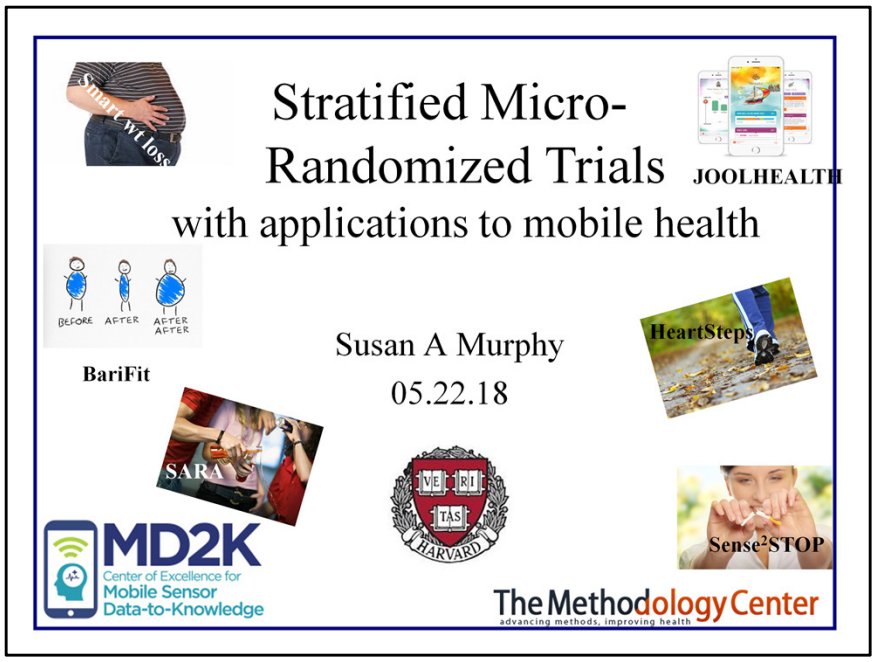


Stratified Micro-randomized Trials with Applications in Mobile Health

Abstract: Technological advancements in the field of mobile devices and wearable sensors make it possible to deliver treatments anytime and anywhere to users like you and me. Increasingly the delivery of these treatments is triggered by detections/predictions of vulnerability and receptivity. These observations are likely to have been impacted by prior treatments. Furthermore the treatments are often designed to have an impact on users over a span of time during which subsequent treatments may be provided. Here we discuss our work on the design of a mobile health smoking cessation study in which the above two challenges arose. This work involves the use of multiple online data analysis algorithms. Online algorithms are used in the detection, for example, of physiological stress. Other algorithms are used to forecast at each vulnerable time, the remaining number of vulnerable times in the day. These algorithms are then inputs into a randomization algorithm that ensures that each user is randomized to each treatment an appropriate number of times per day. We develop the stratified micro-randomized trial which involves not only the randomization algorithm but a precise statement of the meaning of the treatment effects and the primary scientific hypotheses along with primary analyses and sample size calculations. Considerations of causal inference and potential causal bias

incurred by inappropriate data analyses play a large role throughout.

Outline

- Introduction to Mobile Health
- Sense^{2}STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator

2

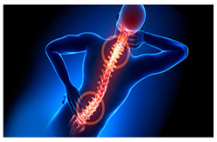
mHealth: Much Promise for Health Behavior Change!

Obesity/Weight Management

(e.gHsu et al, 2014)

Eating disorders

(e.g., Bauer et al. 2010)



Other chronic disorders

(e.g., Kristjánsdóttir et al., 2013)

Smoking cessation

(e.g., MD2K, 2017)

Alcohol use disorders

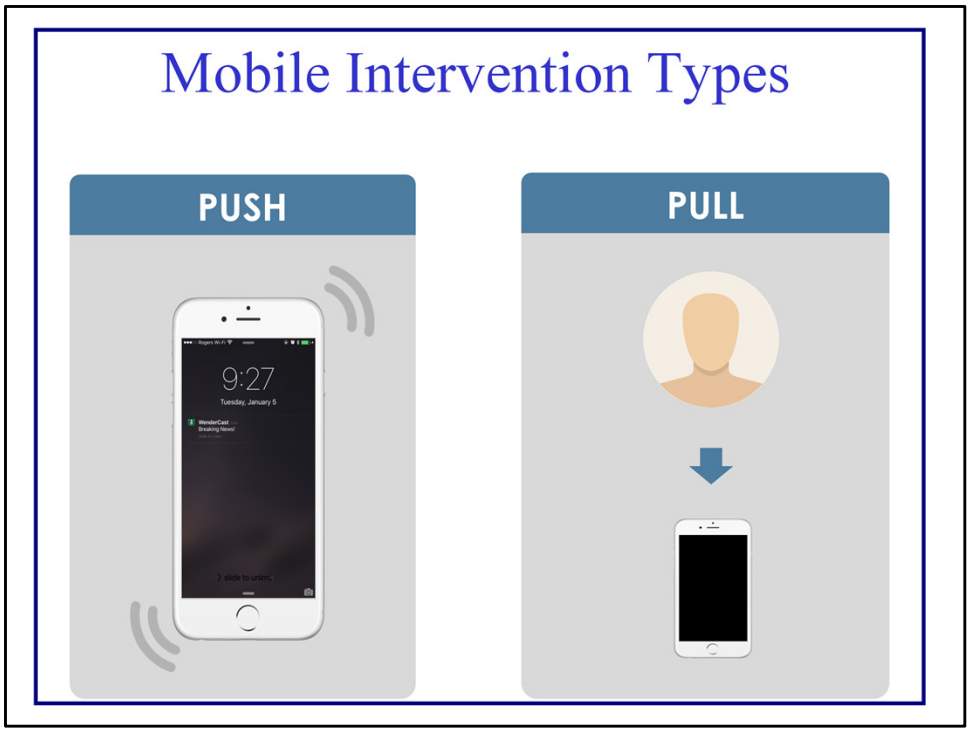
(e.g., Gustafson et al., 2014)

