Real-time, Within Person-
Randomization using
a Bandit Algorithm
in a Clinical Trial

Susan A Murphy
08.01.18
Mobile Health Treatments
HeartSteps (PI Klasnja)

Goal: Develop an mobile activity coach for individuals who are at high risk of adverse cardiac events.

Three iterative studies:

- **V1**: 42 day micro-randomized pilot study with 37 sedentary individuals,
- **V2**: 90 day micro-randomized (partially via a bandit) study,
- **V3**: 365 day personalized study
Data from wearable devices that sense and provide treatments

- On each individual: \( S_1 A_1 R_2, \ldots, S_t, A_t, R_{t+1}, \ldots \)

- \( t \): Decision point

- \( S_t \) : Context accrued after \( t-1 \) and up to/including decision point \( t \) (high dimensional)

- \( A_t \) : Action at \( t^{th} \) decision point (treatment)

- \( R_{t+1} \) : Reward (e.g., utility, cost) accrued after time \( t \) and prior to time \( t+1 \)
HeartSteps V1

Heartsteps MRT to Promote Physical Activity Among Sedentary People

Each day of study
Observations are continuous (except self report)
Randomizations to activity prompts occur 5x/day at likely times for increasing physical activity

<table>
<thead>
<tr>
<th>Observations</th>
<th>Start Intervention Tailored prompt to become physically active</th>
</tr>
</thead>
<tbody>
<tr>
<td>- location (via GPS)</td>
<td>- weather, location, time of day, day of week</td>
</tr>
<tr>
<td>- weather (via Internet)</td>
<td></td>
</tr>
<tr>
<td>- motion (via wrist band)</td>
<td></td>
</tr>
<tr>
<td>- usefulness of prompt (via user indication)</td>
<td></td>
</tr>
<tr>
<td>- self report of activity (via app in evenings)</td>
<td></td>
</tr>
</tbody>
</table>

Next 30 minutes after intervention is delivered
Proximal Outcome
physical activity (steps taken)

Measured via accelerometer throughout study
Distal Outcome
Overall activity in the 42-day study

Following day
Proximal Outcome
physical activity (steps taken)

No intervention
Average of 1x every other day

Driving?
Yes -> No intervention
No -> 1x/day in evening

Walking?
Yes -> Stop
No -> No intervention

Average 3x/day

Start Intervention Prompt planning of next day's activity
Actions

• Contextually tailored activity suggestion (provide yes/no)

• The set of actions may depend on the context, $S_t$
Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to predict succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days, prior 30 min step count, total steps on prior day, current temperature
Some Results from HeartSteps V1

1) The tailored activity suggestion, as compared to no activity suggestion, indicates an initial increase in step count over succeeding 30 minutes by approximately 271 steps but by day 20 this increase is only approximately 65 steps.

2) Features that appear to interact with treatment on succeeding 30 minute step count:
   1) Time in study, recent number of messages sent, location, variability in step count in 60 min window over previous 7 days
Goal

During HeartSteps V2 select binary treatment action---whether to provide tailored activity suggestion--online so as to maximize the sum of rewards for each user over the 90 day study (subject to constraints).

1. 5 decision points per day (set according to user’s work schedule)
2. Reward is the 30 minute stepcount following each decision point $t$. 
A “Bandit” Algorithm

Overview:

1) Initialize parameters in expected reward, \( r(S, A) \), given the treatment, \( A \) and context feature, \( S \).

2) At time point \( t \): input current features, \( S_t \) and select treatment, \( A_t \).

3) After the time point: input the reward, \( R_{t+1} \).

4) The algorithm updates the expected reward, as a function of both the treatment and features.

5) Given the context features at the next time point, \( S_{t+1} \) the algorithm uses the updated expected reward to select next treatment, \( A_{t+1} \). Go to 3) with \( t = t + 1 \).
Linear Thompson Sampling Bandit

1) Linear model for expected reward, e.g. \( r(S_t, A_t) = E[R_{t+1} | S_t, A_t] = \eta^T f(S_t, A_t) \)

2) Initialize \( \eta \) parameters in expected reward with a prior distribution (here a Gaussian).

3) Given \( S_t, A_t, R_{t+1} \) update posterior distribution of \( \eta \). Mean, covariance matrix of this posterior distribution is \( \eta_t, \Sigma_t \).

