# Supplementary Material for "Reward Design For An Online Reinforcement Learning Algorithm For Supporting Oral Self-Care"

Anna L. Trella, Kelly W. Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, Susan A. Murphy

August 10, 2022

## 1 Data and Code

We use two data sets from previous studies: ROBAS 2 [Shetty et al., 2020] and ROBAS 3. The data sets are publically available here for ROBAS 2 and here for ROBAS 3. Code for this paper can be found in a github repository here.

## 2 Simulation Environments

### 2.1 Defining the Target: Brushing Quality $Q_{i,t}$

Brushing quality is the true target that the scientific team cares about in achieving healthy brushing behaviors (desired behavioral health outcome). We define a proxy for brushing quality as $Q_{i,t} = \min(B_{i,t} - P_{i,t}, 180)$. $B_{i,t}$ is the participant's brushing duration in seconds and $P_{i,t}$ is the aggregated duration of over pressure in seconds. $Q_{i,t}$ is truncated to avoid optimizing for over-brushing.

Other sensory information that were considered was zoned brushing duration, as another indicator of brushing quality is brushing more evenly across the six zones. However, zoned brushing duration is not reliably obtainable as it requires Bluetooth connectivity and the participant to stand close enough to the docking station. Therefore, zoned brushing duration only appeared in about 82% of brushing sessions in the pilot study. $Q_{i,t}$ can be interpreted as a non-penalized reward, and the true target capturing the health outcome that the scientific team's wants to maximize.

### 2.2 Baseline Feature Space of the Environment Base Models

We form the following features using domain expert knowledge from behavioral health and dentistry. In this section we discuss how we fit a model for the

baseline reward (i.e., the brushing quality under action $A_{i,t} = 0$). A discussion on how we model brushing quality under action $A_{i,t} = 1$ is in Appendix 2.6.

$g(S_{i,t}) \in \mathbb{R}^6$ denotes the feature space used to fit a model of the baseline reward which is the following:

1. Bias / Intercept Term $\in \mathbb{R}$
2. Time of Day (Morning/Evening) $\in \{0, 1\}$
3. Prior Day Total Brushing Quality (Normalized) $\in \mathbb{R}$
4. Weekend Indicator (Weekday/Weekend) $\in \{0, 1\}$
5. Proportion of Non-zero Brushing Sessions Over Past 7 Days $\in [0, 1]$
6. Day in Study (Normalized) $\in [-1, 1]$

### 2.2.1 Normalization of State Features

We normalize features to ensure that all state features are within a similar range. The Prior Day Total Brushing Quality feature is normalized using z-score normalization (subtract mean and divide by standard deviation). Since each participant in the ROBAS 3 study had varying participant study lengths (due to dropping out), the Day in Study feature (originally in the range $[1 : T_i]$ where $T_i$ represents the number of days user $i$ was in the study) is normalized to be between $[-1, 1]$. Note that when generating rewards, Day in Study is normalized based on Oralytic's anticipated 70 day study duration (range is still $[-1, 1]$).

$$\text{Normalized Total Brushing Quality in Seconds} = (\text{Brushing Quality} - 154)/163$$

$$\text{Normalized Day in Study When Fitting Model for User } i = \left(\text{Day} - \frac{T_i + 1}{2}\right) / \frac{T_i - 1}{2}$$

$$\text{Normalized Day in Study When Generating Rewards} = (\text{Day} - 35.5)/34.5$$

## 2.3 Environment Base Model

Due to the zero-inflated nature of the ROBAS 3 data set, we use a zero-inflated Poisson model to generate brushing quality. $w_{i,b}, w_{i,p}$ are user-specific weight vectors, $g(S_{i,t})$ is the baseline feature vector of the current state defined in Appendix 2.2, and $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