Physical activity

(e.g., Thomas & Bond, 2015)

Mental illnesses

(e.g., Ben-Zeev et al., 2013)

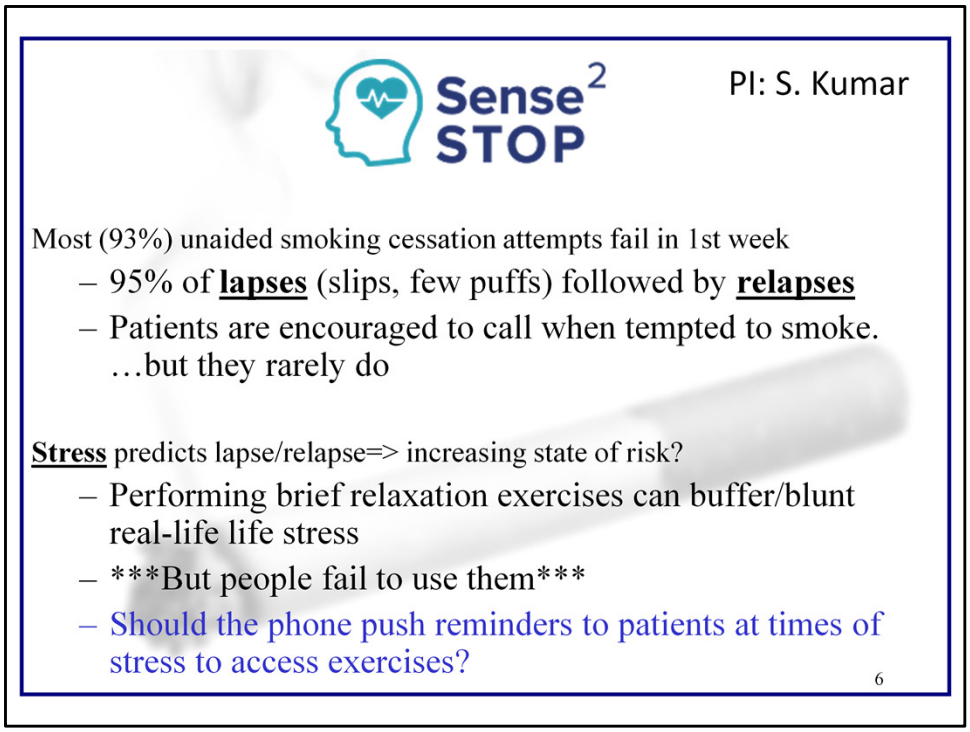


Pushes are designed to help person over a particular time interval

Outline

- Introduction to Mobile Health
- Sense^{2}STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator

5



This is a simplified version of Sense2Stop!

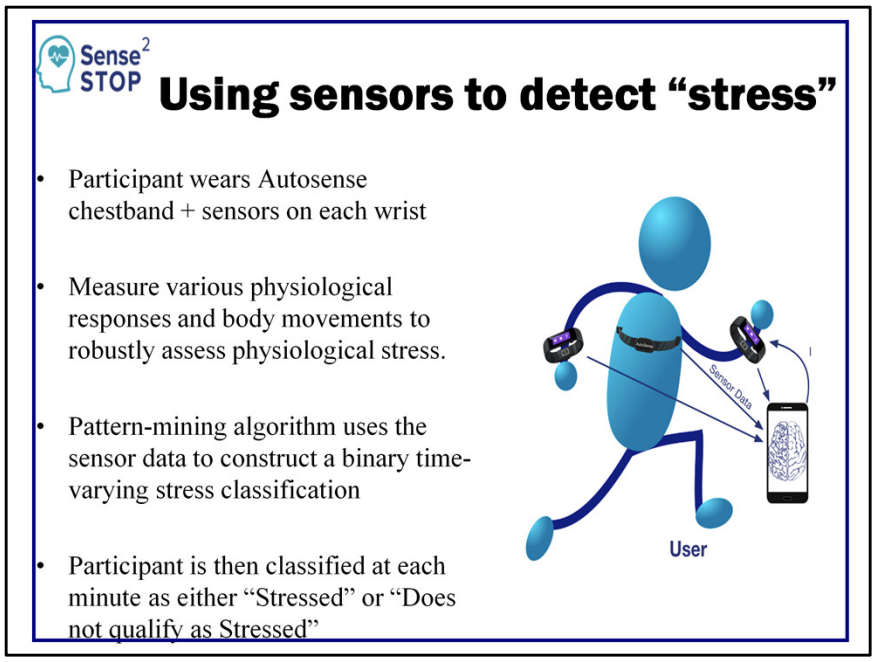
Stress State characterized by a combination of arousal and displeasure
(Kristensen, 1996; Posner et al., 2005)
Can be triggered by various circumstances in real-life

Shiffman et al. (2000) defined lapse as any occasion of smoking, even if only a puff. This was differentiated from relapse, which was defined as smoking at least five cigarettes for three consecutive days

The smoking cessation outcome can be 7-day point prevalence abstinence.

A “relapse” is defined as seven consecutive days of at least one puff per day following a period of total abstinence.
A “lapse” is defined as an isolated smoking episode of not more than six consecutive days followed by at least 24 h of abstinence.
This is from Ramelson, H. Z., Friedman, R. H., & Ockene, J. K. (1999). An automated telephone-based smoking cessation education and counseling system. *Patient*

Education and Counseling, *36*(2), 131-144.

