



Title: 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.

Experimentation, Continual Optimization

- “Iterative nature of experimentation” (RA Fisher & G. Box)
- “At Google, experimentation is practically a mantra; we evaluate almost every change that potentially affects what our users experience.” (4 Google scientists)
- “Online experiments are widely used to compare specific design alternatives, but they can also be used to produce generalizable knowledge and inform strategic decision making. Doing so often requires sophisticated experimental designs, iterative refinement, and careful logging and analysis.” (3 Facebook scientists)

2

Fisher and G. Box in industrial engineering philosophy: “iterative nature of experimentation” when the underlying system is not under control. Always be clear is about what the goal is –what is it you want to optimize.

Data + idea -> deduction (used to test a theory)

Data – idea → induction (used to generate new theory)

Can change the objective after you experiment

http://www.statisticsviews.com/details/video/5018561/The-Iterative-Nature-of-Experimentation---Part-II-by-George-E_P_-Box.html

Quote from paper by 4 Google scientists: At Google, experimentation is practically a mantra; we evaluate almost every change that potentially affects what our users experience. Such changes include not only obvious user-visible changes such as modifications to a user interface, but also more subtle changes such as different machine learning algorithms that might affect ranking or content selection. Our insatiable appetite for experimentation has led us to tackle the problems of how to run more experiments, how to run experiments that produce better decisions, and how to run them faster. Overlapping Experiment Infrastructure: More, Better, Faster Experimentation Diane Tang, Ashish Agarwal, Deirdre O’Brien, Mike Meyer Google, Inc. Mountain View, CA [diane,agarwal,deirdre,mmm]@google.com

Quote from paper by 3 facebook scientists and cs at stanford: “Online experiments are widely used to compare specific design alternatives, but they can also be used to produce generalizable knowledge and inform strategic decision making. Doing so often requires sophisticated experimental designs, iterative refinement, and careful logging and analysis.”

Outline



- Introduction to Mobile Health
- Sense²STOP
- Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

3

Smart wt loss



JOOLHEALTH



BariFit



HeartSteps



SARA



Sense²STOP

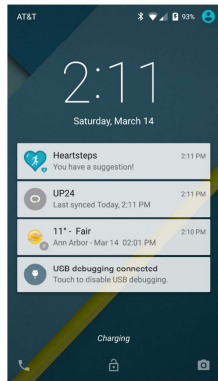
<https://methodology.psu.edu/ra/adap-inter/mrt-projects#proj>

Mobile Intervention Treatments

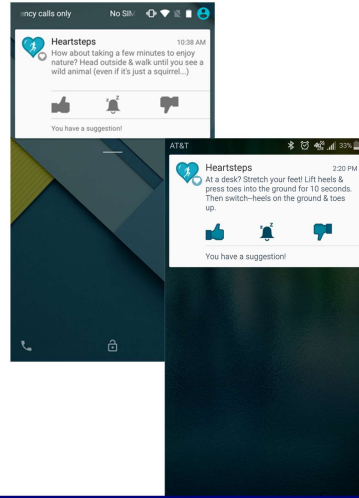


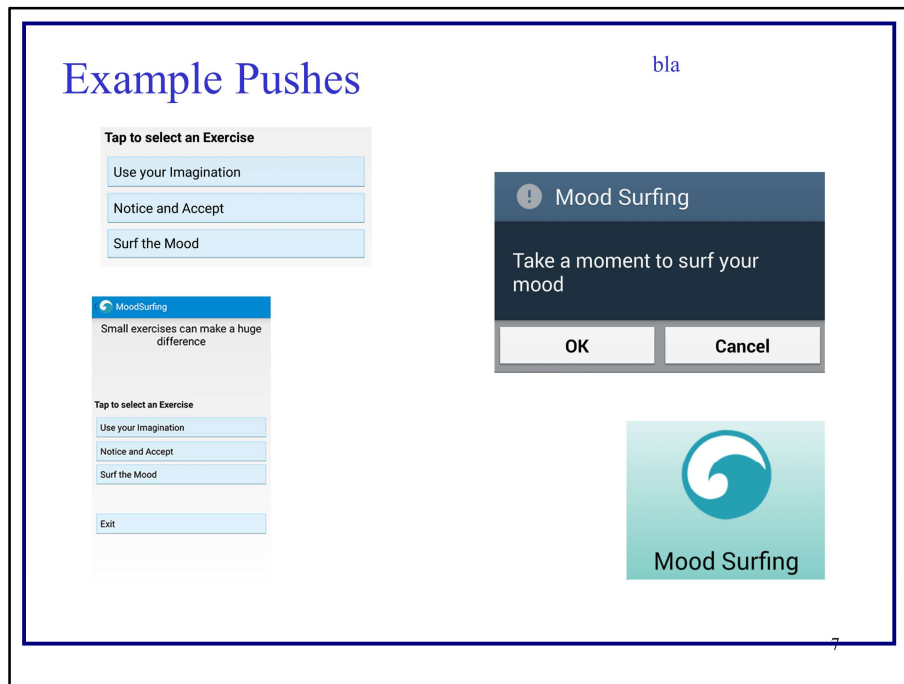
Push has both positive and negative effects

