Randomization Probabilities for the Smoking Study

Peng Liao, Walter Dempsey and Susan Murphy

University of Michigan

1 Setup and Notation

1.1 Defining the Stress and Not Stress Episodes

At every minute, the stress intensity is obtained and Moving Average Convergence Divergence (MACD) is used to estimate the trend based on short-term and long-term Exponential Moving Average of the past stress intensity (imputed if missing). The output of MACD is whether current stress intensity is increasing (+) or decreasing (−); see Hillol et al. (2016) for details. When the signs change from − to +, we mark the second time point (e.g. the one with +) as the start or the end of an episode. When the signs change from + to −, we mark the first time point (e.g. the one with +) as the "peak" (not necessarily the maximal stress intensity) of the current episode. Suppose an episode starts at time a and attains the peak at time b; see Figure 1. If the proportion of missingness from time a to $b+1$ is greater than 50%, then the episode is classified as "Unknown"; otherwise, we calculate the average of stress intensity from time a to $b+1$ and compare it with the threshold: if above the threshold, then classified as "Stress" episode; otherwise, "Not stress". For either "Stress" or "Not Stress" episode, the EMI is randomized at time $b+1$.

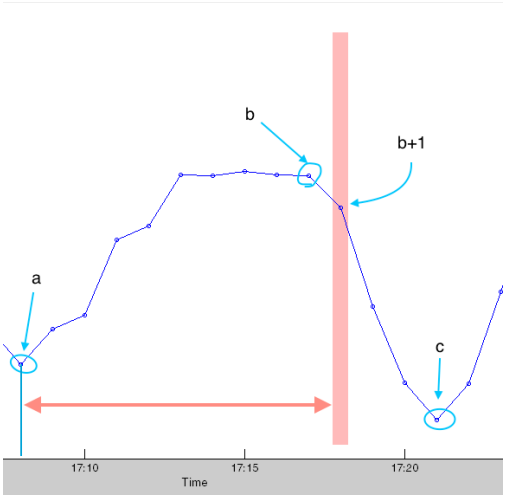

Figure 1: Example of an episode. a is the starting time, b is the peak and c is the end time of the episode. $b+1$ is the time in which classification is performed.

1.2 Setup

Consider the planned smoking cessation trial. Let s index the number of episode. X_s is the classification of s-th episode: $X_s = 1$ means "Not Stress and all conditions are satisfied" , $X_s = 2$ means "Stress and all conditions are satisfied" and $X_s = 3$ means "Unknown or some of conditions are not satisfied". The "conditions" includes: timing constraints for EMA/EMI; data quality in last 5 min is good; not driving; phone battery level $> 10\%$; not currently being physically active, e.g. walking or moving. Δ_s is the duration of s-th episode, e.g. $= c - a$ in Figure 1. A_s is binary variable indicating whether EMI is sent for s-th episode, e.g. $1 =$ "Send EMI" and $0 =$ "Provide Nothing" and T_s is the time at which A_s is randomized; e.g. after the peak of the episode (red line in Figure 1) . Let H be the total hours of a day (e.g. $H = 12$) and $S_H = \max\{s : T_s \leq H\}$ is the total number of episodes with the peak time that would occur in c hours.

In the pre-lapse phase, the average number of Stress and Not Stress episodes per day are 5.84 (3.88) and 29.50 (13.95) in the Minnesota data; see below for details. [1]. In the post-lapse, these are 3.81 (3.65) and 27.59 (11.70). These numbers are calculated for the subset in which the duration of a day is greater than 12 hours and averaged across person-day. It turns out that if considering the constraints imposed between the the times of random EMA, event contingent EMA and post EMIs, it is not possible to guarantee, on average, 1.5 EMIs assigned for Stress episodes in the post lapse. As such, in the post-lapse, we aim to provide on average 1 EMI per day at Stress and 1.5 EMIs at Not Stress; that is

$$\mathbb{E}\left[\sum_{t=1}^{S_H} A_t \mathbb{1}_{\{X_t=1\}}\right] \approx 1.5, \text{ and } \mathbb{E}\left[\sum_{t=1}^{S_H} A_t \mathbb{1}_{\{X_t=2\}}\right] \approx 1.$$

In the pre-lapse, we aim to provide 1.5 EMIs for both Stress and Not Stress episode:

$$\mathbb{E}\left[\sum_{t=1}^{S_H} A_t \mathbb{1}_{\{X_t=i\}}\right] \approx 1.5, \ i = 1, 2.$$

2 Randomization Probability for EMI

2.1 Algorithm

The inputs of the randomization algorithm are the tunning parameters $\mathbf{N} \in \mathbb{R}^2$, $\lambda \in (0, 1)$, together with a function $g(i, r)$ with $i = 1, 2, r > 0$, the latter of which is used to predict number of available Stress (or Not Stress) episodes that would occur in the remaining time r (available means the timing constraints and other conditions are satisfied). At the peak of s-th episode, suppose $T_s \leq c$. The randomization probability to assign EMI is given by

$$p(X_s) = \frac{\mathbf{N}(X_s) - \sum_{t=1}^{s-1} \left[\lambda_t A_t + (1 - \lambda_t) p(X_t)\right] \mathbb{1}_{\{X_t = X_s\}}}{1 + g(X_s, c - T_s)}, \text{ for } X_s = 1, 2 \tag{1}$$

where $\lambda_t = \lambda^{T_s - T_t}$. Note that when $s = 1$, the sum in the numerator is defined as 0, e.g. $p(X_1) = N(X_1)/(1 + g(X_1, c - T_1))$. In addition, we will restrict the randomization

[1] Number in the parenthesis is standard deviation.

probability within the interval $[\epsilon, 1-\epsilon]$ for "Stress" and $[0,1]$ for "Not Stress". [2] In the following, $\epsilon = 0.05$. explain above formula and provide the intuitive definitions of $N(j), j = 1,2,3$ as well as g and λ_t in this section–mainly need to move the text around.