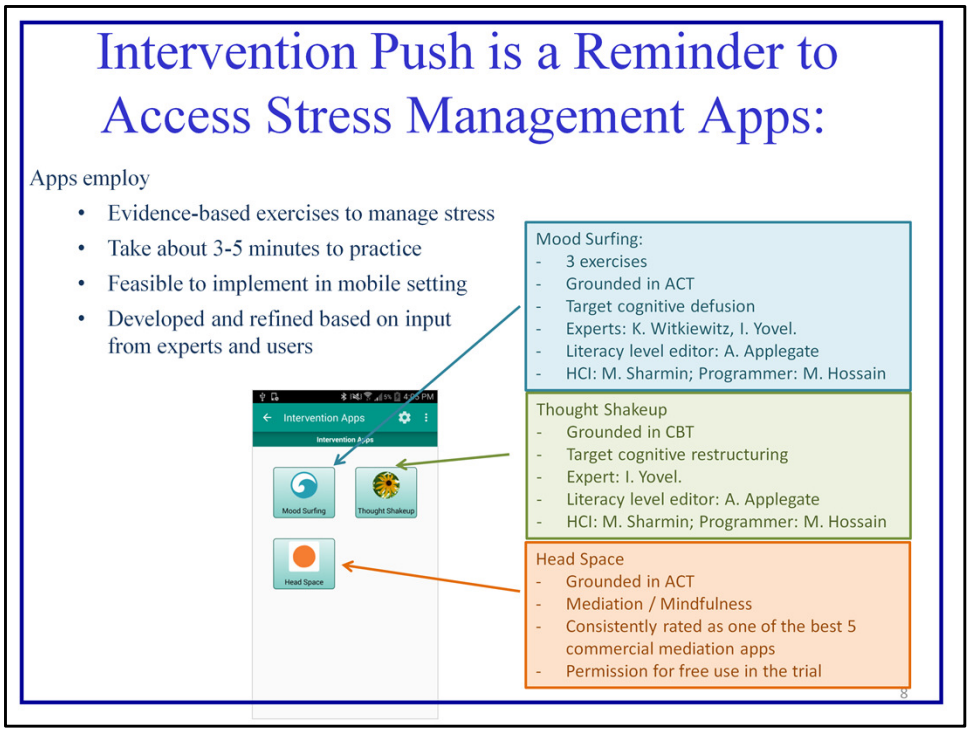


This is a prototype study as chest band is not for real life use.
Also participants are paid a good bit.

cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment
Karen Hovsepian_, Mustafa al'Absiy, Emre Ertin, Thomas Kamarck^
Motohiro Nakajimay, Santosh Kumar at UbiComp'15,September 7-11, 2015,

Ertin, E., Stohs, N., Kumar, S., Raij, A., al'Absi, M., and
Shah, S. Autosense: Unobtrusively wearable sensor
suite for inferring the onset, causality, and consequences
of stress in the field. In Proc. of ACM SenSys (2011),
274–287.

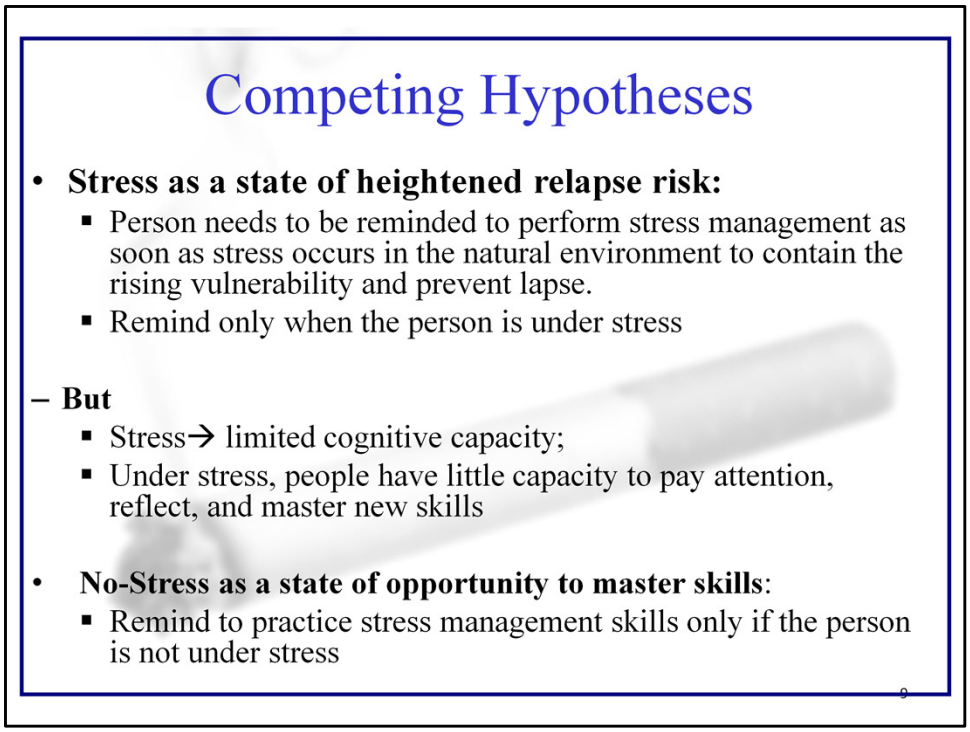
Sarker H, Hovsepian K, Chatterjee S, Nahum-Shani I, Murphy SA, Spring B, Ertin E, al'Absi M, Nakajima M and Kumar S (2017), *"From Markers to Interventions: The Case of Just-in-Time Stress Intervention"*, In Mobile Health: Sensors, Analytic Methods, and Applications. , pp. 411-433. Springer International Publishing.