Example Pushes



In-the-Moment Tailored Activity Suggestions





EMA types:

- Random Experience Sampling
- Event Contingent Recording
- End of Day Recording

Intervention types:

- Moodsurfing
- Thought shakeup
- Headspace

Prompts: These are the messages that would appear on your lock screen

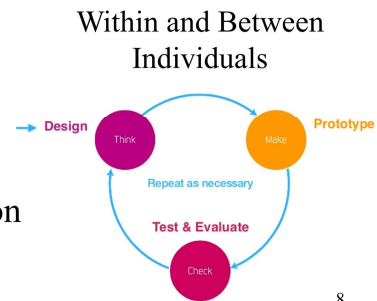
- Time to take a Survey
- Take a moment to surf your mood
- Take a moment to make some head space
- Take a moment to shake up your thoughts

Experimentation, Continual Optimization

- Learning/experimentation prior to mHealth intervention evaluation

→

- Learning/experimentation during mHealth intervention implementation



Underpins the way we are thinking in terms of causality and treatment effects

Iterative design or design thinking. Problem, express, test, feedback, cycle, and restart.
Experimentation

Design and test new products.

Iterative design https://en.wikipedia.org/wiki/Iterative_design in **Human computer interfaces**

Design Thinking: Express–Test–Cycle

Outline



- Introduction to Mobile Health
- **Sense²STOP**
- Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

9



PI: S. Kumar

Most (93%) unaided smoking cessation attempts fail in 1st week

- 95% of **lapses** (few puffs) followed by **relapse**
- Patients are encouraged to call when tempted to smoke.
...but they rarely do

Stress predicts lapse/relapse=> increasing state of risk

- Performing brief relaxation exercises can buffer/blunt real-life life stress
- ***But people fail to use them***
- **Should the phone push reminders to individuals at times of stress to access exercises on smartphone?**

10

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.

 **Using sensors to detect “stress”**

- Participant wears Autosense chestband + sensors on each wrist
- Measure various physiological responses and body movements to robustly assess physiological stress.
- Pattern-mining algorithm uses the sensor data to construct a binary time-varying stress classification
- Participant is then classified at each minute as either “Stressed” or “Does not qualify as Stressed” or “UK”



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,

Uses ECG features, Respiration features at one min. level

inspiration duration, expiration duration, respiration duration, i-e duration, ratio, stretch, respiratory sinus

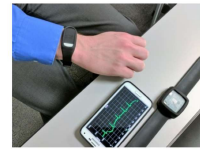
arrhythmia (RSA), breath-rate²; inspiration minute volume² use mean, median, 80th percentile, quartile, deviation

Table 2. All of the base and aggregate RIP features, which are computed by our system. 1: RSA is a hybrid feature that uses both RIP and ECG signals. 2: The aggregate features breath rate and inspiration minute volume are computed without any other base features, but rather using just the number of respiration cycles in a minute.

SVM uses radial basis functions and then Platt's scaling to turn output into probabilities

Ertin, E., Stohs, N., Kumar, S., Rajj, 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.



- In the near term the reminder push notification should reduce:
 - Near time, proximal, stress
- Primary: Should the smartphone push a reminder to utilize app directed stress-management exercises when the user is stressed?
 - Does the effect vary with time or with current context?

12

Observations at different time scales

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
- Sense²STOP
- Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

13

Stratified Micro-Randomized Trial

- On each participant: $O_1, A_1, \dots, O_t, A_t, \dots$
- t : 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

Observation, treatment action, observation, treatment action,.....