#### 2.3.1 Zero-Inflated Poisson Model for Brushing Quality

To generate brushing quality for user $i$ at decision time $t$, we use the following model:

$$Z \sim \text{Bernoulli}\left(1 - \text{sigmoid}(g(S_{i,t})^T w_{i,b})\right)$$

$$Y \sim \text{Poisson}\left(\exp\left(g(S_{i,t})^T w_{i,p}\right)\right)$$

$$\text{Brushing Quality} : Q_{i,t} = ZY$$

### 2.4 Fitting the Environment Base Model

We use ROBAS 3 data to fit the brushing quality model under no intervention ($A_{i,t} = 0$). We fit one model per user and each user model was fit using MAP with a prior $w_{i,b}, w_{i,p} \sim \mathcal{N}(0, I)$ as a form of regularization because we have sparse data for each user. We jointly fit parameters for both the Bernoulli and the Poisson components. Weights were chosen by running random restarts and selecting the weights with the highest log posterior density.

### 2.5 Checking the Quality of the Environment Base Model

#### 2.5.1 Checking Moments

[ALT: ANNA TODO]

### 2.6 Imputing Treatment Effect Sizes for Simulation Environments

Recall that the ROBAS 3 data set does not have data under interventions (sending a message). Therefore we impute effect sizes to model the reward under action 1.

#### 2.6.1 Treatment Effect Feature Space

The treatment effect (advantage) feature space was made after discussion with domain experts on which features are most likely to interact with the intervention (action). $h(S_{i,t}) \in \mathbb{R}^5$ denotes the features space used for the treatment effect which is the following:

1. Bias/Intercept Term $\in \mathbb{R}$
2. Time of Day (Morning/Evening) $\in \{0, 1\}$
3. Prior Day Total Brushing Quality (Normalized) $\in \mathbb{R}$
4. Weekend Indicator (Weekday/Weekend) $\in \{0, 1\}$
5. Day in Study (Normalized) $\in \mathbb{R}$

### 2.6.2 Environment Model Including Effect Sizes

We impute treatment effects on both the Bernoulli component which represents the user's intent to brush and the Poisson component which represents the user's brushing quality when they intend to brush. After incorporating treatement effects, brushing quality $Q_{i,t}$ under action $A_{i,t}$ in state $S_{i,t}$ is:

$$Z \sim \text{Bernoulli}(1 - \text{sigmoid}(g(S_{i,t})^T w_{i,b} + A_{i,t} * h(S_{i,t})^T \Delta_{i,B}))$$
$$Y \sim \text{Poisson}(\exp(g(S_{i,t})^T w_{i,p} + A_{i,t} * h(S_{i,t})^T \Delta_{i,N}))$$
$$Q_{i,t} = ZY$$

$\Delta_{i,B}, \Delta_{i,N}$ are user-specific effect sizes. $g(S_{i,t})$ is the baseline feature vector as described in Appendix 2.2, and $h(S_{i,t})$ is the feature vector that interacts with the effect size specified above.

### 2.6.3 Procedure For Imputing Effect Sizes

We consider a unique realistic effect size for each user. We first construct a population level effect size for the Bernoulli and Poisson components, $\Delta_B, \Delta_N$ respectively. We then use $\Delta_B, \Delta_N$ to sample user-specific effect sizes $\Delta_{i,B}, \Delta_{i,N}$.

Recall that for the environment base model, we fit a user-specific model for the brushing quality and obtained user-specific parameters $w_{i,b}, w_{i,p} \in \mathbb{R}^6$ (values of the fitted parameters can be found here). For mobile health interventions, we expect the treatment effect to be on the order of or smaller than the effect of the baseline features. Therefore we use the fitted parameters to form the population effect sizes as follows:

Zero-Inflated Models' Effect Sizes:

- $\Delta_B = \mu_{B,\text{avg}}$ where $\mu_{B,\text{avg}} = \frac{1}{5} \sum_{d \in [2:\,6]} \frac{1}{N} \sum_{i=1}^{N} |w_{i,b}^{(d)}|$.
- $\Delta_N = \mu_{N,\text{avg}}$ where $\mu_{N,\text{avg}} = \frac{1}{5} \sum_{d \in [2:\,6]} \frac{1}{N} \sum_{i=1}^{N} |w_{i,p}^{(d)}|$.