2.2 Using Minnesota dataset

We will use Minnesota dataset to approximate a transition probability matrix for the X_s's (classification of the episode), the distribution of the Δ_s's (length of the episodes), and the peak time for each Stress and Not Stress episode for the planned smoking study. need to clarify which part of minnesota data set is used here so no confusion results when you use a subset of the minnesota data to evaluate the method; you need a cite to a paper describing the minnesota data set This model does not incorporate the constraints on the timings of the EMAs, EMIs; the model does not incorporate the effects of the EMIs on either the transition probability matrix, peak time in which classification is detected, or the length of the episodes. To distinguish the classifications in the Minnesota trial from the classifications in our planned smoking study, we use $\tilde{X}_s$ to denote the classification of $s-$th episode in the Minnesota trial. Thus $\tilde{X}_s = 1$ means "Not Stress", $\tilde{X}_s = 2$ means "Stress", and $\tilde{X}_s = 1$ means "Unknown". We assume a Markov transition model for $\tilde{X}_s$ and assume a parametric model for Δ_s given $\tilde{X}_s$: Gamma distribution when $\tilde{X}_s = 1, 2$ and Log Normal distribution when $\tilde{X}_s = 3$. Denote r_s by the ratio of the peak time to the length of episode, e.g. $(b+1-a)/(c-a)$ in Figure 1. We discretize $r_s \in [0,1]$ by $\{[0.05(k-1), 0.05k), k = 1, \ldots, 20\}$ and fit a Multinomial distribution. All of above estimation will be done separately for pre-lapse and post-lapse; see appendix for details of the estimated models.

The estimated transition model based on the Minnesota data will serve two purposes. First, we will use it to calculate the average number of Stress/Not Stress episodes that will occur in a given period of time. This will serve as a basis to obtain the g function. Secondly, we will use the estimated model to generate simulation data so as to evaluate the performance of the algorithm.

2.3 On the selection of N, λ and g

The selection of tuning parameters $\mathbf{N}$ and λ, together with g function is based on three criteria: 1) the average number of assigned EMIs at Stress and Not Stress episodes approximately satisfy our requirements (e.g. 1.5 for both Stress and Not Stress in the pre-lapse and 1 and 1.5 for Stress and Not Stress in the post-lapse); 2) the total number of assigned EMIs in a day has a low probability of receiving no EMI and very many EMI's (ideally, on most days the user should receive 1 or 2 EMIs at Stress and similarly for at Not Stress episodes in a day); and 3) the average numbers of EMIs assigned at each one-hour block in a day are approximately equal. Just as the estimation of transition model, the tuning parameters and g function will be selected separately for pre-lapse and post-lapse. In the following, we discuss how we select $\mathbf{N}$, λ and function g to meet these three criteria.

It can be seen from (1) that $\mathbf{N}$ roughly controls the number of EMIs that could be sent in a day. Thus $\mathbf{N}$ is selected to meet the first criterion. We propose to use a small value

[2] To clarify, $p(X_s)$ is equal to the truncation of the RHS in (1). Note each $p(X_t)$ in the numerator on the RHS of the display has already been truncated.

for λ to meet the second criterion. The reason why λ should not be set to 1 can be readily seen from (1): when $\lambda = 1$, the randomization probability at a not stressed episode $X_s = 1$, becomes

$$p(1) = \frac{\mathbf{N}(1) - \sum_{t=1}^{s-1} A_t \mathbb{1}_{\{X_t=1\}}}{1 + g(1, c - T_s)}.$$

Thus whenever we have already provided $\mathbf{N}(1)$ EMIs at Not Stressed episode in the past, we stop sending any more EMIs in the rest of the Not Stressed episodes. Even though this can guarantee we will have exactly $\mathbf{N}$ EMIs in a day, this will cause some trouble in the data analysis as there will be episodes with randomization probability equal to 0. On the other hand, the reason why we include λ, e.g. instead of $\lambda = 0$, is that we found that the variance in number of EMIs per day is high (but the average number of EMIs is still close to what we want). This is true because when $\lambda = 0$, we do not take into account how many EMIs has been already sent in the past, but only the probability of sending the EMI, since the probability is determined by

$$p(1) = \frac{\mathbf{N}(1) - \sum_{t=1}^{s-1} p(X_t) \mathbb{1}_{\{X_t=1\}}}{1 + g(1, c - T_s)}.$$

Hence, using $\lambda \in (0, 1)$ allows us to smoothly transferring between these two extreme unwanted cases so as to meet the above second criterion.