In clinical trials many people use blocking and stratification interchangeably:

Stratification (clinical trials)

From Wikipedia, the free encyclopedia

Stratification of [clinical trials](#), is the partitioning of subjects and results by a factor other than the treatment given.

Stratification can be used to ensure equal allocation of subgroups of participants to each experimental condition. This may be done by gender, age, or other demographic factors. Stratification can be used to control for [confounding variables](#) (variables other than those the researcher is studying), thereby making it easier for the research to detect and interpret relationships between [variables](#).^[1] For example, if doing a study of fitness where age or gender was expected to influence the outcomes, participants could be stratified into groups by the confounding variable. A limitation of this method is that it requires knowledge of what variables need to be controlled.^[1]

Types of stratification

[Stratified random sampling](#) designs divide the population into [homogeneous](#) strata, and an appropriate number of participants are chosen at random from each strata.^[1] Stratification is sometimes called blocking, and may be used in [randomized block design](#).^[1]

From

<http://mathforum.org/kb/thread.jspa?forumID=67&threadID=213792&messageID=747629>

Advanced placement for high school students!

The defining difference is that blocking is an experimental design concept, and stratification is a sampling concept.

In the randomized block design one looks for homogeneous chunks of experimental material -- the blocks. To these blocks, each experimental treatment is assigned. The idea is that if the experimental material is similar in some respect, any difference on average between treatments cannot be due to any big difference in the material. Suppose, as we frequently do in Iowa, that the most important thing in the world is tall corn. Toward that end we want to see whether fertilizer A and fertilizer B are equally good or if one is better. If we (by chance) assigned more of fertilizer A to "excellent" soil and more of fertilizer B to "top notch" soil, the (presumed) success of fertilizer A could be due not to characteristics of fertilizer A, but to the better soil. (We don't have any soil worse than top notch.)

In order to get a more truthful picture of the two fertilizers, we create blocks of soil equally wonderful. We then make sure to use fertilizer A and fertilizer B in this soil. (No, not at the same time -- we might have plots on the same section, with A and B randomly assigned.) Then, since fertilizers A and B appear equally in different qualities of soil, soil quality difference cannot explain differences in the height of the corn.

Now let's turn to stratification. Suppose that we are interested in the second most important thing in the world, the sweetness of sweetcorn. I have no idea what makes sweetcorn sweet, but again let us suppose that it is related to the type of soil. If I wish to estimate the sweetness this year's sweetcorn crop, I could take a random sample of sweetcorn fields, and measure the sweetness. (It is a little known fact that sweetness is measured on the Audrey Hepburn scale, where 0 is the sweetness of Attila the Hen, and 5 is the sweetness of Audrey Hepburn. For those of less than a certain age, think of 0 = Brittany Spears now, and 10 = Britany Spears a few years ago.)

Anyway, if we take a simple random sample of fields of sweetcorn, it is possible that we might accidentally get an over-representation of sweet fields, and an under-representation of the other fields (sour?). We will get a better estimate of the sweetness if we divide the State of Iowa into squares, and sample from each square -- b/c we select samples from (say) each county in Iowa, it is more likely that we will get a sample that is representative of the variety of sweetness around the state. More to the point here, within each county the variability of sweetness will be less, and the resulting mean sweetness (acquired by taking a weighted average of the sweetnesses in each county) will have a sampling distribution that is less variable, leading to greater precision in the estimate.

With blocking, the homogeneity in the blocks leads to a smaller amount of extraneous

variation in the experiment, which gives one a smaller denominator for the t , and leads to greater power. With stratification, the homogeneity in the strata leads to a smaller amount of variation in each sub-sample, and when combined over the whole leads to a smaller amount of total variation, which results in smaller confidence intervals in the estimation procedure.

Stratified Micro-Randomized Trial

- On each participant: $O_1, A_1, \dots, O_t, A_t, \dots$
- t : Times at which a treatment might be provided
 - Sense²STOP: 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
 - Sense²STOP: $X_t = 1$ 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
 - Sense²STOP: $X_t = 1$ each minute if classified as stressed and =0, otherwise
 - $I_t \subset O_t$: availability indicator
 - Sense²STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; =0 otherwise

Available if data is good, not very physically active at this time, 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
 - Sense²STOP: $I_t = 1$ if not treated in prior hour and if online classification is possible; =0 otherwise

Randomization occurs only if $I_t=1$

- A_t : Randomized treatment at time t
 - Sense²STOP: $A_t = 1$ if reminder is pushed to participant; =0 otherwise


Stratified Micro-Randomized Trial

- A_t : Randomized treatment at time t
 - Sense²STOP: $A_t = 1$ 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 times
 - Sense²STOP: fraction of time stressed in $\Delta=60$ 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  randomization ensures that we can assess causal effects of the reminder.
- Sequential randomization facilitates investigating:
 - Does the reminder vs. no reminder impact stress management in the near term?
 - Does this effect vary with time and/or current context?