We use $w_{i,b}^{(d)}, w_{i,p}^{(d)}$ to denote the $d^{\text{th}}$ dimension of the vector $w_{i,b}, w_{i,p}$ respectively; we take the average over all dimensions excluding $d = 1$, which represents the weight for the bias/intercept term.

Now to construct the user-specific effect sizes, we draw effect sizes for each each user from a truncated normal distribution:

[ALT: ********QUESTION: so $\Delta_B$ has to be negative because the Bern draw has prob. $1 - \text{sigmoid}(BLAH)$. How can I do truncated Normal but only keep negative values?] [ALT: *******Right now we get lucky and we sample all negative values for $\Delta_{i,B}$ and all positive values for $\Delta_{i,N}$ just by sampling from a regular Normal but I don't know a good way of writing it.]

$$\Delta_{i,B} \sim \text{Truncated Normal}(\Delta_B, \sigma_B^2)$$

$$\Delta_{i,N} \sim \text{Truncated Normal}(\Delta_N, \sigma_N^2)$$

We truncate the effect size at 0 because it would not make sense in our setting to have a negative effect size. Having a negative effect size would mean that in the current context, not sending a message will yield a higher reward than sending a message. This is unreasonable because the only negative consequences of sending a message is the diminishing the responsivity to future messages. Therefore we constructed the effect sizes to be nonnegative.

$\sigma_B^2, \sigma_N^2$ are the empirical standard deviation (SD) over the average of the fitted parameters and are set by the following procedure:

- $\sigma_B$ is the empirical SD over $\{\mu_{i,B}\}_{i=1}^N$ where $\mu_{i,B} = \frac{1}{5}\sum_{d\in[2:\,6]} |w_{i,b}^{(d)}|$.
- $\sigma_N$ is the empirical SD over $\{\mu_{i,N}\}_{i=1}^N$ where $\mu_{i,N} = \frac{1}{5}\sum_{d\in[2:\,6]} |w_{i,p}^{(d)}|$.

After the procedure described above, we found $\Delta_B = -0.743, \Delta_N = 0.227$, $\sigma_B = 0.177$, and $\sigma_N = 0.109$ (values are rounded to the nearest 3 decimal places).

## 2.7 Simulating Delayed Effects

### 2.7.1 Simulating Unresponsiveness

We define unresponsiveness as shrinking the user's environment effect size by some scalar $E$ where $0 < E < 1$. At time $t$, we check if the unresponsiveness condition has been fulfilled. If so, then starting at time $t+1$, the effect sizes of the user environment $\Delta_{i,B}, \Delta_{i,N}$ will be shrunk by value $E$.

### 2.7.2 Simulating Disengagement

We define disengagement at time $t$ as the RL algorithm no longer obtaining rewards after time $t$ (i.e., $R_{i,t+1} = R_{i,t+2} = ... = R_{i,T} = 0$). To simulate disengagement, we consider a user-specific probability of disengagement $\pi^i_{\text{disengage}}$ sampled from a Beta distribution with mean equal to the population probability of disengagement $\mu_{\text{disengage}}$. To construct each $\pi^i_{\text{disengage}}$:

$$\begin{aligned} \pi^i_{\text{disengage}} &\sim \text{Beta}(a, b) \\ a &= \left(\frac{1-\mu_{\text{disengage}}}{\sigma^2_{\text{disengage}}} - \frac{1}{\mu_{\text{disengage}}}\right)\mu_{\text{disengage}}, \\ b &= a\left(\frac{1}{\mu_{\text{disengage}}} - 1\right) \end{aligned} \tag{1}$$

$\mu_{\text{disengage}}$ is a variant of the simulation environment and $\sigma_{\text{disengage}} = \frac{\sigma_B + \sigma_N}{2}$ (measure of between-user variance).

# 3 RL Algorithm

For the RL algorithm, we use a contextual bandit algorithm with Thompson sampling, a Bayesian Linear Regression reward function, and full pooling.

## 3.1 Feature Space of the RL Algorithm

$S_{i,t} \in \mathbb{R}^d$ represents the $i$th participant's state at decision time $t$, where $d$ is the number of features describing the participant's state.

