CDT Summer School

- Find Github
- Find your breakout group

- Designing Oralytics: RL for Real Life
  - Lecture 30 min

- Breakout in groups + Discussion (20 min)

1

https://github.com/StatisticalReinforcementLearningLab/Stat-ML-CDT-2023/tree/main

Lecture 4: 3 hours 15 min, 13.45 to 17:00

CDT Summer School

- Practicum for data analyses when the RL algorithm pools data across users in order to select actions
  - Overview Lecture (15 min)
  - Coding in breakout groups (30 min)
  - Lecture (15 min)
  - Coding and discussion in breakout groups (30 min)

- Break (10 min)

CDT Summer School

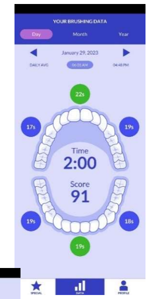
- Secondary analyses tools to help with individualized inference
  - Lecture (35 min)
  - Discussion (10 min)

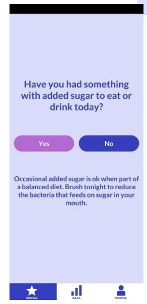
Oralytics:
RL for Real Life!

Susan A Murphy

Kelly Zhang

Raaz Dwivedi





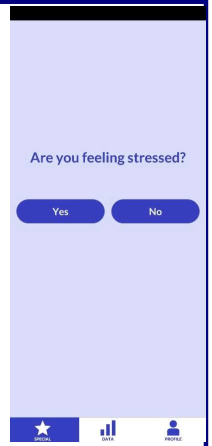
optimizing adaptive interventions, by design

To Think About!

- Think about warm-starting an RL algorithm
- Think about how to construct the RL algorithm's reward
- Think about how to monitor the RL algorithm

5

Oralytics Re-Cap



Oralytics V1, 10-week trial in Aug 2023 involving $n \approx 70$ users:

- App on smartphone + Bluetooth enabled toothbrush
- "Micro-randomization" via a reinforcement learning (RL) algorithm

Oralytics V2, 20-week, randomized-control trial (RCT) involving $n \approx 260$ users, early 2025

6

Both within study learning and between study learning
Need to be able to transfer to a potentially different population….

V2 will involve two groups. One with RL and one without RL…

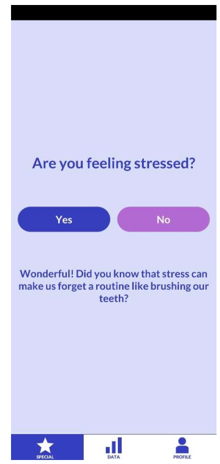
Oralytics Re-Cap

- Decision Times, t: 2 time points per day (prior to the individual's usual brushing time)

- State, S_t: app engagement, prior brushing quality, time of day, weekend, prior # messages,…

- Action, A_t: An engagement message (deliver or not deliver)

- Reward, R_{t+1}: brushing quality score



Brushing Quality =min(180, brushing duration-over pressure)
The RL alg will use a surrogate reward –this talk.
ONLY 140 decision times per user over 70 day study

Oralytics Re-Cap: Linear Thompson-Sampling

- Learning Algorithm: Bayesian Algorithm
  - Inference in parameters in a linear model for $r(s,a) = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$
- Optimization Algorithm: Posterior Sampling
  - $\pi_t^{\mathcal{L}}(\cdot\,|s)$ is usually the posterior probability that
    $$r(s,1) - r(s,0) > 0$$
  - $A_t \sim \pi_t^{\mathcal{L}}(\cdot\,|S_t; H_{t-1})$

8

Features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (=1 if the user has the app open and in focus (i.e. not in background

Oralytics: Linear Thompson-Sampling

- Learning Algorithm: Bayesian Algorithm
  - Inference in parameters in a linear model for $r(s, a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
  - Model for $r(s, a) = \alpha^T f(s) + \beta^T f(s) a$

$f(S_t) = (1, D_t, \bar{B}_t, \bar{A}_t, E_t)$ 9

Think about the use of the linear model for $r(s, a)$

Definitely the wrong model, why?