19

Ill- posed questions!

Why Stratify?

- Fraction of time points in one strata is low compared to other strata
 - Stratify the randomization to ensure sufficient treatment/no treatment in each strata.
- Sense²STOP:
 - On average 1 minute stressed for each 6 minutes not stressed

20

From Peng

Originally there are 64 person day satisfying at least 12 hours. I did not consider the 10 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 times for each strata.
 - $E[\sum_{t=1}^T A_t 1_{X_t=x} I_t] \cong \tilde{N}_x$
 - Budget is usually due to participant burden concerns
- In  Sense² STOP
 - the budget is $E[\sum_{t \in \text{day}} A_t 1_{X_t=x} I_t] \cong 1.5$, for $x=0,1$

This is pre-relapse budget; post-relapse there are not enough detectable stress times to meet this budget.

X=0 means not classified as stressed

X=1 means classified as stressed

Optimization Criterion for Randomization Probabilities

Subject to the budget constraint

$$E \left[\sum_{t \in \text{day}} A_t 1_{X_t=x} I_t \right] \cong 1.5$$

Minimize deviance from uniform randomization
across all person-time points at which $X_t = x$, $I_t =$
1,

- Minimize Kullback-Leibler divergence of
randomization probabilities from uniform
distribution.

Randomization probabilities

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

$$p_t(H_t) = \frac{\tilde{N}_x - C_{t,\lambda}(x)}{1 + g(x | H_t)}$$

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

We used a forecast built from a markov chain model for episodes, parametric models for episode duration and multinomial model for time to peak. The only data used as an input to this forecast was current x and time remaining in day.

For $\lambda=1$ you will end up with low variability in number of treatments across days – this one counts no. of treatments already provided today.

If forecast is perfect, then $\lambda=1$ gives you exactly 1.5 treatments per day and in $\lambda=0$ then you get uniform distribution of treatments across stress times

However another alternative might be a poisson regression model:

training set with the outcome being the count of

future "Stress" episodes and the input features: the remaining time of the day, the numbers of

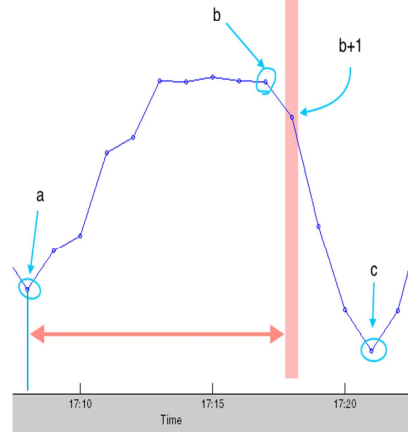
"Stress", "Not Stress" and the "Unknown" episodes so far in the day and the indicator of lapse. To account for

availability, these forecasts are further discounted by a constant

Forecast is based on a no-treatment-effect model

Forecast based on data collected in an observational, no treatment, smoking cessation study of 45 cigarette smokers wearing the same sensor suite

- For each episode type (i.e., $x \in \{0,1\}$), estimate the probability that the next episode will be a stress episode to form Markovian transition matrix
- For each episode type (i.e., $x \in \{0,1\}$), use a parametric model for the episode length



Input to prediction is remaining time in day and whether person is currently in $X=1$, or $X=0$ state.

Markov renewal process.

The estimated transition model based on the Minnesota data will be used 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.

we use X_s to denote the classification of s th episode in the Minnesota trial.

Thus $X_s = 1$ means "Not Stress", $X_s = 2$ means "Stress", and $X_s = 3$ means "Unknown".

We assume a Markov transition model for X_s and assume a parametric model for Δ_s (length of episode) given

X_s

Δ_s : Gamma distribution when $X_s = 1; 2$ and Log Normal distribution when $X_s = 3$.