**Advantage Feature Space** $f(S_{i,t}) \in \mathbb{R}^4$ denotes the feature space used to predict the advantage (i.e. the immediate treatment effect) which is the following:

1. Bias / Intercept Term $\in \mathbb{R}$
2. Time of Day (Morning/Evening) $\in \{0, 1\}$
3. Exponential Average of Brushing Over Past 7 Days (Normalized) $\in \mathbb{R}$
4. Exponential Average of Messages Sent Over Past 7 Days $\in [0, 1]$

The normalization procedure for Prior Day Brushing Quality is the same as the one described in Appendix 2.2.1. Features 3 and 4 are $\bar{B} = c_\gamma \sum_{j=1}^{14} \gamma^{j-1} B_{i,t-j}$ and $\bar{A} = c_\gamma \sum_{j=1}^{14} \gamma^{j-1} A_{i,t-j}$ respectively, the same $\bar{B}, \bar{A}$ used in the cost term of the reward.

**Baseline Feature Space** $m(S_{i,t}) \in \mathbb{R}^5$ denotes the feature space used to predict the baseline reward function which contains all the above covariates and the following:

5. Weekend Indicator (Weekday/Weekend) $\in \{0, 1\}$

The feature space used by the RL algorithm candidates is different than the feature space used to model the reward in the simulation environments in order to test the robustness of the RL algorithm despite having a misspecified reward model.

## 3.2 Bayesian Linear Regression with Action Centering

Recall that our model for the reward is a Bayesian Linear Regression (BLR) model with action centering. Since we consider an algorithm that does full pooling (clustering with cluster size $K = N$), the algorithm is shared between all users in the study and therefore shares parameters $\alpha_0, \alpha_1, \beta$.

$$R_{i,t} = m(S_{i,t})^T \alpha_0 + \pi_{i,t} f(S_{i,t})^T \alpha_1 + (A_{i,t} - \pi_{i,t}) f(S_{i,t})^T \beta + \epsilon \tag{2}$$

where $\pi_{i,t}$ is the probability that the RL algorithm selects action $A_{i,t} = 1$ in state $S_{i,t}$ for participant $i$ at decision time $t$. There are priors on $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\alpha_0 \sim \mathcal{N}(\mu_{\alpha_0}, \Sigma_{\alpha_0})$, $\alpha_1 \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$, $\beta \sim \mathcal{N}(\mu_\beta, \Sigma_\beta)$. We discuss how we set informative prior values for $\mu_{\alpha_0}, \Sigma_{\alpha_0}, \mu_\beta, \Sigma_\beta, \sigma^2$ using the ROBAS 2 data set in Appendix 3.3.

Let $\phi(S_{i,t}, A_{i,t}) = [m(S_{i,t}),\ \pi_{i,t} f(S_{i,t}),\ (A_{i,t} - \pi_{i,t}) f(S_{i,t})]^T$ be the joint feature vector and $\theta_i = [\alpha_0, \alpha_1, \beta]$ be the joint weight vector. Notice that

Equation 2 can be vectorized in the form: $R_{i,t} = \phi(S_{i,t}, A_{i,t})\theta_i + \epsilon$. [ALT: may need to change this notation if the vector contains data from multiple users] Let $\Phi_{i,1:t-1} \in \mathbb{R}^{Kx3}$ be the matrix of all stacked vectors $\{\phi(S_{i,s}, A_{i,s})\}_{s=1}^{t-1}$, and $\mathbf{R}_{i,1:t-1} \in \mathbb{R}^{K}$ be a vector of stacked rewards $\{R_{i,s}\}_{s=1}^{t-1}$, where we have batch data of the $t-1$ decision times before the current update time. Notice that $\Phi_{i,1:t-1}$ and $\mathbf{R}_{i,1:t-1}$ contains data from other users because we are performing full pooling.