4) Given \( S_{t+1} \), the probability of selecting treatment, \( A_{t+1} = a \), is given by the posterior probability that treatment \( a \) has the highest expected reward.
Challenges in Mobile Health

1) **Noisy data**

Our solution: Bandit algorithm

- Bandit algorithms learn faster than full RL algorithms
- The bandit acts as a regularizer (discount rate is 0): trade speed of learning (reduced variance) with bias
- Use a low dimensional parameterization of the expected reward: linear model in context and treatment. $E[R_{t+1}|S_t, A_t] = \eta^T f(S_t, A_t)$
- Use a Gaussian prior on $\eta$ with mean, variance based on the data from Heartsteps V1 and a baseline no-treatment week of data from Heartsteps V2
Challenges

2) Nonstationarity: Over longer periods of time, the expected reward function will likely change.
   - Due to inability to fully sense, known and unknown, aspects of user’s current context.

Our solution:

- Promote continual exploration: Use a Gaussian process prior in the model for the expected reward, e.g.
  \[ E[R_{t+1}|S_t, A_t] = f(S_t, A_t)^T \eta_t \text{ where } \eta_{t+1} = \mu_0 + \gamma(\eta_t - \mu_0) + \varepsilon_t \]
  \[ \varepsilon_t \sim N(0, 1 - \gamma^2) \]
Challenges

3) The immediate effects are primarily positive; the delayed effects are primarily negative. →
   - Algorithm may falsely learn that “always treat” is best, yet there are better policies.

Our Solution
- Add a low variance proxy for the “value function” to the current reward
Challenges

4) Need to ensure ability to conduct “off-policy learning” and causal inference after bandit study

Our solution:

• Use explicit randomization to explore: Thompson Sampling Bandit
• Ensure the no-treatment selection probability lies in an interval bounded away from 0 and 1; here [.2, .9]
Challenges

5) Expected reward, $E[R_{t+1} | S_t, A_t]$ is likely a complex function of context, $S_t$

Proposed Solution: Center the treatment indicator by binary treatment selection probability, $\pi_t$
Proposal: For binary $A_t$:
replace $E[R_{t+1}|S_t = s, A_t = a] = \eta^T f(s, a)$
with
$E[R_{t+1}|S_t = s, A_t = a] = b_t(s) + \eta^T f(s)(a - \pi_t)$
where
$b_t(s)$ is an unspecified baseline (maybe nonlinear, non-stationary)

$\eta^T f(s)(a - \pi_t)$ is centered since $\pi_t$ is the probability of selecting treatment $A_t = a = 1$

In the Thompson-Sampling update of expected reward use a working (but likely mis-specified) approximation for $b_t(s)$. 
Median Regret
500 simulated users

Quartiles of Regret
500 simulated users

Context, $s$, is 3 dimensional
True $b_t(s)$ is nonlinear
Linear working model for $b_t(s)$

No proxy value
Only Gaussian prior
No Gaussian process prior
Discussion

Challenges:

– Online accommodation/use of missing data
– High between user variance in performance of online algorithms

Randomization assists in forming causal inferences based on minimal structural assumptions in after study data analyses

– The bandit algorithm is one way to conduct randomization
– Randomization can also be based on forecasts or predictions
Collaborators!

samurphy@fas.harvard.edu
Notes

Context $s_t = [I, s_{t2}, s_{t3}, s_{t4}]^T$

nonlinear generative model

$$r_t = I(a_t = 1)\theta_1^T s_t + I(a_t = 2)\theta_2^T s_t + [1.32, 1.26, .58][1, [s_{t3}], [s_{t4}]]^T + 2I([s_{t2}] \leq 0.8)$$

Analysis model used for action centering.

$$r_t = (I(a_t > 0) - \pi_t)[I(\bar{a}_t = 1)\theta_1^T s_t + I(\bar{a}_t = 2)\theta_2^T s_t] + \pi_t[I(\bar{a}_t = 1)\zeta_1^T s_t + I(\bar{a}_t = 2)\zeta_2^T s_t] + \eta^T s_t$$

Total params: 20.

Analysis model for std Thompson.

$$r_t = I(a_t = 1)\theta_1^T s_t + I(a_t = 2)\theta_2^T s_t + \eta^T s_t$$

Total params: 12.