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 length increments -- $k = 1; \dots; 20$ and use a

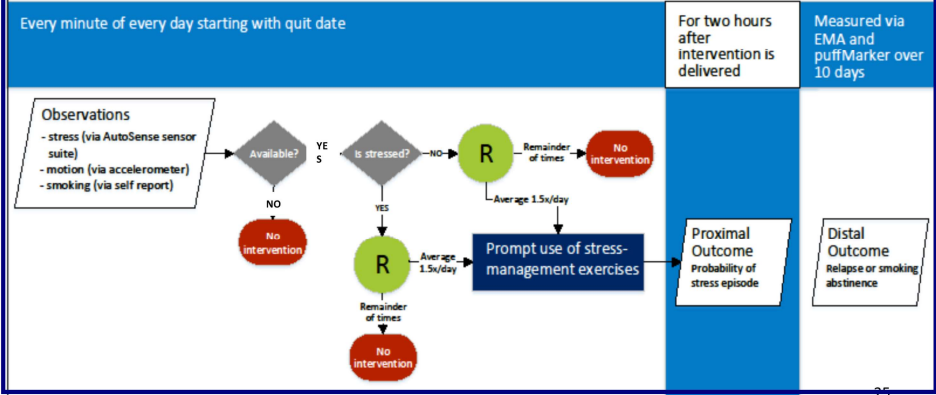
Multinomial

distribution. All of above estimation will be done separately for pre-lapse and post-lapse



10 day study
from
Quit Date

MRT for Stress Management in Newly Abstinent Smokers



Aiming for n=75

Observations at different time scales

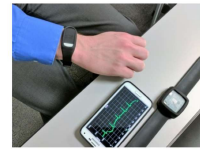
Statistically speaking, 28% of the time data is lost due to sensor loosening, sensor detachment, battery down, taking the sensor off, noisy data due to jerks, etc. For stress assessment, 23% of the time people are physically active, and it takes 7% additional time for physiology to recover from activity.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

Outline



- Introduction to Mobile Health
- Sense²STOP
- Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

26



- Primary: Should the smartphone push a reminder to utilize app directed stress-management exercises when the user is stressed?
 - The reminder push notification should reduce near time stress compared to no reminder.
 - Does the effect vary with time or with current context?

View through the lens of Continual Learning/Experimentation

27

Observations at different time scales

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.

To make English precise use potential outcomes

- $\bar{A}_t = \{A_1, \dots, A_t\}$ (random treatments);
- $\bar{a}_t = \{a_1, \dots, a_t\}$ (realizations of treatments)
- $Y_{t,\Delta}(\bar{a}_{t+\Delta-1})$ (one potential proximal response)
Recall $Y_{t,\Delta} = \Delta^{-1} \sum_{s=1}^{\Delta} 1[X_{t+s} = 1]$
- $I_t(\bar{a}_{t-1})$ (one potential availability indicator)
- $H_t(\bar{a}_{t-1})$ (one potential history vector)

Need to figure out how to communicate that you want to view this within the lens of sequential experimentation/learning/decision making

Treatment effect

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

$$Y_{t,\Delta}(\bar{A}_{t-1}, a_t = 1, a_{t+1} = 0, \dots, a_{t+\Delta-1} = 0) \\ - Y_{t,\Delta}(\bar{A}_{t-1}, a_t = 0, a_{t+1} = 0, \dots, a_{t+\Delta-1} = 0)$$

Causal

Recall definition of $Y_{t,\Delta}$

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

Treatment effect

The individual level effect is a contrast between two excursions from the present course, \bar{A}_{t-1} :

$$Y_{t,\Delta}(\bar{A}_{t-1}, a_t = 1, a_{t+1} = 0, \dots, a_{t+\Delta-1} = 0) \\ - Y_{t,\Delta}(\bar{A}_{t-1}, a_t = 0, a_{t+1} = 0, \dots, a_{t+\Delta-1} = 0)$$

a short trip or outing to some place, usually for a special purpose
deviation from a direct, definite, or proper course

Treatment effect

The individual level effect is a contrast between two excursions:

$$Y_{t,\Delta}(\bar{A}_{t-1}, 1, \bar{0}_{t+1:t+\Delta-1}) - Y_{t,\Delta}(\bar{A}_{t-1}, 0, \bar{0}_{t+1:t+\Delta-1})$$

Absent assumptions, individual level effect 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

Absent strong assumptions, individual level effect is not estimable.....