## 3.3 Fitting the Prior for the Reward Function

We use the ROBAS 2 dataset to inform priors on $\alpha_0, \alpha_1, \beta$ as well as setting the noise variance term $\sigma^2$ in Equation 2. We follow the procedure as mentioned in [Liao et al., 2019]. Values are shown in Table 1.

| Parameter | Value |
|---|---|
| $\sigma^2$ | 3396.449 |
| $\mu_{\alpha_0}$ | $[0, 4.925, 0, 0, 82.209]^T$ |
| $\Sigma_{\alpha_0}$ | $\text{diag}(29.090^2, 30.186^2, 29.624^2, 12.989^2, 46.240^2)$ |
| $\mu_\beta$ | $[0, 0, 0, 0]^T$ |
| $\Sigma_\beta$ | $29.624^2 \cdot I_4$ |

Table 1: **Prior values for the RL algorithm informed using the ROBAS 2 data set.** Values are rounded to the nearest 3 decimal places. After performing Generalized Estimating Equations' (GEE) analysis, we found the Prior Day Total Brushing Duration feature and the bias term to be significant.

### 3.3.1 Fitting $\sigma^2$

$\sigma^2$ is set using ROBAS 2 and fixed for the entire study. To choose the value, we fit a separate Generalized Estimating Equations' (GEE) linear regression model per user and set $\sigma^2$ to be the average of the empirical variance of the residuals for each user model across the 32 user models.

### 3.3.2 Fitting $\mu_{\alpha_0}, \Sigma_{\alpha_0}, \mu_\beta, \Sigma_\beta$

To set values for $\mu_{\alpha_0}, \Sigma_{\alpha_0}$, we follow the procedure described in [Liao et al., 2019]. We do the following: 1) conduct GEE regression analyses using all participant's data in ROBAS 2 in order to assess the significance of each feature; 2a) for features that are significant in the GEE analysis, the prior mean is set to be the point estimate found from GEE analysis and the prior standard deviation is set to be the standard deviation across user models from GEE analysis; 2b) for features that are not significant, the prior mean is 0 and we shrink the standard deviation by half. For feature 3. Exponential Average of Brushing Over Past 7 Days, we imputed that value with the average of past brushing obtained so far for the first week. Recall that ROBAS 2 had no data under $A_{i,t} = 1$, so for feature 4. Exponential Average of Messages Sent Over Past 7 Days, we set the

prior mean to 0 and set the standard deviation to the average prior SD of the other features.

Notice that $\Sigma_{\alpha_0}$ is a diagonal matrix where the diagonal values are the prior variances described above. Notice that ROBAS 3 had a more sophisticated sensory suite than ROBAS 2, so while the scientific team chose brushing quality as the reward for the RL algorithm, we only have brushing duration from ROBAS 2. However, brushing duration is a reasonable guess for brushing quality, especially because average pressure duration is small. Therefore, when performing GEE analysis, the target is brushing duration.

Again, because ROBAS 2 only had no data under $A_{i,t} = 1$ so we cannot use the same procedure as described above to set $\mu_\beta, \Sigma_\beta$. Instead, we set $\mu_\beta = \mathbf{0}$ and set $\Sigma_\beta = \sigma_\beta^2 I$ (diagonal matrix where each entry on the diagonal is $\sigma_\beta^2$). $\sigma_\beta$ is equal to the average prior SD of the other features discussed above.

### 3.4 Posterior Updating

[ALT: ANNA TODO]

### 3.5 Action Selection

[ALT: ANNA TODO] [ALT: add in action selection prob. [0.35, 0.75]]

## 4 Experiments

### 4.1 Grid Search Values

We use bounds for $\xi_1, \xi_2$ equal to $[0, 120]$.

### 4.2 Simulating Study

[ALT: talk about the fitting b bar and a bar issue. what to do for the first week in the study?]

- for the first week we impute values for b bar and a bar to be the average of values we have obtained thus far (this includes the value of b bar used in the state and the reward)

## References

[Liao et al., 2019] Liao, P., Greenewald, K. H., Klasnja, P. V., and Murphy, S. A. (2019). Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *CoRR*, abs/1909.03539.

[Shetty et al., 2020] Shetty, V., Morrison, D., Belin, T., Hnat, T., and Kumar, S. (2020). A scalable system for passively monitoring oral health behaviors using electronic toothbrushes in the home setting: Development and feasibility study. *JMIR Mhealth Uhealth*, 8(6):e17347.