Now we discuss how to obtain a desired g function to meet the third criterion. One intuitive and simple choice of the g function is the conditional expectation of average number of Stress (Not Stress) episodes that will occur in the remaining time. However there are at least two problems with this choice. The first one is that this conditional expectation doesn't take into account availability, that is, the constraints in our planned study on the time between EMIs and between EMIs and EMAs (especially the constraint that two EMIs need to be separated by 60 mins). We found that using the conditional expectation to predict the number of available stress episodes in the future is likely too large, resulting in too low a probability for providing EMIs at Stress episodes. The other problem is due to the right? skewness of the distribution of remaining number Stress (Not Stress) episodes; this means that the conditional expectation is not a good representation of the distribution of the remaining number of Stress episodes. To overcome these issues, we incorporate the standard deviation of the remaining number of Stress (Not Stress) episodes into the g function:

$$g(i, r) = \mathbb{E}\left[\sum_{j=2}^{\max\{k:T_k \le r - \delta(\tilde{X}_1)\}} \mathbb{1}_{\{\tilde{X}_j = i\}} \,\middle|\, \tilde{X}_1 = i \right] - \boldsymbol{\eta}(i) \times \mathbf{sd}\left[\sum_{j=2}^{\max\{k:T_k \le r - \delta(\tilde{X}_1)\}} \mathbb{1}_{\{\tilde{X}_j = i\}} \,\middle|\, \tilde{X}_1 = i \right],$$

where $\delta(\tilde{X}_1)$ is the time gap between the peak time to the end of first episode, $\mathbf{sd}$ is the standard deviation and $\boldsymbol{\eta}(i) > 0, i = 1, 2$ are the tuning parameters. The above expectation and the standard deviation will be calculated by Monte Carlo method using the transition probability matrix and the distribution of episodes (e.g. Δ_t) constructed using the Minnesota data. The tuning parameters $\boldsymbol{\eta}$ is selected so as to satisfy the third criterion.

3 Simulation Study

In the following simulation, random EMAs are assigned according to the proposal for our smoking cessation study. In the post-lapse phase the event contingent EMA are also assigned according to the proposal. For simplicity we simulate smoking events in each four-hour window block with a uniform distribution in the post-lapse phase. We also consider the timing constraints on the EMAs/EMIs. The transition probability matrix as well as the distribution of the episodes used to simulate data was constructed using the Minnesota data set. We evaluate the algorithm on the simulated data. We also evaluate the algorithm directly by sampling from Minnesota dataset.

3.1 Pre-lapse Phase

In the pre-lapse, the tuning parameters are optimized to approximately meet the criteria discussed in section 2.3. The tuning parameters are chosen as $\mathbf{N} = (1.6, 2.25)$, $\lambda = 0.4$, $\boldsymbol{\eta} = (1.1, 0.5)$. The performance of this choice of tuning parameters on the simulated dataset and Minnesota dataset will be discussed in the next two sections. give some words about how you optimized–be upfront that this was informal in that there is no criterion function

3.1.1 Simulated Dataset

Recall that the simulated data is based on the transition probability matrix as well as the distribution of the episodes (peak time and length) constructed from the Minnesota data set. We consider two cases for the starting episodes: one is Unknown, and the other one is randomly sampled according to the estimated initial distribution in the pre-lapse (0.426, 0.213 and 0.362 for Not Stress, Stress and Unknown). To evaluate the performance of the algorithm, we generate 10,000 person-day's of data and for each person-day, we assign random EMA as discussed in the beginning of section 3, and assign EMIs sequentially based on (1). There is no event contingent EMA involved since it is for pre-lapse. The simulation result is given in Figure 2.

When the starting episode is randomly selected according to the estimated starting distribution from the ? subset of the Minnesota data (whole), the probability of receiving EMI at stress episode in the first hour is much higher than the rest of the day. This is due to the impact of the initial distribution; see Figure 3 for the distribution of Stress Episodes across hour block for the simulated data. When starting episode is Unknown, the probability of receiving EMI throughout the day are relatively flat.

3.1.2 Minnesota Dataset

As a further evaluation of the updating algorithm, we apply the algorithm to a subset of person-day's from Minnesota dataset. need to clarify how this subset is different/same as the data used to build the generative model and used to provde average numbers of stress and not-stressed episodes per day (This is NOT the same dataset to build generative model; But this is same for calculating the number of stress/not stress per day in section 1.2.) This subset is the subset of person-day's for which the duration of a day is longer than 12 hours (and we truncate at 12 hours) and the participant is in pre-lapse phase. The total

number of person-day's in the Minnesota data set meeting these requirements is 32. We sample 10,000 times (with replacement) from this subset and for each sampled person-day, we assign random EMA and assign EMIs sequentially based on (1) during that day.

The simulation result is given in Figure 4. The time trend of the average number of EMIs in each hour block might be due to the fact that the Stress/Not Stress episodes are not quite uniform across each hour-block in the (selected) subset; see Figure 5 for the distribution of Stress/Not Stress Episodes across hour block for the selected subset.