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

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

Marginal, Conditional & Causal

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

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

- Effect is conditional on availability and in strata x at time t : $I_t = 1, X_t = x$
- Effect is marginal over past data, H_t
- Why are we focusing on contrasts between excursions?

Why are we considering excursions?

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:

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

- $p_j(H_j)$ is randomization probability

Outline



- Introduction to Mobile Health
- Sense²STOP and Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

35

Primary Hypothesis



- 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

Experiment was later revised and we used subsequent 2 hours

Test Statistic

- Primary Hypothesis Test:

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

(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

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

Trade off bias and power via low dimensional alternatives.

--- capture a decreasing effect with day in study in a coarse manner:

$$H_1: \beta(t; x) = f_t(x)' \beta, \quad \beta \neq 0$$

- $f_t(x)' = (x, x d_t, x d_t^2, (1-x), (1-x) d_t, (1-x) d_t^2)$
or
- $f_t(x)' = (x, x d_t, (1-x), (1-x) d_t)$

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:

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

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

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

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

Explain why $A_t - 0.5$?!

Estimand for Test Statistic

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

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

$\beta(t; x)$ is time varying causal excursion effect.

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

Test Statistic

To construct the test statistic calculate

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

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

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

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

Hypothesis test

The rejection region for the test

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

is:

$$\left\{ n\hat{\beta}'\hat{\Sigma}^{-1}\hat{\beta} > \frac{q(n - (q' + 1))}{n - (q' + q)} F_{q, n - (q' + q); 0}^{-1}(1 - \alpha_0) \right\}$$

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, $g_t(H_t)$

Stand on the shoulders of the scientists who came before you!

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

Outline



- Introduction to Mobile Health
- Sense²STOP
- Stratified Micro-Randomized Trials
- The Causal Treatment Effect (a.k.a, causal excursions)
- Test Statistic for Primary Hypothesis
- Discussion

44

Comments on Sample Size



- Type I error = 0.05; Power = 0.80
- Targeted Alternative: $\beta(t; x) = f_t(x)' \beta$
- Working model for average proximal response $g_t(H_t)' \alpha$
 - where $f_t(x)' = g_t(x)' = (x, x d_t, x d_t^2, (1-x), (1-x) d_t, (1-x) d_t^2)$

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

$\bar{\beta}$ is a standardized average effect size

(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

Intuition Behind the Sample Sizes

- Low dimensional alternative: \uparrow power and \downarrow sample size--test is able to use within participant contrasts
- Small effect sizes: \downarrow power and \uparrow sample size
- Short study with small no of treatments per day: \downarrow power and \uparrow sample size

Continual Learning/Experimentation

- Where is all of this going?
A mobile health intervention that incorporates continual learning/personalization.
- Treatment design and experimental design are intertwined when the goal is to continually learn and personalize.
- Algorithms are not only ***part*** of the experimental design they are treatment components in the mobile health intervention.

47

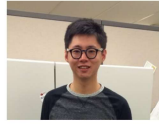
pinpointing where in stress episode to intervene

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

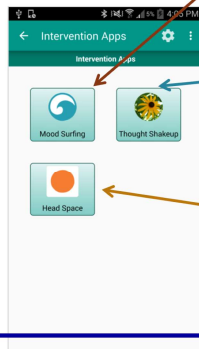
Collaborators!



Intervention Push is a Reminder to Access Stress Management Apps:

Apps employ

- Evidence-based exercises to manage stress
- Take about 3-5 minutes to practice
- Developed and refined based on input from experts and users



Mood Surfing:

- 3 exercises
- Grounded in ACT
- Target cognitive defusion
- Experts: K. Witkiewitz, I. Yovel.
- Literacy level editor: A. Applegate
- HCI: M. Sharmin; Programmer: M. Hossain

Thought Shakeup

- Grounded in CBT
- Target cognitive restructuring
- Expert: I. Yovel.
- Literacy level editor: A. Applegate
- HCI: M. Sharmin; Programmer: M. Hossain

Head Space

- Grounded in ACT
- Mediation / Mindfulness
- Consistently rated as one of the best 5 commercial mediation apps
- Permission for free use in the trial

49

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]

Cole Porter's "Experiment"

(Verse)

Before you leave these portals
To meet less fortunate mortals
There's just one final message
I would give to you

