1, 3, 4 and 9 days of the 14 day study). Thus, 49 participants remained in the sample (see Fig. 1 for the CONSORT diagram).

A red traffic light

Description automatically generated

Figure 1: CONSORT diagram.

Table 1: Baseline characteristics of Sense2Stop participants.



### ***Included Decision Points***

Each of the 49 participants generated an average of 198 decision points (average of total decision points per participant) over the course of the 10 day micro-randomized trial, with a daily median as indicated by the boxplots in Figure 2.

A screenshot of a cell phone

Description automatically generated

Figure 2: Boxplots representing the number of generated decision points per day.

* 1. ***Available Decision Points***

There were 9704 total generated decision points across all days of the micro-randomized trial phase of the Sense2Stop study and across all of the 49 participants. Of these, 5044 were available decision points. Across the 5044 available decision points, 630 of these were used to deliver reminders to practice stress-management exercises (12.5%). On average across all participants, participants received 1.3 reminders each day (*SD* = 0.9). For a breakdown of the average number of reminders sent to participants (across all participants) per day based on episode type and lapse phase see Table 2 below (Tables 6 and 7 on how Table2 was calculated).

Table : Average number of reminders sent to participants per day.



**What Decision Points to include in the analysis?**

We propose to include all generated decision points that do not correspond to a red or yellow shaded day in Table 3.

* 1. ***Control Covariates***

We include control covariates in the analysis to reduce the noise/variance and thus be better able to detect the effect of sending of the reminder to practice stress-reduction exercises on the subsequent 2-hour indication of stress. The control covariates we consider are:

* previous 2-hour detection of stress
* has the participant lapsed
  + **definition of lapse vs relapse**
    - A **smoking relapse** is resumption of regular smoking after a period of abstinence
    - A smoking relapse is frequently preceded by isolated smoking incidents known as **lapses**.
    - A lapse (or smoking episode) according to puffMarker output is detection of 4 consecutive puffs.
    - Stray puffs from puffMarker output are regarded as false positives.
  + **which data stream to use for detection of lapse and relapse**

Self-reported smoking was used for determination of pre-lapse/ post-lapse during an available decision time in the online cStress algorithm used for the Sense2Stop study.

**MM**: Check self-reported smoking against record of pre-lapse/post-lapse in phone log files to fill in gaps for lapse at an available decision time.

* days since first lapse
* day in study
* type of current episode
  1. ***Data Missingness***

We decided that the **primary outcome** is a triad: (#min. classified as stressed, #min. not-able-to-detect-as-stressed, #min. physically active). If there were no missing data then these three counts should sum to 120min.

Define

S = Number of minutes detected as stressed in 120 minutes following an available decision

point

NS = Number of minutes not able to detect as stressed in 120 minutes following an available

decision point

A = Number of minutes detected as physically active in 120 minutes following an available

decision point

The proximal outcome variable takes on three values. Specifically,

Y = {“detected as stressed” if max(S, NS, A) = S,

“not able to detect as stressed” if max(S, NS, A) = NS,

“physically active” if max(S, NS, A) = A}.

**3.5.1 Missingness at the episode level**

For every minute of the 120 minute window following an available decision time, we know whether this minute is part of one of the following four types of episodes:

1. detected stressed
2. not-able-to-detect-as-stressed
3. unknown due primarily to physical activity \*\*This is not missing data\*\*

* *~~Rule 1~~*~~: mark these minutes as non-classifiable (NC) minutes if we do not consider a triad outcome. Best causal inference would be if we consider a triad outcome~~.