Harris (2009) explains that cognitive defusion means:
Looking *at* thoughts rather than *from* thoughts
Noticing thoughts rather than becoming caught up in thoughts
Letting thoughts come and go rather than holding onto them
The general purpose of cognitive defusion is to:
Notice the *true nature* of thoughts – they are words or images in your mind
Respond to thoughts in terms of taking workable action – take action based on what "works" rather than what is "true"
Notice the actual *process of thinking* – recognize that thoughts do not dictate behaviors
Use cognitive defusion when thoughts are acting as a barrier to living in accordance with your true values

Cognitive restructuring (CR) is a psychotherapeutic process of learning to identify and dispute irrational or maladaptive thoughts known as cognitive distortions,[1] such as all-or-nothing thinking (splitting), magical thinking, over-generalization, magnification,[1] and emotional reasoning, which are commonly associated with many mental health disorders.[2] CR employs many strategies, such as Socratic questioning, thought recording, and guided imagery, and is used in many types of therapies, including cognitive behavioral therapy (CBT) and rational emotive therapy (RET). A number of studies demonstrate considerable efficacy in using CR-based

therapies.[3][4][5]



This is a simplified version of Sense2Stop!

State characterized by a combination of arousal and displeasure (Kristensen, 1996; Posner et al., 2005)
Can be triggered by various circumstances in real-life

Shiffman et al. (2000) defined lapse as any occasion of smoking, even if only a puff. This was differentiated from relapse, which was defined as smoking at least five cigarettes for three consecutive days

The smoking cessation outcome can be 7-day point prevalence abstinence.

It has been described that up to 80% of quitters eventually relapse within 6 months.
A “relapse” is defined as seven consecutive days of at least one puff per day following a period of total abstinence.
A “lapse” is defined as an isolated smoking episode of not more than six consecutive days followed by at least 24 h of abstinence.
This is from Ramelson, H. Z., Friedman, R. H., & Ockene, J. K. (1999). An automated

telephone-based smoking cessation education and counseling system. *Patient Education and Counseling*, *36*(2), 131-144.

Sense^2 STOP

- Primary: Should the smartphone notify the user with a reminder to utilize app directed stress-management exercises when the user is (not) stressed?
  - Does this effect vary with time or with current context?
- In the near term the reminder notification should reduce:
  - Near time, proximal, stress

10

Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day. Given these expectation, can we come up with a threshold for missing data from stress assessment?

Stratified Micro-Randomized Trial

- On each participant: $O_1, A_1, \ldots, O_t, A_t, \ldots$
- t : Decision time (times at which a treatment might be provided)
- O_t: observations after time $t-1$ and up to and including time t
  - $X_t \subset O_t$: time-varying stratification variable

Delta will be 120

Stratified Micro-Randomized Trial

- On each participant: $O_1, A_1, \ldots, O_t, A_t, \ldots$
- t : Decision time (times at which a treatment might be provided)
  - Sense^{2}STOP: each minute
- O_t: observations after time $t-1$ and up to and including time t
  - $X_t \subset O_t$: time-varying stratification variable
  - Sense^{2}STOP: $X_t = 1$ each minute if classified as stressed and =0, otherwise

Stratified Micro-Randomized Trial

- O_t: observations after time $t-1$ and up to and including time t
  - $X_t \subset O_t$: time-varying stratification variable
  - Sense^{2}STOP: $X_t = 1$ each minute if classified as stressed and =0, otherwise
  - $I_t \subset O_t$: availability indicator
  - Sense^{2}STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; =0 otherwise

Available if data is good, if t is at top of episode, if not driving, if no ema in prior 10 min, no treatment in prior hour.

Stratified Micro-Randomized Trial

- O_t: observations after time $t-1$ and up to and including time t
  - $I_t \subset O_t$: availability indicator
  - Sense^{2}STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; =0 otherwise
- A_t: Randomized Treatment at decision time t
  - Sense^{2}STOP: $A_t = 1$ if reminder is pushed to participant; =0 otherwise

Stratified Micro-Randomized Trial

- A_t: Randomized Treatment at decision time t
  - Sense^{2}STOP: $A_t = 1$ if reminder is pushed to participant; =0 otherwise
- $Y_{t,\Delta}$: Proximal response is a known function of participant's data within subsequent window of length Δ decision times
  - Sense^{2}STOP: fraction of time stressed in Δ=60 minutes