$f(S_t)$ Features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (=1 if the user has the app open and in focus (i.e. not in background

SA Murphy

Oralytics: Linear Thompson-Sampling

- Optimization Algorithm: Posterior Sampling

  - $\pi_t^{\mathcal{L}}(\cdot\,|s)$ is usually the posterior expectation
    $$\mathrm{E}[1\{\beta^T f(s) > 0\}|H_{t-1}]$$

  - Instead of $1\{x > 0\}$ Oralytics uses a smooth allocation function, $\rho(x)$

$x = \beta^T f(s)$

10

$r(s,1) - r(s,0) = \beta^T f(s)$

H_{t-1} is all data from all individuals collected prior to the Sunday night before this user's decision time t

$\rho(x)$ is a generalized logistic function $\rho(x) = L\_\{\backslash min\} + \backslash frac\{ L\_\{\backslash max\}$ - $L\_\{\backslash min\} \}\{ \backslash big[1 + c \backslash exp(\text{-}b\ x) \backslash big]\}$

$L\_\{\backslash min\}$=0.2; $L_{\{\backslash max\}} = 0.8$

C=5, b=0.515

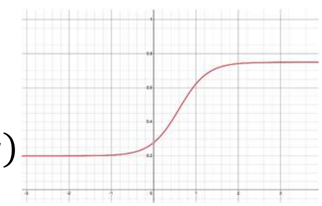
Oralytics: Linear Thompson-Sampling

- Optimization Algorithm: Posterior Sampling
  - Instead of $1\{x > 0\}$ Oralytics uses a smooth allocation function, $\rho(x)$
  - $\pi_t^{\mathcal{L}}(\cdot\,|s)$ the posterior expectation

$$E\left[\rho\left(\beta^T f(s)\right)|H_{t-1}\right]$$

  - $A_t \sim \pi_t^{\mathcal{L}}(\cdot\,|S_t; H_{t-1})$

$\rho(x)$



x

11

H_{t-1} is all data from all individuals collected prior to the Sunday night before this user's decision time t

$r(s,1) - r(s,0) = \beta^T f(s)$

Warm-Start for RL Algorithm

12

RL Algorithm Warm-Start

Role of Prior in Bayesian online RL algorithm

1. Prior distribution on parameters, α, β, can incorporate subjective knowledge based on scientific expertise, previous data
   – Model: $r(s,a) = \alpha^T f(s) + \beta^T f(s)a$

   – When data is sparse, there is a shrinkage of inference to subjective prior (posterior distribution is close to prior distribution).

13

Shrinkage is critical when data is noisy/sparse and trades bias with variance

$f(S_t)$ Features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (=1 if the user has the app open and in focus (i.e. not in background

Think about the use of the linear model for $r(s,a)$

Definitely the wrong model, why?

RL Algorithm Warm-Start

Role of Prior in Bayesian online RL algorithm

2. Prior distribution on parameters, α, β, acts as a warm-start for the online optimization algorithm
   – If this prior distribution is "centered" at the true parameters in the linear model for $r(s, a)$, then optimal actions will be more likely to be sampled early in trial.

14

Transfer learning….
Reduce disengagement by users

RL Algorithm Warm-Start

Role of Prior in Bayesian online RL algorithm

3. Gaussian prior distribution on parameters, α, β, acts as an L_2 regularizer
   – Stabilizes computations when data is sparse

- Rebuild prior between trials to warm-start the RL alg

Transfer learning….
Reduce disengagement by users

Constructing the Prior

Rules of thumb:

- Use existing MRT data from 9 pilot users.
  - MRT deploys the same actions and collects the same state data
- Fit the RL alg's model to *each* user in the MRT data.

16

For each pilot user $i \in [1 : 9]$, we fit a linear model with action-centering for the reward given state and action. Notice that to prevent numerical instability, we fit each model using L2 regularization with $\lambda = 10^{-3}$. The linear model with action-centering contains 15 parameters.

Constructing the Prior

Rules of thumb:

- Use existing MRT data from 9 pilot users.
  - MRT deploys the same actions and collects the same state data
- Fit the RL alg's model to *each* user in the MRT data.

Linear model: $r(s,a) = \alpha^T f(s) + \beta^T f(s)a$

17

Think about the use of the linear model for $r(s,a)$

Definitely misspecified……

Advantage is defined as $r(s,1) - r(s,0) = \beta^T f(s)$

10 regression coefficients including intercept (15 if you use action-centering)

features are:
1. D_t: Time of Day (Morning/Evening) $\in \{0, 1\}$
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) $\in$ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) $\in [-1, 1]$
4. E_t: Prior Day App Engagement $\in \{0, 1\}$ (if the user has the app open and in focus (i.e. not in background

Constructing the Prior

- Calculate the mean and variance across users of the estimated α, β parameters.
  - Do EDA plots
  - Discuss with scientific team
- Decide what statistics constitute evidence that a parameter is likely not close to the null value (i.e. 0)
  - Only 9 pilot users

18

In prior sessions I called these parameters, weights.