Theta1 = [.116, -.275, -.233, .0425]

Theta2 = [.116, .275, -.233, .0425].
Long Term Goal: Continually learning mHealth App

The learning algorithm is part of the mHealth app

– Incorporate continual learning in the implementation of a mHealth application.
– Learning algorithm makes structural assumptions so as to trade bias and variance in learning
Challenges to RL

• State is large yet partially observed: “unknown-unknowns”
  – Non-stationary reward
• Treatment actions that tend to have positive effects on immediate rewards but negative impact on future rewards via user habituation/burden.
• High noise within/between user
• Clinical populations (e.g., small numbers of users)
• Off-policy causal inference to further develop behavioral science
Micro-randomized Trial

• Micro-randomized trial = combination of factorial experimental design with explicitly controlled exploration
• Exploration via use of online forecasting (and RL algorithms)
• Multiple treatment factors occurring at different time scales and which target different rewards
• Probabilistic budgets on # of treatment pushes to manage habituation/burden
• Off-policy, after study is over, causal inference
BariFit MRT to Promote Weight Maintenance Among People Who Received Bariatric Surgery

Prior to the study
Each participant randomized 2x at baseline

Set Goals
- Use 60th percentile of daily step counts over 10 prior days as goal

Set Goals
- Use variable percentiles of daily step counts over 10 prior days as goal

Daily Goals
- Rest (no goal) 1 of every 7 days on average

Daily Goals
- No rest: receive a goal every day of the study

Each day of study

Observations
- Step count (sensor)
- Whether participant tracks food (sensor)
- Interaction with app (sensor)
- Weight (self report)
- Food intake (self report)

Start Intervention
- Tailored text message to become physically active

Proximal Outcome
- Physical activity (steps taken)

Following day

Distal Outcome
- Growth in step count

30 minutes after prompt

Proximal Outcome
- Weather
- Time of day
- Day of week

Start Intervention
- Remind participant to track food

Average 1.5x/day

Average 3.5x/day

Average 1x everyday

Average 1x every other day

R

R

PI: P Klasnja
Location & Funding: Kaiser Permanente
Sense$^2$Stop

Sense$^2$Stop MRT for Stress Management in Newly Abstinent Smokers

Every minute of every day starting with quit date

For two hours after intervention is delivered

Measured via EMA and puffMarker over 10 days

Observations
- stress (via AutoSense sensor suits)
- motion (via accelerometer)
- smoking (via self-report)

Available? NO

is stressed? NO

Available? YES

is stressed? YES

Remainder of times

Prompt use of stress-management exercises

Remainder of times

No intervention

Average 1.5x/day

Proximal Outcome
Probability of stress episode

Distal Outcome
Release or smoking abstinence

PI: S Kumar

Location: Northwestern University, B. Spring, (P.I.)

Funding: NIBIB through funds provided by the trans-NIH Big Data to Knowledge initiative U54EB020404
SARA

Data Collection MRT to Promote Engagement in Substance Use Research

Each day of study
This chart represents two of four engagement MRTs embedded in the SARA app

Observations
- location via (GPS)
- motion (via accelerometer)
- number of surveys completed

After 6 pm

Did user complete evening survey?

YES

R

Engagement Intervention
Reward message for completing survey

50% 

NO

Stop

50%

No Intervention

1/x day between 12 pm and 6 pm

No Intervention

50%

R

Engagement intervention
Youth-targeted message encouraging completion of assessment later in the day

50%

The day after the engagement intervention

Proximal Outcome
survey completed

Distal Outcome
Completion rate of surveys during study

Day 30

The evening of the engagement intervention

Proximal Outcome
survey completed

PIs: M Walton, S Murphy, and M Rabbi Shuvo
Location: University of Michigan
Funding: Michigan Institute for Data Science (PI S. Murphy), University of Michigan Injury Center (PI M. Walton)
Engagement with JOOL

MRT to Promote Engagement with Purpose-driven Well-being App

Each day of study

Within 24 hours of push notification

Future

Observations
- activity (via accelerometer)
- surveys (via app)

Select time of day:
1/3 chance for each on weekdays
1/5 chance on weekends
8:30 am (weekday only)
12:30 pm
5:30 pm
6:30 pm
7:30 pm or
8:30 pm

Received notification recently?
YES

Stop

NO

Stop

Stop

50%

50%

Push notification
Tailored health message to encourage engagement with the app

Proximal Outcome
Engaged app

Distal Outcome
More consistent self monitoring

PI: Victor Strecher, PhD, MPH, CEO of JOOL Health
Location & Funding: Ann Arbor, MI
URL: https://www.joolhealth.com