1. missing due primarily to bad data quality

* *Rule 2*: If prior episode is detected-stressed and next episode is detected-stressed, impute minutes as detected-stressed minutes.
* *Rule 3*: If prior episode is detected-stressed and next episode is not-able-to-detect-as-stressed, impute minutes as not-able-to-detect-as-stressed minutes.
* *Rule 4*:If prior episode is not-able-to-detect-as-stressed and next episode is detected-stressed, impute minutes as not-able-to-detect-as-stressed minutes.
* *Rule 5*:If prior episode is not-able-to-detect-as-stressed and next episode is not-able-to-detect-as-stressed, impute minutes as not-able-to-detect-as-stressed minutes.
* *Rule 6*: If prior minutes are not classifiable (i.e., due to physical activity) and next minutes are within an episode detected as stressed, mark minutes as not classifiable.
* *Rule 7*: If prior minutes are not classifiable (i.e., due to physical activity) and next minutes are within an episode not-able-to-detect-as-stressed, mark minutes as not-able-to-detect-as-stressed as we know physical activity to be stress reducing.
* *…*

**3.5.2 Missingness at the minute level**

For every minute of the 120-minute window following an available decision time, we know whether this minute is part of one of the following four types of labels (i.e. a classifiable minute):

1. detected stressed
2. not-able-to-detect-as-stressed
3. unsure (as coded within the data) due to physical activity \*\*This is not missing data\*\*

Table 5 shows the classifiable minutes out of the 120 minutes following an available decision time using both (i) stress episode data; and (ii) cStress minute level labeled data. Section S.1 provides detail on the definition of a classifiable minute.

At this time, there is no interpolation being conducted for the minute level labeled data, since at first we are interested in understanding exactly how much missing data is present. Due to the large amount of missingness resulting from using the minute level labeled data, we opt to use the episode data in order to construct the data for the primary outcome.

**3.5.3 Imputation Strategy**

As evident by Table 5, the average fraction of classifiable minutes in the 120-minute window following available decision times using data streams for both the episode level and minute level detected-stress classifications are *very* low.

**Data issue currently under investigation:** E.g., for individual 202 on day 3, the episode level data stream provides an average fraction of 0.43 for classifiable minutes in the 120 minute window following an available decision time. The corresponding value using the data stream providing minute level cStress labels is 0.

The detected-stress minute level classification time series suffer from much missing data and a variety of noise sources, including sensor detection error, are present in the data. Therefore, pre-processing steps like data imputation are important to achieve high-quality data for reliable inferences.

Perhaps the missing data is missing at random (MAR). This would be the case if stress can be explained by other known contextual variables such as day of the week, time of the day, previous stress levels, and the slope and intercept of previous time-series samples.

**Proposal**: We aim to conduct multiple imputation of our longitudinal categorial data (i.e., detected-stressed minute, not-able-to-detect-as-stressed minute, physical activity minute) through Bayesian mixture latent Markov models (Vidotto et.al., 2019).

**3.5.4 Predicting missing episodes**

As a first pass, we build a logistic mixed effects model (with a random effect on participant ID) in order to identify features of current/past history that may be useful in predicting missing minutes.

We create the following variables for consideration:

Y = episode type {1 = missing episode, 0 = classified episode}

X1 = Day in MRT

X2 = episode length in minutes (standardized using z-transformation)

X3 = previous episode type {1 = missing episode, 0 = classified episode}

X4 = previous episode length in minutes (standardized using z-transformation)

X5 = participant ID

In R, we use glmer() to fit a simple logistic mixed effects model. We also use ggeffects::ggpredict()which returns marginal effects on the response scale and thus the predicted values are predicted probabilities (these are shown in the following four plots). For this mixed model, these predicted values are at the population level.









Given that we identify clear trends of the predictors in terms of predicting the response, these variables could be considered within an imputation model for filling in missing episodes. We plan to conduce a more thorough analysis in coming weeks.

***3.6 Effects of Intervention Reminder***

<TODO>

* 1. ***Sensitivity Analyses***

<TODO>

## **Discussion**

## **References**

Boruvka, A., Almirall, D., Witkiewitz, K. and Murphy, S.A., 2018. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, *113*(523), pp.1112-1121.

Dempsey, W., Liao, P., Kumar, S. and Murphy, S.A., 2017. The stratified micro-randomized trial design: sample size considerations for testing nested causal effects of time-varying treatments. *arXiv preprint arXiv:1711.03587*.