10 regression coefficients including intercept

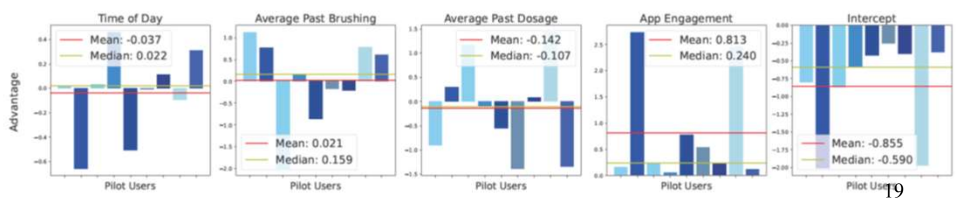
features are:
1. D_t: Time of Day (Morning/Evening) $\in \{0, 1\}$
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) $\in$ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) $\in [-1, 1]$
4. E_t: Prior Day App Engagement $\in \{0, 1\}$ (if the user has the app open and in focus (i.e. not in background

Discount rate=γ =13/14 $\quad {}^{1}/_{1-\gamma} = 14$ (one week) --used to form $\bar{B}_{i,t}, \bar{A}_{i,t}$

Constructing the Prior

- Decide what statistics constitute evidence that a parameter is likely not close to the null value (i.e. 0)
  – Oralytics parameters are α, β in $r(s,a) = \alpha^T f(s) + \beta^T f(s)a$

- Only 9 pilot users, focus on β:
  – Statistics are standardized effect sizes, e.g. ${}^{\beta_i}/_{\sigma_i}$



19

σ_i is the ith user's reward variance

If average ${}^{\beta_i}/_{\sigma_i}$ across users has a modulus greater than 0.15 we considered this feature "significant"

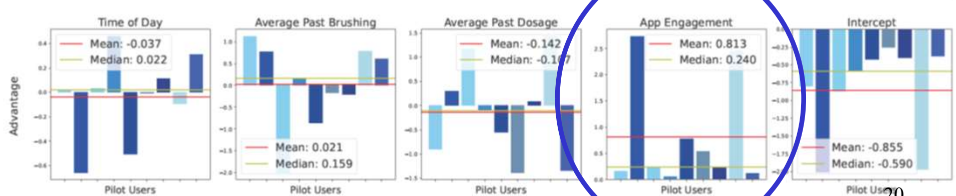
In other studies (HeartSteps) we have many more pilot users so we used standard statistics..

features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (if the user has the app open and in focus (i.e. not in background

Discount rate=γ =13/14 $\;{}^{1}/_{1-\gamma} = 14$ (one week) --used to form $\bar{B}_{i,t}, \bar{A}_{i,t}$

Constructing the Prior

- If the average across the 9 standardized effect sizes, $\frac{1}{9}\sum_{i=1}^{9} {}^{\beta_i}/_{\sigma_i} \geq 0.15$, then some "evidence" against the null value (i.e., 0)
  - Construct an informative, subjective prior for these parameters



20

σ_i is the ith user's reward variance. If average ${}^{\beta_i}/_{\sigma_i}$ across users has a modulus greater than 0.15 we considered this feature "significant"

In other studies (HeartSteps) we have many more pilot users so we used standard statistics..

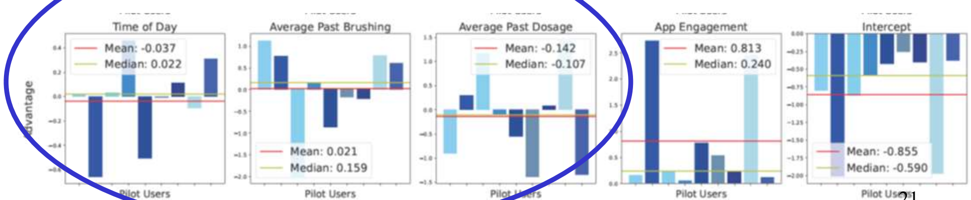
Setting Prior Means and Prior Variances For significant features, we set the prior mean to the empirical mean parameter value for that feature across 9 users. For significant features, we set the prior SD to the empirical SD for that feature across 9 users.