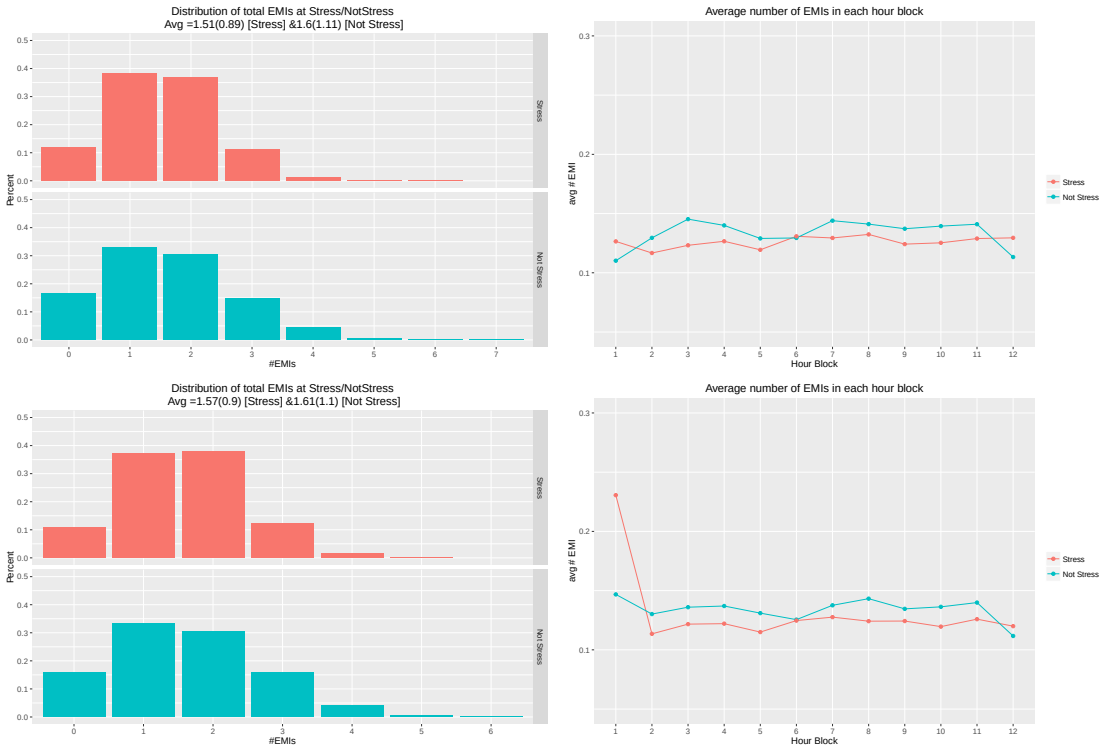

Figure 2: On the Simulated Dataset. Pre-lapse. Starting with Unknown (top) and estimated initial distribution (down). The left two figures are the empirical distribution of total EMIs at Stress (red) and Not Stress (blue) across the replications. The right figure is the average number of assigned EMIs in each hour block across the replications.

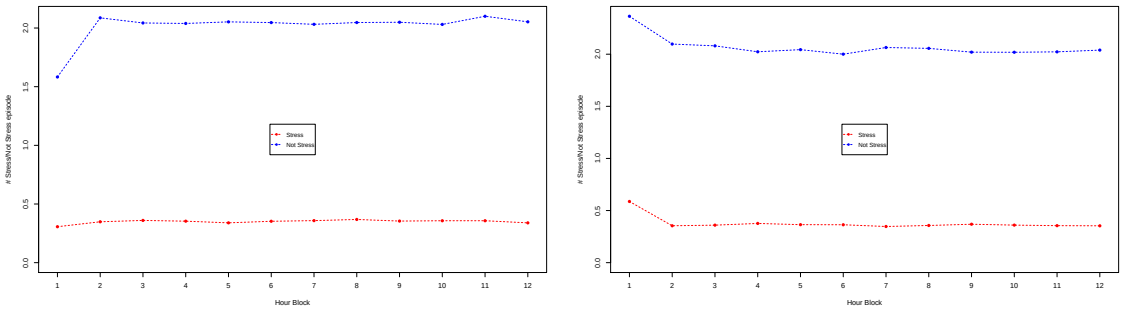

Figure 3: Distribution of Stress/Not Stress Episode across hour block in the simulated dataset. Starting with Unknown (left) and estimated initial distribution (right).

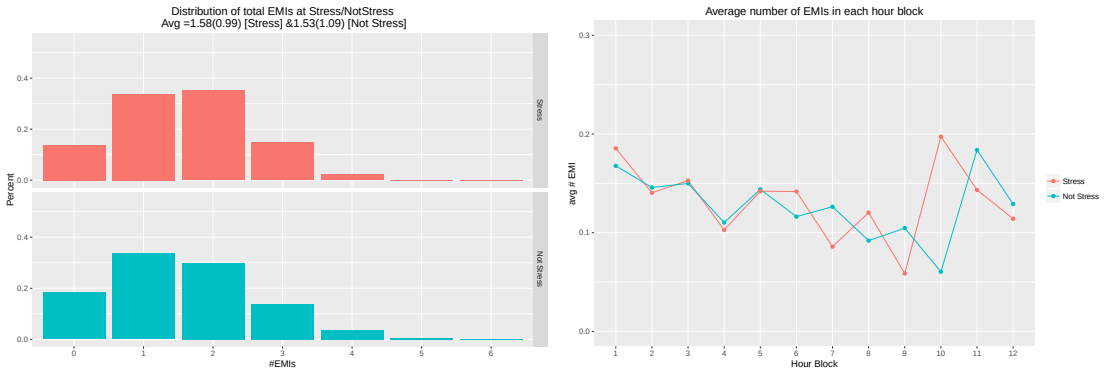

Figure 4: Using resampled person-day's from the 35 pre-lapse person-day's in the Minnesota Dataset. The left figure is the empirical distribution of total EMIs at Stress (red) and Not Stress (blue) across the resamplings. The right figure is the average number of assigned EMIs in each hour block across the resamplings.