$$Y_{t,\Delta} = \Delta^{-1} \sum_{s=1}^{\Delta} 1[X_{t+s} = 1]$$

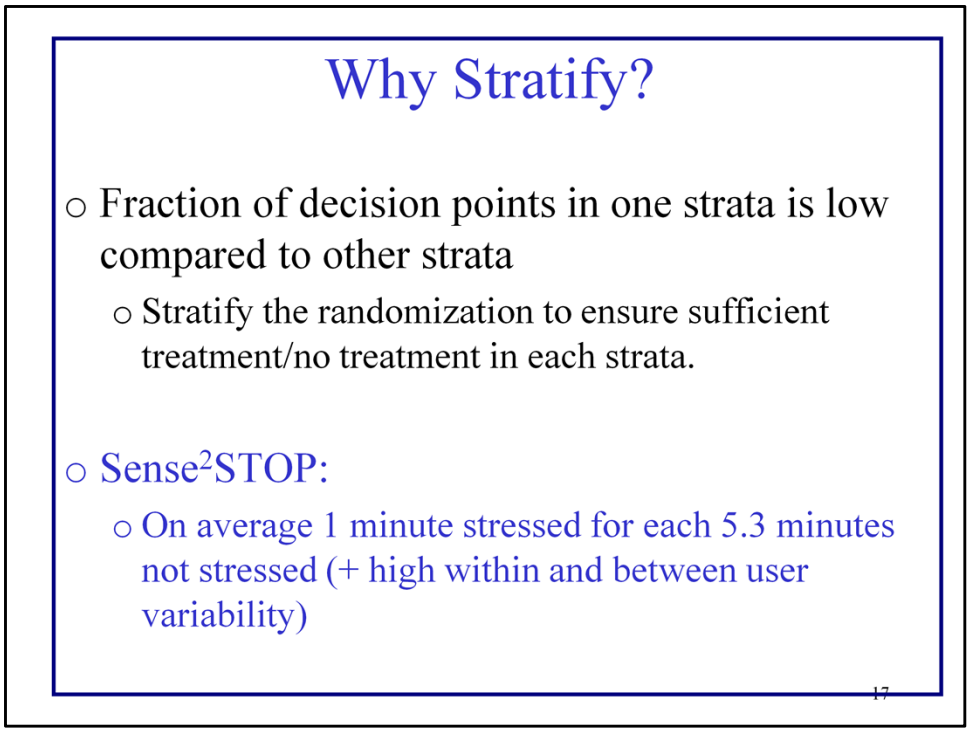
Delta will be 120 but in the sizing of the study we were thinking 60 min.

Why Micro-Randomize?

- In Sense^2 STOP randomization ensures that we can assess causal effects of the reminder.
  - Should the smartphone remind you to practice a stress-regulation exercise when (not) stressed?
  - Does this effect vary with time and/or current context?
- Sequential randomization due to sequences of treatment

16

We don't want to send these reminders if you don't need them!



From Peng
Originally there are 64 person day satisfying at least 12 hours. I did not consider the 10 the person-day in which there is no episode classified as Stress, thus resulting in 54 person days (I did this because the KL divergence is undefined in these cases..)
Below is the summary of stress and non-stress episodes.

(1) the entire 64 person-days

Stress

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	2.000	4.000	4.828	7.000	17.000

Not Stress

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.00	19.00	27.50	28.55	37.50	55.00

(2) after removing 10 person-days with no Stress

Stress

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.000	3.000	5.000	5.722	7.750	17.000

Not Stress

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.00	22.25	29.00	29.98	39.00	55.00

From our paper each stress episode lasts around 12 min and each non-stress episode lasts 11 min.

Stratified Micro-Randomized Trial

- Generally there is a "soft" budget, $\tilde{N}_x$, for the number of treatments that can be provided over T decision times for each strata of decision times.
  - $E\left[\sum_{t=1}^{T} A_t 1_{X_t=x} I_t\right] \cong \tilde{N}_x$
  - Budget is usually due to participant burden concerns
- In Sense^2 STOP
  - The budget is $E\left[\sum_{t\in day} A_t 1_{X_t=x} I_t\right] \cong 1.5$, for $x=0,1$

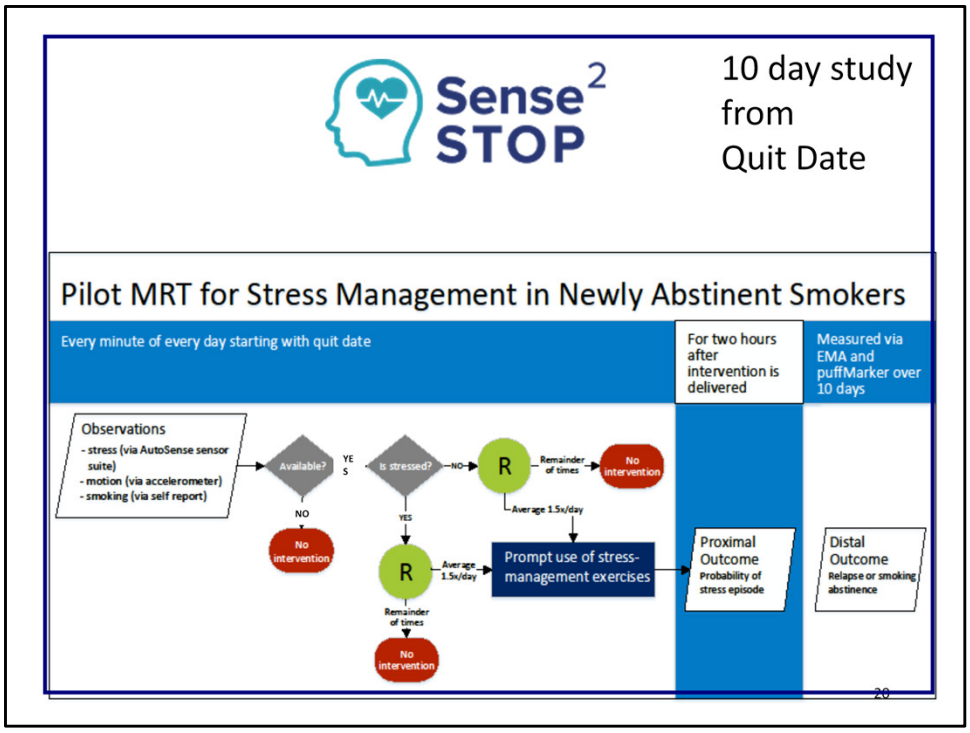
Delta will be 120

Randomization probabilities

- Given $H_t = \{O_1, A_1, \ldots, O_t\}$, $\mathrm{X_t} = x$ and $\mathrm{I}_t = 1$, we deliver the treatment at time t with probability

$$\mathrm{p_t}(H_t) = \frac{\widetilde{N}_x - \sum_{s=1}^{t-1} C_{t,\lambda}(x)}{1 + g(x \mid H_t)}$$

- $\widetilde{N}_x$ is desired average no. of treatments
- $C_{t,\lambda}(x)$: soft version of the number of treatments that have already been delivered that day
- $g(x \mid H_t)$: forecast of number of available decision points at level x remaining during day



Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day.