features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (if the user has the app open and in focus (i.e. not in background

Discount rate=γ =13/14 ${}^{1}/_{1-\gamma} = 14$ (one week) --used to form $\bar{B}_{i,t}, \bar{A}_{i,t}$

Constructing the Prior

- If the average across the 9 standardized effect sizes, $\frac{1}{9}\sum_{i=1}^{9} {}^{\beta_i}/_{\sigma_i} < 0.15$, then little "evidence" against the null value (i.e., 0)
  - Construct a "weakly informative" prior for remaining parameters



21

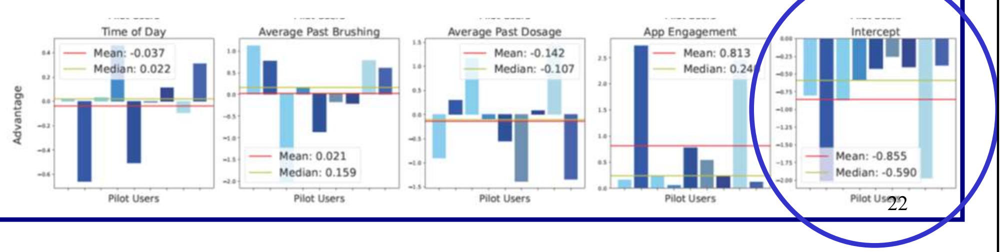
Setting Prior Means and Prior Variances. For non-significant parameters, we set the prior mean to be 0. For non-significant parameters, we set the prior SD to the empirical SD divided by 2. Notice that we are reducing the SD of the non-significant weights because we want to provide more shrinkage to the prior mean of 0. (i.e., more data is needed to overcome the prior). However the reduction value of 2 was an arbitrary choice.

features are:
1. D_t: Time of Day (Morning/Evening) ∈ {0, 1}
2. $\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days (Normalized) ∈ R
3. $\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days (Normalized) ∈ [−1, 1]
4. E_t: Prior Day App Engagement ∈ {0, 1} (if the user has the app open and in focus (i.e. not in background

Discount rate=γ =13/14 $\quad {}^{1}/_{1-\gamma} = 14$ (one week) --used to form $\bar{B}_{i,t}, \bar{A}_{i,t}$ prior to normalization

Constructing the Prior

- Average across the 9 standardized effect sizes, $\frac{1}{9}\sum_{i=1}^{9} \beta_i/\sigma_i < -0.15$, and contradicts the domain science……
  – Construct a "weakly informative" prior for remaining parameters



22

Notice that the calculated standard effect size of the intercept in the advantage has an average magnitude greater than the threshold of 0.15. This is because scientifically the intervention messages should either not affect or should improve brushing quality (the intercept in the advantage should be nonnegative) and thus our team decided to declare this intercept feature insignificant. For non-significant parameters, we set the prior SD to the empirical SD divided by 2 (Table 2). Notice that we are reducing the SD of the non-significant weights because we want to provide more shrinkage to the prior mean of 0. (i.e., more data is needed to overcome the prior). However the reduction value of 2 was an arbitrary choice.

After using these guidelines, we determined "Time of Day", "Average Past Dosage" and "Intercept" to be significant for the baseline and "App Engagement" to be significant for the advantage

Table 1: Finalized Prior Using Oralytics Pilot Data. Values are rounded to the nearest integer. The ordering of the β features in $r(s,a) = \alpha^T f(s) + \beta^T f(s)a$: :

Time of Day, Exponential Average of Brushing Over Past 7 Days (Normalized), Exponential Average of

Messages Sent Over Past 7 Days, Prior Day App Engagement, Intercept Term.