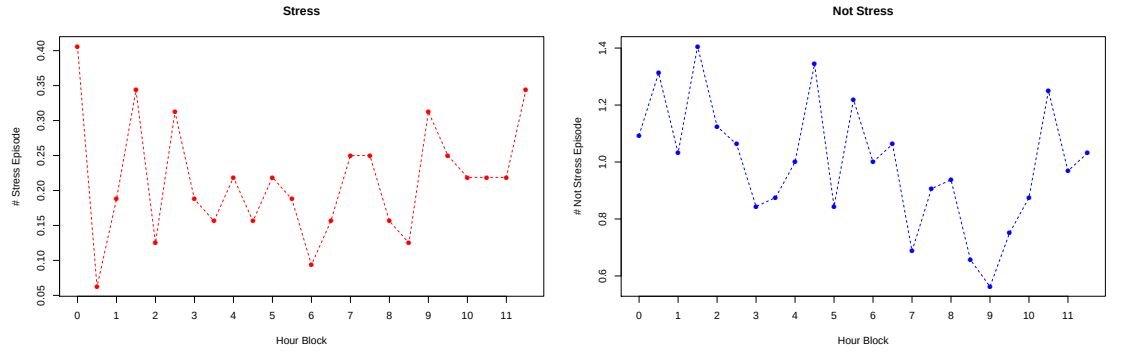

Figure 5: Distribution of Stress/Not Stress Episode across hour block in the selected subset of Minnesota dataset. See section 3.1.2 for how the subsets are selected.

3.2 Post-lapse Phase

In the post-lapse, the tuning parameters are optimized to approximately meet the criterion discussed in section 2.3. The tuning parameters are chosen as $\mathbf{N} = (1.65, 3)$, $\lambda = 0.4$, and $\boldsymbol{\eta} = (1.2, 0.5)$.

3.2.1 Simulated Dataset

As in the pre-lapse, the simulated data is generated according to the transition model estimated from the Minnesota dataset in the post-lapse phase and all of the starting episodes are "Unknown". We also consider two cases for the starting episodes: one is Unknown, and the other one is randomly sampled according to the estimated initial distribution in the post-lapse (0.491, 0.070 and 0.439 for Not Stress, Stress and Unknown) of the ? subsample of the Minnesota data (whole). We also need to simulate the smoking events. For simplicity, the numbers of smoking events in each four-hour window are assumed the same and are uniformly sampled in each time window. Recall that we aim to provide on average 1 EMI at Stress episodes and 1.5 EMIs at Not Stress episodes. We use 10,000 person-day's generated from the simulation model. The simulation results with differing numbers of smoking events are provided in Figure 6 and Figure 7. As in the pre-lapse, the initial distribution also has some impact on the probability of receiving EMI in the first hour block.

3.2.2 Minnesota Dataset

Same as in the pre-lapse, we select the subset of the person-day data in the post-lapse satisfying that the duration of a day is longer than 12 hours as test dataset. The total numbers of person-day data meeting these requirement is also 32. The results based on resampling 10,000 person-day's from this subset and with different numbers of smoking events is given in Figure 8. Again, the time trend of the average number of EMIs is related to the distribution of the Stress/Not Stress Episodes across hour block in the selected dataset; see Figure 10.