Outline

- Introduction to Mobile Health
- Sense^{2}STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator

21

Sense^2 STOP

- Primary: Should the smartphone notify the user with a reminder to utilize app directed stress-management exercises when the user is (not) stressed?
  - Does this effect vary with time or with current context?
- In the near term the reminder notification should reduce:
  - Near time, proximal, stress

22

Observations at different time scales

Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity. Hence, a total of 42% data is expected to be available from 16 hours of sensor wearing per day. Given these expectation, can we come up with a threshold for missing data from stress assessment?

To make English precise use potential outcomes

- $\bar{A}_t = \{A_1, \ldots, A_t\}$ (random treatments)
- $\bar{a}_t = \{a_1, \ldots, a_t\}$ (realizations of treatments)
- $Y_{t,\Delta}(\bar{a}_{t+\Delta-1})$ (one potential proximal response)
- $I_t(\bar{a}_{t-1})$ (one potential availability indicator)
- $\bar{H}_t(\bar{a}_{t-1})$ (one potential history vector)

Treatment effect

Define the individual level effect as a contrast between two excursions:

$$Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0\right) - Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, a_{t+1}, = 0 \ \ldots, a_{t+\Delta-1} = 0\right)$$

Treatment effect

Define the individual level effect as a contrast between two excursions:

$$Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0\right) - Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, a_{t+1}, = 0 \ \ldots, a_{t+\Delta-1} = 0\right)$$

Causal

Treatment effect

Define the individual level effect as a contrast between two excursions:

$$Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0\right) - Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, a_{t+1}, = 0 \ \ldots, a_{t+\Delta-1} = 0\right)$$

Absent assumptions above is not estimable

Because we are designing a study, we attempt to impose minimal assumptions. This is so we don't end up accidentally constraining the variety of analyses other scientists would like to do.

Treatment effect

$$Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, a_{t+1} = 0, \ldots, a_{t+\Delta-1} = 0\right) - Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, a_{t+1}, = 0 \ \ldots, a_{t+\Delta-1} = 0\right)$$

Define $\beta(t; x) =$
$E\left[\left(Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, \bar{0}\right) - Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, \bar{0}\right)\right) | I\left(\bar{A}_{t-1}\right) = \mathbf{1}, X_t\left(\bar{A}_{t-1}\right) = x\right]$

$\beta(t; x)$: causal excursion effect at time t beginning in strata x

Marginal, Conditional & Causal

$\beta(t; x)$: effect at time t in strata x

$$E\left[\left(Y_{t,\Delta}(\bar{A}_{t-1}, 1, \bar{0}) - Y_{t,\Delta}(\bar{A}_{t-1}, 0, \bar{0})\right) | I(\bar{A}_{t-1}) = 1, X_t(\bar{A}_{t-1}) = x\right]$$

- Expectation is over the distribution of the potential outcomes
- Effect is marginal over past, $\bar{H}_t(\bar{a}_{t-1})$
- Effect is conditional on availability and in strata x at time t.

Expression for Treatment Effect $\beta(t;x)$

- The sequential randomization + randomization probabilities bounded away from 0, 1 imply that $\beta(t;x)$ can be expressed in terms of the data distribution:

$$\mathrm{E}\left[E\left[\left(\prod_{j=t+1}^{t+\Delta-1}\frac{\mathbf{1}[A_j=0]}{p_j(H_j)^{A_j}\left(1-p_j(H_j)\right)^{1-A_j}}\right)Y_{t,\Delta}\mid A_t=1,H_t\right]\mid I_t=1,X_t=x\right]$$

$$-\mathrm{E}\left[E\left[\left(\prod_{j=t+1}^{t+\Delta-1}\frac{\mathbf{1}[A_j=0]}{p_j(H_j)^{A_j}\left(1-p_j(H_j)\right)^{1-A_j}}\right)Y_{t,\Delta}\mid A_t=0,H_t\right]\mid I_t=1,X_t=x\right]$$

- $p_j(H_j)$ is randomization probability

Outline

- Introduction to Mobile Health
- Sense^{2}STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator

30

Primary Hypothesis

Sense^2 STOP

- Consider decision points at which the individual is classified as stressed.
- We aim to contrast two treatment "excursions:"
  - (A) treatment now, no further treatment over subsequent 1 hour versus
  - (B) no treatment now, no further treatment over subsequent 1 hour
- Proximal response is fraction of time stressed over subsequent 1 hour.

Very imprecise english

Test Statistic

- Test statistic to test

$$H_0: \{\beta(t;x)\}_{t=1,\ldots,T,x=0,1} = 0$$

(e.g. is there anything going on here?!)

- Construct test statistic to target particular alternatives; consider alternatives of the form:
  - $\beta(t;x) = f_t(x)'\beta$ where $f_t(x) \in R^q$ is feature vector depending on t and x

Test Statistic

- Primary Hypothesis

$$H_0: \{\beta(t;x)\}_{t=1,\ldots,T,x=0,1} = 0$$

- Focus on alternatives of the form:
  - $\beta(t;x) = f_t(x)'\beta$ where $f_t(x) \in R^q$ is feature vector depending on t and x
  - If we suspect a decreasing effect with day in study, we might try to capture this is a coarse manner:
    - $f_t(x)' = \left(x, xd_t, xd_t{}^2, (1-x), (1-x)d_t, (1-x)d_t{}^2\right)$ or
    - $f_t(x)' = (x, xd_t, (1-x), (1-x)d_t)$

d_t is day in study at decision point t

Explain what you do if you want to just focus on x=1, stressed times.