$\mu\alpha 0$: prior mean on the baseline state features [18, 0, 30, 0, 73]
$\Sigma\alpha 0$: prior variance on the baseline state features diag(732, 252, 952, 272, 832)
$\mu\beta$: prior mean on the advantage state features [0, 0, 0, 53, 0]
$\Sigma\beta$: prior variance on the advantage state features diag(122, 332, 352, 562, 172)

Reward for RL Algorithm

23

Surrogate Reward

- Proximal Health Outcome is Brushing Quality
  - Q_t= min(180, brushing duration minus over-pressure duration) for user at time t

- Why consider a surrogate reward for the RL alg?
  - Negative delayed effects of sending an engagement message leading to disengagement…..
    - Habituation
    - Treatment Burden
  - Model misspecification

24

Why surrogate reward? Delayed effects, model misspecification

Habituation occurs under repeated stimuli --eventually you don't even perceive the stimuli. Think about traffic or trains near your home and how eventually you don't even notice the noise.

Treatment burden due to intervening as someone goes about their daily life…. People don't like getting pinged by their phone all the time.

Surrogate Reward

- Proximal Health Outcome is Brushing Quality
  - Q_t= min(180, brushing duration minus over-pressure duration) for user at time t

- Surrogate Reward
  - $R_{t+1} = Q_t - A_t C_t$

- C_t is a "proxy" for ?
  - Think about the target in MDPs….

25

Surrogate Reward

- Surrogate Reward
  - $R_{t+1} = Q_t - A_t C_t$

- C_t is a "proxy" for

$E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 0\right] - E[V^{\pi^*}(S_{t+1})|S_t, A_t = 1]$

where π^* is optimal policy.

- Why?

26

Q_t=Brushing Quality =min(180, brushing duration-over pressure)

Best target is $Q_t + V^{\pi^*}(S_{t+1})$

Note that $E\left[Q_t + V^{\pi^*}(S_{t+1})\middle|S_t, A_t\right] = E[Q_t|S_t, A_t] +$
$A_t\left(E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 1\right] - E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 0\right]\right) +$
$E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 0\right]$

Thus $-A_t C_t = A_t\left(E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 1\right] - E\left[V^{\pi^*}(S_{t+1})\middle|S_t, A_t = 0\right]\right)$

Surrogate Reward

- Proximal Health Outcome is Brushing Quality
  - Q_t= min(180, brushing duration minus over-pressure duration) for a user at time t

- Surrogate Reward
  - $R_{t+1} = Q_{i,t} - A_t C_t$

- RL alg uses Surrogate Reward
- Use Q_t to evaluate RL alg

27

Why surrogate reward? Delayed effects, model misspecification

Surrogate Reward

- Surrogate reward
  - $R_{t+1} = Q_t - A_t C_t$
  - $C_t = \xi_1 1_{\{\bar{B}_t > 111\}} 1_{\{\bar{A}_t > 0.5\}} + \xi_2 1_{\{\bar{A}_t > 0.8\}}$
  - $\bar{B}_t$ is exponentially discounted brushing quality over prior week
  - $\bar{A}_t$ is exponentially discounted number of messages over prior week.
- Tune ξ_1, ξ_2 using ROBAS3 based simulation testbed

28

111, is the 50th-percentile of user brushing durations in ROBAS 2,

- 0:5, represents a rough approximation of the user getting a message 50% of the time (rough approximation

because we are using an exponential average mean)

- 0:8, represents a rough approximation of the user getting a message 80% of the time (rough approximation

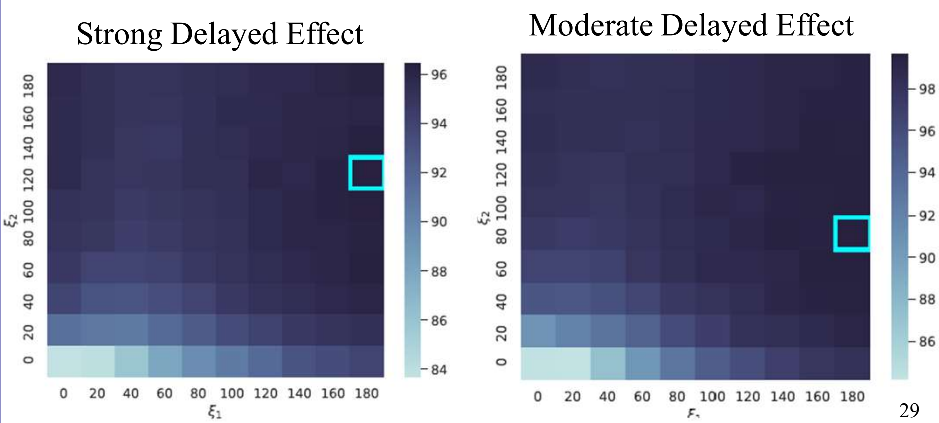
because we are using an exponential average mean)