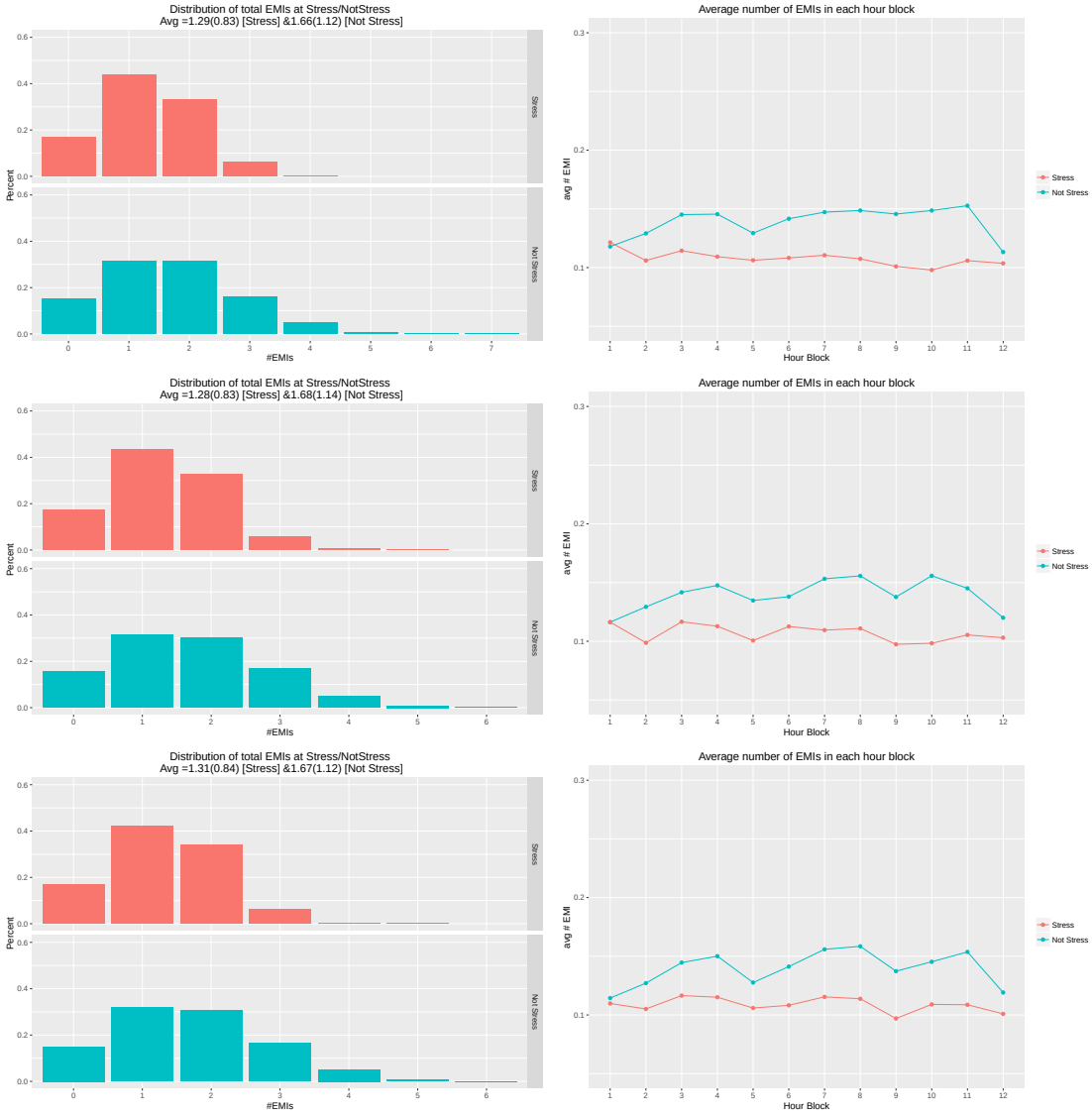

Figure 6: On the Simulated Dataset. Post-lapse. The initial episode is Unknown. Number of smoking events in each 4-hour window is 1 (top), 2 (Middle) and 3 (bottom). The left figures are the empirical distribution of total EMIs at Stress (red) and Not Stress (blue) across the replications. The right figure is the average number of assigned EMIs in each hour block across the replications.

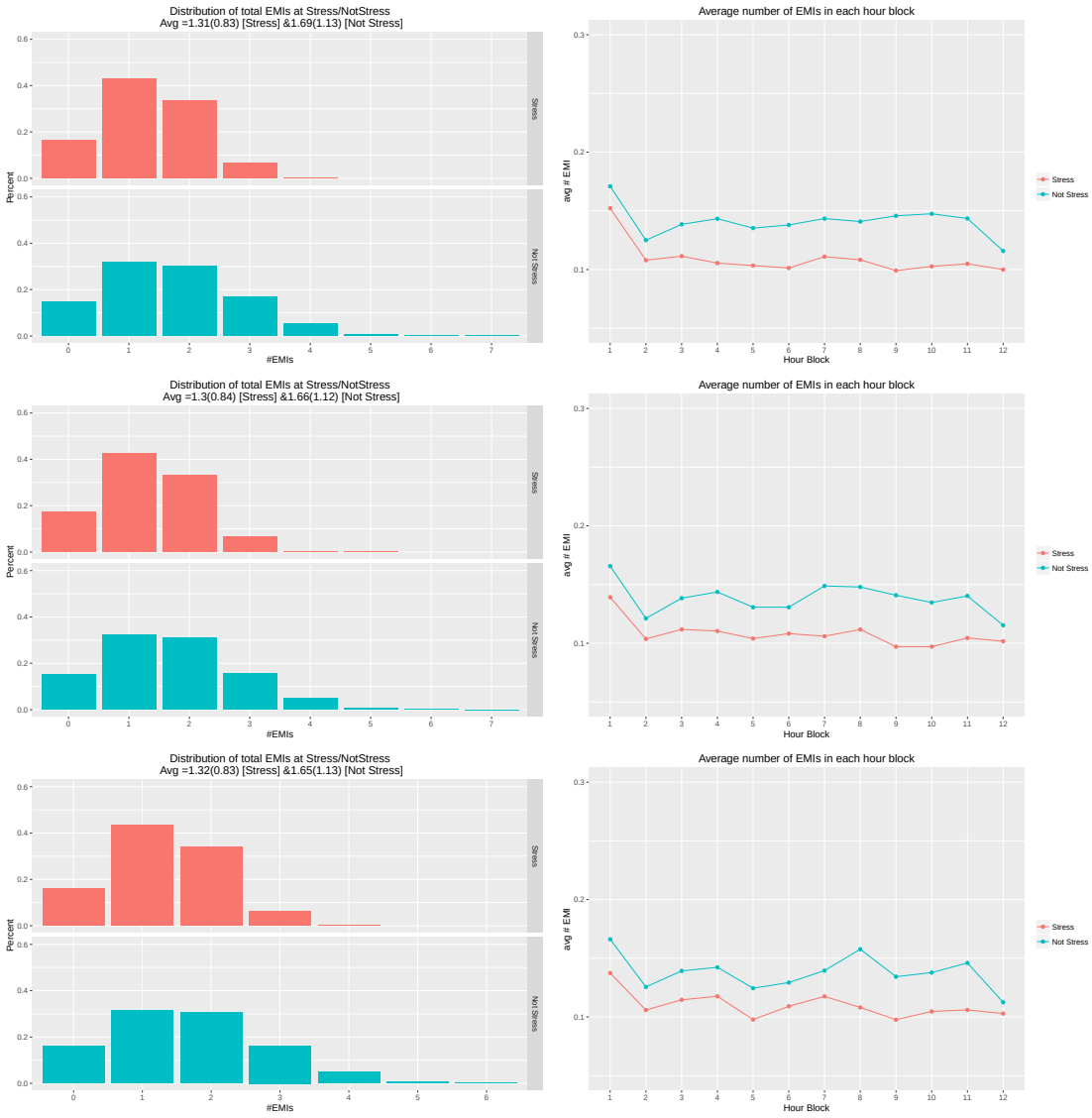

Figure 7: On the Simulated Dataset. Post-lapse. Starting with the estimated initial distribution. Number of smoking events in each 4-hour window is 1 (top), 2 (Middle) and 3 (bottom). The left figures are the empirical distribution of total EMIs at Stress (red) and Not Stress (blue) across the replications. The right figure is the average number of assigned EMIs in each hour block across the replications.