Control variables

Used to reduce the variance/increase power

- Control variables will be used in a working model for the average proximal response:

$$E\left(.5Y_{t,\Delta}\left(\bar{A}_{t-1}, 1, \bar{0}\right) + .5Y_{t,\Delta}\left(\bar{A}_{t-1}, 0, \bar{0}\right)\middle| H_t\right) = g_t(H_t)'\alpha$$

for $g_t(H_t)$, a vector of summaries of prior data.

- The test statistic/Type 1 error rate will be robust to the mis-specification of this working model.

Weighted-centered least squares criteria

To construct the test statistic calculate

$$\arg\min_{\alpha,\beta} P_n \left[\sum_{t=1}^{T} I_t w_t(H_{t+\Delta-1}) \left(Y_{t,\Delta} - g_t(H_t)'\alpha - (A_t - .5) f_t(X_t)'\beta \right)^2 \right]$$

- P_n means average over participants' data
- $w_t(H_{t+\Delta-1}) = \frac{\prod_{s=1}^{\Delta-1} \mathbf{1}[A_{t+s}=0]}{\prod_{s=0}^{\Delta-1} p_{t+s}^{A_{t+s}} (1-p_{t+s})^{1-A_{t+s}}}$
- p_{t+s} is the randomization probability used in the study

Explain why A_t-0.5?!

Weighted-centered least squares criteria

To construct the test statistic calculate

$$\arg\min_{\alpha,\beta} P_n \left[\sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1}) \left(Y_{t,\Delta} - g_t(H_t)'\alpha - (A_t - .5) f_t(X_t)'\beta \right)^2 \right]$$

This results in $\hat{\beta}$.

An Aside

$\hat{\beta}$ is an estimator of

$$\beta^{\star} = \arg\min_{\beta} E\left[\sum_{t=1}^{T} I_t(\beta(t; X_t) - f_t(X_t)'\beta)^2\right]$$

Recall $f_t(x) \in R^q$ is feature vector depending on t and x

Test Statistic

To construct the test statistic calculate

$$\arg\min_{\alpha,\beta} P_n \left[\sum_{t=1}^{T} I_t w_{ct}(H_{t+\Delta-1}) \left(Y_{t,\Delta} - g_t(H_t)'\alpha - (A_t - .5) f_t(X_t)'\beta \right)^2 \right]$$

This results in $\hat{\beta}$.

We also construct an estimator of the standard error of $\sqrt{n}\hat{\beta}$ (n is the sample size): $\hat{\Sigma}$

This standard error must allow for unspecified correlation across time in the $Y_{t,\Delta}$

Test Statistic

$$T_n = n\hat{\beta}'\widehat{\Sigma}^{-1}\hat{\beta}$$

We also construct an estimator of the standard error of $\sqrt{n}\hat{\beta}$ (n is the sample size):

$$\hat{\Sigma}$$

This standard error allows for unspecified correlation across time in the $Y_{t,\Delta}$'s

\hat\Sigma is a robust standard error --sandwich formula

Hypothesis test

The rejection region for the test

$$H_0: \{\beta(t;x)\}_{t=1,\ldots,T,x=0,1} = 0$$

is:

$$\left\{T_n > \frac{q(n-(q'+1))}{n-(q'+q)} F^{-1}_{q,n-(q'+q);0}(1-\alpha_0)\right\}$$

where α_0 is the Type I error rate, q is the size of $f_t(x)$ and q' is the size of the controls, $g_t(H_t)$

Use two different small sample corrections, one in test statistic and the other in the critical value

Outline

- Introduction to Mobile Health
- Sense^{2}Stop and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Sample Size Calculator

41

Sample size formula

- Define $\gamma = (\beta^\star)'\Sigma^{-1}(\beta^\star)$
- Then the sample size is the smallest value n such that

$$1 - F_{q,n-(q'+q);n\gamma}\left(\frac{n-(q'+1)}{n-(q'+q)} F^{-1}_{q,n-(q'+q);0}(1-\alpha_0)\right) \geq 1-\beta_0$$

where α_0 is the Type I error rate and $1-\beta_0$ is the power

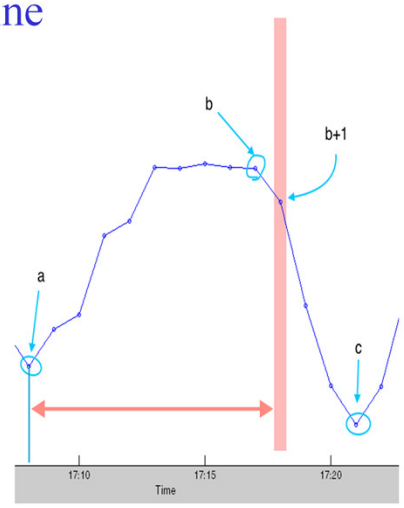
Inputs into the sample size calculation

- Desired Type I and Type II error rates,
- Targeted alternative $\{\beta(t;x)\}_{t=1,\ldots,T,x=0,1}$,
- Selected "control variables" $g_t(H_t)$,
- The randomization formula used to determine $P(A_t = 1 \mid h_t)$ given any history, $H_t = h_t$ and
- A generative model for $\{H_t\}_{t=1,\ldots,T}$.