Dürichen, R., Pimentel, M.A., Clifton, L., Schweikard, A. and Clifton, D.A., 2014. Multitask Gaussian processes for multivariate physiological time-series analysis. *IEEE Transactions on Biomedical Engineering*, *62*(1), pp.314-322.

Team, R.C., 2013. R: A language and environment for statistical computing.

Vidotto, D., Vermunt, J.K. and Van Deun, K., 2019. Multiple imputation of longitudinal categorical data through bayesian mixture latent Markov models. *Journal of Applied Statistics*, pp.1-19.

# **Supplementary Material**

**S.1 Definitions for Supplementary Material**

Tables 3 and 4 in Section S.2 below consists of the following additional information:

* **Highlighted yellow cells**, which correspond to days in which there was no available data quality measurements (i.e., the data quality stream was empty for a participant on a given day).
* **Highlighted red cells**, which correspond to days in which there was <2% good quality data available that day. This percentage is the number of 2-second windows that correspond to good quality data out of the 21600 2-second windows throughout a participant’s 12 hour day (i.e., percentage of data that is not missing due to sensor detachment or sensor off the body, low phone or sensor battery, momentary wireless data loss or software crash).
* **A red box** indicates that decision points on the given day have begun to include the detection of a participant’s post-lapse phase of the trial. This is recorded at an available decision time (i.e., at the peak of a detected-stressed or not-able-to-detect-as-stressed episode given the participant is available) based on previous records of self-reported smoking from the individual. The lapse could have occurred after the available decision times on the previous day, or at some time during the spread of the available decision times on the current day.
* Participant 213 withdrew on day 6 and only generated decision times on days 1 and 2.
* A **classifiable minute** (i.e., one that is not regarding missing) is one which belongs to either:
  + *USING EPISODE DATA*:
    - a detected-stressed episode
    - a not-able-to-detect-as-stress episode
    - an unknown episode that is due to physical activity (*due to the current state of the data, at the moment we use a proxy for this value, that is, a minute that is detected as a physically active minute that is also inside an unknown episode. This will change once we have access to the offline data from the Memphis team.*)
  + *USING CSTRESS MINUTE LEVEL LABELED DATA*:
    - a detected-stressed minute
    - a not-able-to-detect-as-stress minute
    - an “unsure” (as coded in data files) minute that is due to physical activity.

**S.2 Distribution of Decision Times**

Table 3: Total Generated Decision Times (left); Total Available Decision Times (right). Refer to Section S.1 above for detail on how to interpret a red box, a highlighted red box and a highlighted yellow box.



Table 4: Total Generated Decision Times (left); Average fraction of classifiable minutes out of the 120 minutes following an available decision time (right). Refer to Section S.1 above for detail on how to interpret a red box, a highlighted red box, a highlighted yellow box and the definition of a classifiable minute.



Table 5: Average fraction of classifiable minutes out of the 120 minutes following an available decision time: (i) using stress episode data (left); and (ii) using cStress minute level label data (right). Refer to Section S.1 above for detail on the definition of a classifiable minute.



Table 6: Average Number of Interventions Sent on days where there was more than 1 available detected-as-stressed decision time.



Table 7: Average Number of Interventions Sent on days where there was more than 1 available not-able-to-detect-as-stressed decision time.



Table 8: Left: minutes detected to be not physically active within the 2 hour period following an available decision time. Right: minutes detected to be physically active within the 2 hour period following an available decision time.

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

1. Enough good quality data corresponds to not having more than 33% of the minute’s corresponding data missing due to sensor detachment or sensor off the body, low phone or sensor battery, momentary wireless data loss or software crash. [↑](#footnote-ref-1)
2. Physical activity was determined at each generated decision point by using activity recognition algorithms that automatically analyze data from the AutoSense-based accelerometer to classify participants’ current activity. [↑](#footnote-ref-2)