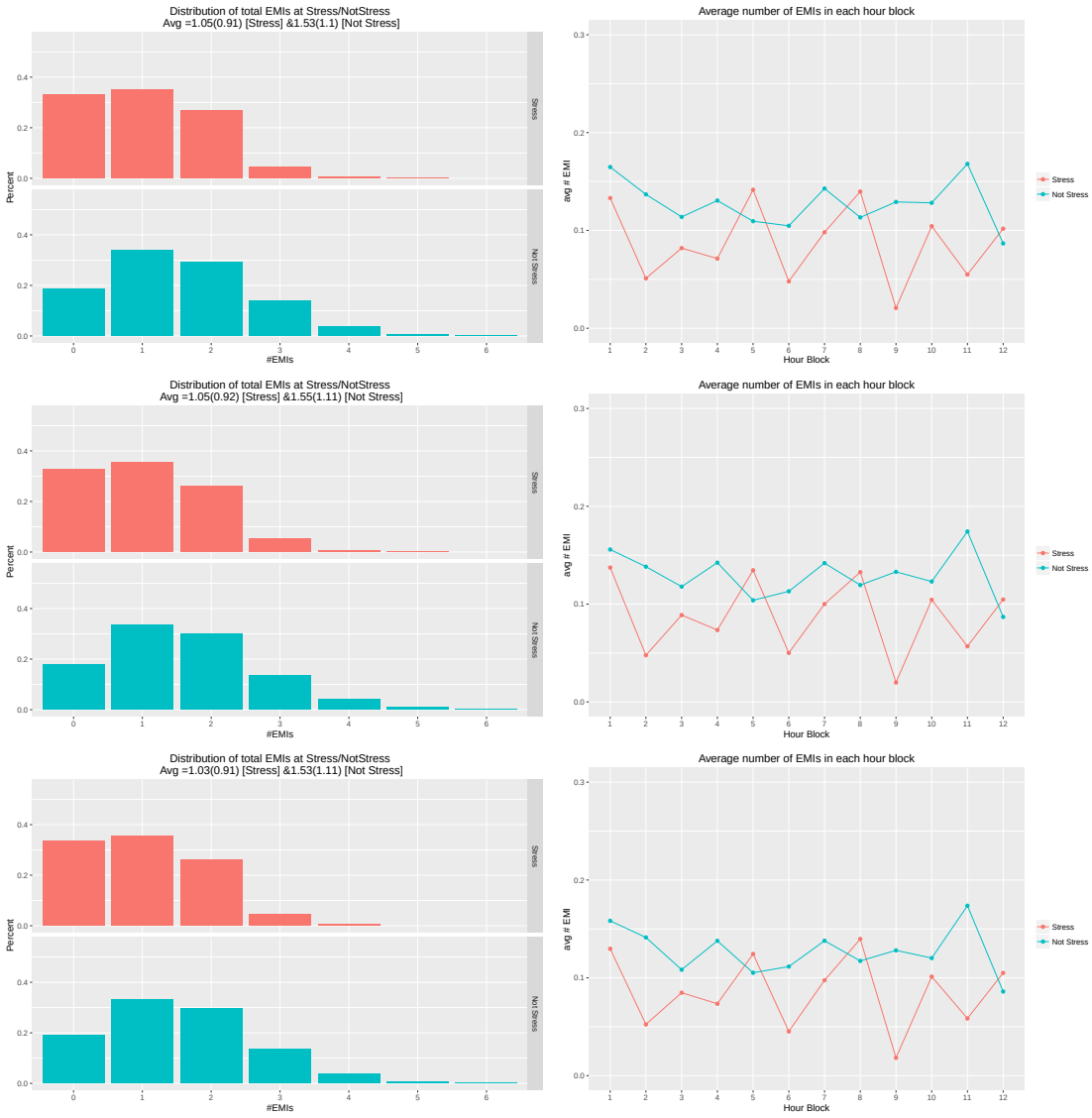

Figure 8: On the selected subset of Minnesota Dataset. Post-lapse. Number of smoking events in each 4-hour window is 1 (top), 2 (Middle) and 3 (bottom). The left figures are the empirical distribution of total EMIs at Stress (red) and Not Stress (blue) across the replications. The right figure is the average number of assigned EMIs in each hour block across the replications.

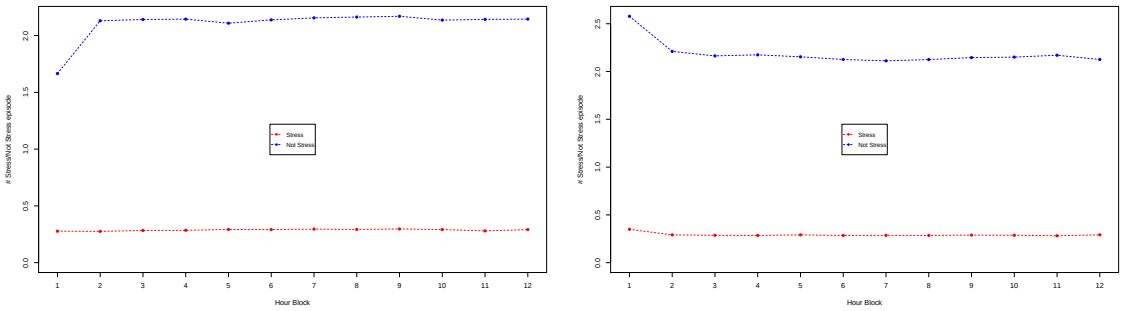

Figure 9: Distribution of Stress/Not Stress Episode across hour block in the simulated dataset. Post-lapse. Starting with Unknown (Left) and estimated initial distribution (right).