"Data-driven" baseline generative model

We construct summary statistics using subset of the data collected in an observational, no treatment, smoking cessation study of 50 cigarette smokers

- For each episode type (i.e., $x \in \{0,1\}$), estimate the probability that the next episode will be a stress episode – i.e., a 2 by 1 vector $\bar{W}$
- For each episode type (i.e., $x \in \{0,1\}$), estimate the average episode length–i.e., a 2 by 1 vector $\bar{Z}$



"Data-driven" baseline generative model

We model the joint process (X_t, U_t) by a Markov chain

- X_t : Episode classification
- U_t: Stage of episode

Inputs:

- $\bar{W} = (.067, .519)$ (probability next episode is "stress")
- $\bar{Z} = (10.9, 12.0)$ (average duration of episode in min.)

TABLE 1

$P^{(0)}$: Transition Matrix for the Markov chain, V_t, under No Treatment

		Non-stress			Stress		
		Pre-peak	Peak	Post-peak	Pre-peak	Peak	Post-peak
Non-stress	Pre-peak	0.80	0.20	0.00	0.00	0.00	0.00
	Peak	0.00	0.00	1.00	0.00	0.00	0.00
	Post-peak	0.19	0.00	0.80	0.01	0.00	0.00
Stress	Pre-peak	0.00	0.00	0.00	0.82	0.18	0.00
	Peak	0.00	0.00	0.00	0.00	0.00	1.00
	Post-peak	0.09	0.00	0.00	0.09	0.00	0.82

Two values for \bar W, \bar Z correspond to not-stress, stress starting state

Generative model under treatment

- (X_t, U_t), is a Markov chain
  - X_t : Episode classification
  - U_t: Stage of episode
- We compute the alternative transition matrix that achieves the targeted alternative. This becomes the transition matrix for 60 time points after a treatment.

Type I=.05
Type II=.2

Sample Size Calculation

- Type I error = 0.05; Type II error = 0.20
- Targeted Alternative: $\beta(t;x) = f_t(x)'\beta$
- Working model for average proximal response $g_t(H_t)'\alpha$
  - where $f_t(x)' = g_t(x)' = \left(x, xd_t, xd_t^2, (1-x), (1-x)d_t, (1-x)d_t^2\right)$

Results:

Table 1: Estimated sample size, n, and achieved power.

	Sample size	Power
$\bar{\beta} = 0.030$	50	80.6
$\bar{\beta} = 0.025$	67	80.7
$\bar{\beta} = 0.020$	127	80.6

(1) an initial conditional effect= 0, (2) the day of maximal effect = 5 days and (3) the average conditional treatment effect is the same for both x=0 (not stress) and x=1 (stress)—these are the rows in table 1

Increasing Robustness to misspecification of the transition matrix

Recall (X_t, U_t), is a Markov chain

- X_t : Episode classification
- U_t: Stage of episode

We compute the largest sample size under the inputs:

- $\bar{W} = (.067 \pm \epsilon, .519 \pm \epsilon)$
- $\bar{Z} = (10.9 \pm \epsilon', 12.0 \pm \epsilon')$

Table 1: Estimated sample size, n, and computed power under $\epsilon = 0.01$ and $\epsilon' = 2$.

	Sample size	Minimum Power
$\bar{\beta} = 0.030$	69	81.9
$\bar{\beta} = 0.025$	107	80.4
$\bar{\beta} = 0.020$	208	80.5

This is work in progress

For the average durations \bar Z we have:
Mean for NS, S: 10.9, 12.0 (in paper) with standard errors (0.12 for NS and 0.28 for S).
SD for NS, S: 6.89, 6.48

Choosing the new \bar Z should reflect sufficient deviation from the current \bar Z, but stay within the SDs of around 6.5 to 7.0. Thus the choice of 2

For \bar W, I can calculate the SD for the estimated fractions (i.e., \bar W) obtained from the data. This is 0.005 for NS and 0.03 for S. Given this the value 0.01 made sense as potential deviations. I wanted to choose one level across both S and NS, and thus opted for the two I chose.

Intuition Behind the Sample Sizes

- Because the alternative hypothesis targets a low dimensional alternative instead of all possible functions of time, t and stress, x, the calculator is able to use within participant contrasts to increase power and ↓ sample size

- $Sense_2STOP$ is a short study of only 10 days with few treatments per day, e.g. 3; this decreases power and ↑ sample size
- Small effect sizes decreases power and ↑ sample size

49

Micro-Randomized Trials: When are they (not) useful?

- NOT USEFUL. When circumstances are rare: Want to learn the best type of alert to prevent suicide attempt
- USEFUL. When circumstances change rapidly: Stress, urges to smoke, adherence, physical activity
- NOT USEFUL. Proximal outcome cannot be easily sensed or observed: suicide ideation, craving
- USEFUL. Proximal outcome can be unobtrusively sensed or unobtrusively self-reported.

50

Last two are in a state of flux and depend on the current sensor technology.

What are we learning?

- Communication is *CRITICAL!*
  - Health Scientist requires a translation of the steps in the data algorithms ("machine learning algorithms/data analytics") into English
  - Health Scientist feeds back his/her understanding of the algorithm to Data Scientist
- The data algorithms are ***part*** of the clinical trial protocol.
- Data Scientist's criterion for excellence is not always the same as the Health Scientist's criterion for excellence

51

pinpointing where in stress episode to intervene
Language regarding what times constitute not stressed times
How to handle missing data
Tradeoff between sensitivity/specificity of stress classifications and the number of times during the day as which you can check if an intervention push is useful

Collaborators!