$\bar{B}_t$: Exponential Average of Brushing Over Past 7 Days
$\bar{A}_t$: Exponential Average of Messages Sent Over Past 7 Days

Discount rate=γ =13/14 $^1/_{1-\gamma} = 14$ (1 week) --used to form $\bar{B}_{i,t}, \bar{A}_{i,t}$ (not normalized)

Tune Surrogate Reward

Average of Users' Average Brushing Quality
(dark is better)



29

Heavily penalizes for messages if user is doing well

$C_{i,t} = \xi_1 1_{\{\bar{B}_{i,t}>111\}} 1_{\{\bar{A}_{i,t}>0.5\}} + \xi_2 1_{\{\bar{A}_{i,t}>0.8\}}$

This is results from tuning the surrogate reward for the pilot study.

Simulation Env. Is ROBAS 3. Informative prior on RL alg is ROBAS 2

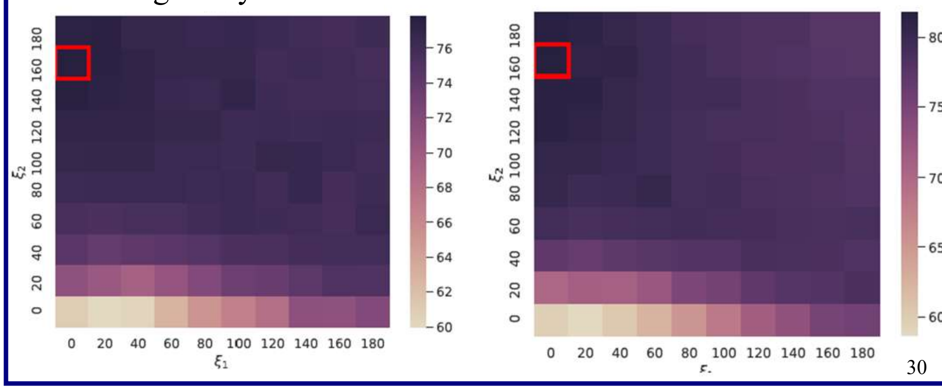
100 monte carlo trials 70 users

T=140 decision times

Tune Surrogate Reward

25th Percentile of Users' Average Brushing Quality
(dark is better)

Strong Delayed Effect　　Moderate Delayed Effect



30

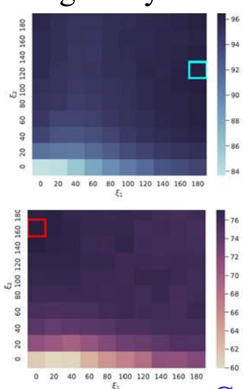
$$C_{i,t} = \xi_1 1_{\{\bar{B}_{i,t}>111\}} 1_{\{\bar{A}_{i,t}>0.5\}} + \xi_2 1_{\{\bar{A}_{i,t}>0.8\}}$$

This is results from tuning the surrogate reward for the pilot study.
Simulation Env. Is ROBAS 3. Informative prior on RL alg is ROBAS 2
100 monte carlo trials

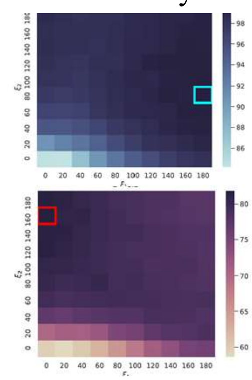
Tune Surrogate Reward

Users Brushing Quality

Strong Delayed Effect



Moderate Delayed Effect



Select $\xi_1 = \xi_2 = 100$

31

$$C_{i,t} = \xi_1 1_{\{\bar{B}_{i,t}>111\}} 1_{\{\bar{A}_{i,t}>0.5\}} + \xi_2 1_{\{\bar{A}_{i,t}>0.8\}}$$

This is results from tuning the surrogate reward for the pilot study.

Simulation Env. Is ROBAS 3. Informative prior on RL alg is ROBAS 2

100 monte carlo trials

20 min for Discussion!

Break & Discussion

- How might you use statistical methods to monitor the online RL algorithm?
  - In a clinical trial in which the *entire intervention* must be pre-specified.
  - In implementation by a health care system or insurance company?
- Are there analyses you might do between trials to check if the online RL algorithm is learning?