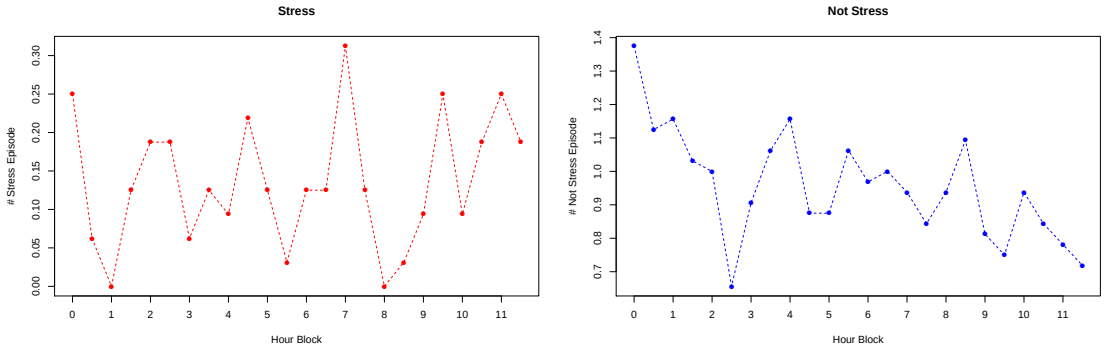

Figure 10: Distribution of Stress/Not Stress Episode across hour block in the selected subset of Minnesota dataset. See section 3.1.2 for how the subsets are selected.

Appendix

1. Estimated Markov transition matrices.

   Recall that $\tilde{X}_s = 1$ means "Not Stress", $\tilde{X}_s = 2$ means "Stress", and $\tilde{X}_s = 1$ means "Unknown". Below is the estimated (time-stationary) transition matrix $\left(\Pr(\tilde{X}_{s+1} = j|\tilde{X}_s = i), i, j = 1, 2, 3\right)$ for pre-lapse and post-lapses:

$$\mathbf{P}_{\text{pre}} = \begin{pmatrix} 0.667 & 0.053 & 0.280 \\ 0.289 & 0.325 & 0.387 \\ 0.421 & 0.097 & 0.482 \end{pmatrix} \qquad \mathbf{P}_{\text{post}} = \begin{pmatrix} 0.700 & 0.036 & 0.264 \\ 0.357 & 0.310 & 0.333 \\ 0.418 & 0.092 & 0.490 \end{pmatrix}$$

2. Estimated models for length of episodes

   In Table 1 we report the estimated models for the length of Stress, Not Stress and Unknown episodes. For Gamma distribution, θ_1 and θ_2 are shape and rate parameters. For Log Normal, θ_1 and θ_2 are mean and standard error parameters. Number in the parenthesis is the estimated standard errors.

Table 1: Estimated Model for length of episodes.

	Pre-lapse		Post-lapse	
	θ_1	θ_2	θ_1	θ_2
Gamma (Not Stress)	2.700 (0.088)	0.263 (0.009)	2.539 (0.107)	0.238 (0.011)
Gamma (Stress)	3.186 (0.243)	0.270 (0.022)	4.589 (0.543)	0.380 (0.048)
Log Normal (Unknown)	2.870 (0.023)	0.776 (0.016)	2.881 (0.032)	0.798 (0.023)

3. Emipircal Distribution ratio of the peak time of episodes

   Recall that the ratio of the peak time to the length of episode is defined as $(b + 1 - a)/(c - a)$; see Figure 1. The empirical distribution of the ratios with 20 equally separated breakpoints are provided in Figure 11.

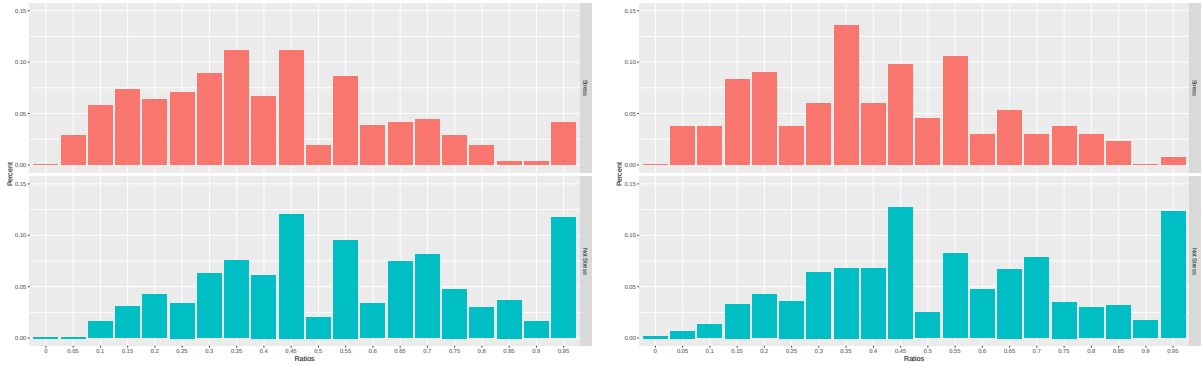

Figure 11: Estimated ratio of the peak time for Not Stress and Stress episodes in pre-lapse (left two) and post-lapse (right two).