You all have learned reliance
On the sacred teachings of
science
So I hope, through life, you
never will decline
In spite of Philistine defiance
To do what all good scientists
do

(Chorus)

Experiment, make it your motto day
and night
Experiment and it will lead you to
the light
The apple on the top of the tree is
never too high to achieve
So take an example from Eve
Experiment

Be curious
Though interfering friends may
frown
Get furious
At each attempt to hold you down

If this advice you always employ
The future can offer you infinite joy
And merriment
Experiment
And you'll see

50

Performed in 1933 but listen to Tony Bennett's recording at
<https://www.youtube.com/watch?v=9bp1VYMrPMA>



Micro-Randomized Trials

SARC (Smartphone Addiction Recovery Coach) MRT

This project tests the feasibility and effectiveness of providing, via a smartphone, messages designed to encourage use of the ecological momentary interventions (EMIs) on the phone so as to support adolescents as they recover from substance use. The resulting data will be used to inform the development of a JITAI to support recovery.

This is a 6 month study. The distal outcome, frequency of substance use during the 6 months, is measured via the 6 month Substance Frequency Scale score. Ranging from 0 to 100% the Substance Frequency Scale ($\alpha=.85$; test-retest $\rho=.94$) is based on the average percent of days reported of alcohol, cannabis, stimulants, opioids, and other substance use, days of heavy substance use, and days of problems from substance use. The participants are youth ages 15-18 enrolled in outpatient substance use program; in particular a total of 300 adolescents will be recruited after 2 sessions of adolescent outpatient substance use disorder treatment within a 6 week period. The participants are expected to be at least 25% female, 25% non-white, 25% white and approximately 80% between the ages of 15-18. The primary substances used included cannabis, tobacco and alcohol, with subsets also using opioids/prescription meds and stimulants. Over half are expected to one or more co-occurring mental health disorders including ADHD, CD, Mood, Anxiety, and Trauma related disorders.

After each of 5 EMAs per day, a participant is randomized 1:1 to an encouraging message or a simple thank you. The encouraging messages are randomly selected from one of 4 message framing bins: Gain/proximal, Loss/proximal, Gain/distal, and Loss Distal. Regardless of randomization, the participant is subsequently provided links to a variety of the EMIs available on the phone appear. The proximal outcome of the encouraging message is EMI use within 1 hour.



Stratified Micro-Randomized Trials

SARC (Smartphone Addiction Recovery Coach) MRT

This project tests the feasibility and effectiveness of providing, via a smartphone, messages designed to encourage use of the ecological momentary interventions (EMIs) on the phone so as to support adolescents as they recover from substance use. The resulting data will be used to inform the development of a JITAI to support recovery.

This is a 6 month study. The distal outcome, frequency of substance use during the 6 months, is measured via the 6 month Substance Frequency Scale score. Ranging from 0 to 100% the Substance Frequency Scale ($\alpha=.85$; test-retest $\rho=.94$) is based on the average percent of days reported of alcohol, cannabis, stimulants, opioids, and other substance use, days of heavy substance use, and days of problems from substance use. The participants are youth ages 15-18 enrolled in outpatient substance use program; in particular a total of 300 adolescents will be recruited after 2 sessions of adolescent outpatient substance use disorder treatment within a 6 week period. The participants are expected to be at least 25% female, 25% non-white, 25% white and approximately 80% between the ages of 15-18. The primary substances used included cannabis, tobacco and alcohol, with subsets also using opioids/prescription meds and stimulants. Over half are expected to one or more co-occurring mental health disorders including ADHD, CD, Mood, Anxiety, and Trauma related disorders.

After each of 5 EMAs per day, a participant is randomized 1:1 to an encouraging message or a simple thank you. The encouraging messages are randomly selected from one of 4 message framing bins: Gain/proximal, Loss/proximal, Gain/distal, and Loss Distal. Regardless of randomization, the participant is subsequently provided links to a variety of the EMIs available on the phone appear. The proximal outcome of the encouraging message is EMI use within 1 hour.

Continual Mobile Intervention Optimization

Is not mundane!!

Exploration produces variation

- used by us to learn and optimize
- within person exploration → within person variation
 - can reduce user habituation, boredom
 - can increase user novelty, positive arousal
 - can provide variable reinforcement
 - enables dynamic, within person, learning

Variable reinforcement is less likely to extinguish

The systematic variation will locate contexts in which message is most helpful but from the user's perspective still may improve