A Semiparametric Inverse Reinforcement Learning Approach to Characterize Decision Making for Mental Disorders

Xingche Guo [1] Donglin Zeng [2] Yuanjia Wang [1,3]

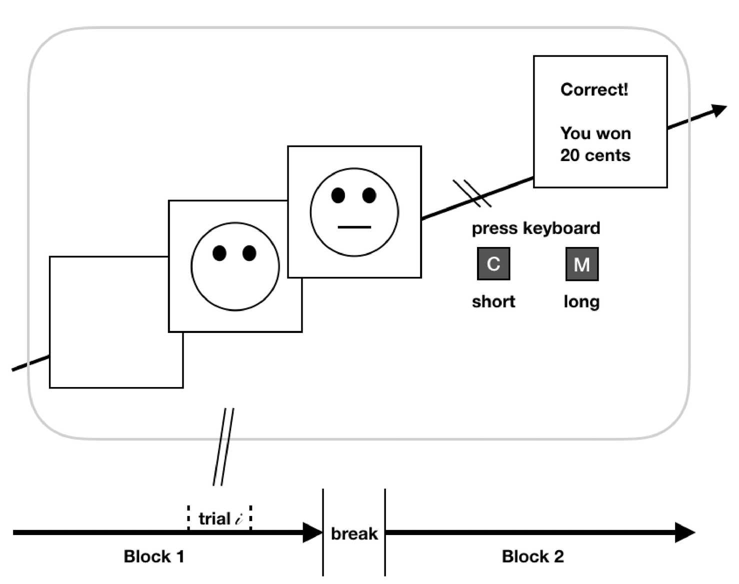
[1]Dept. of Biostatistics, Columbia University [2]Dept. of Biostatistics, Michigan University [2]Dept. of Psychiatry, Columbia University

Probabilistic reward task and EMBARC study

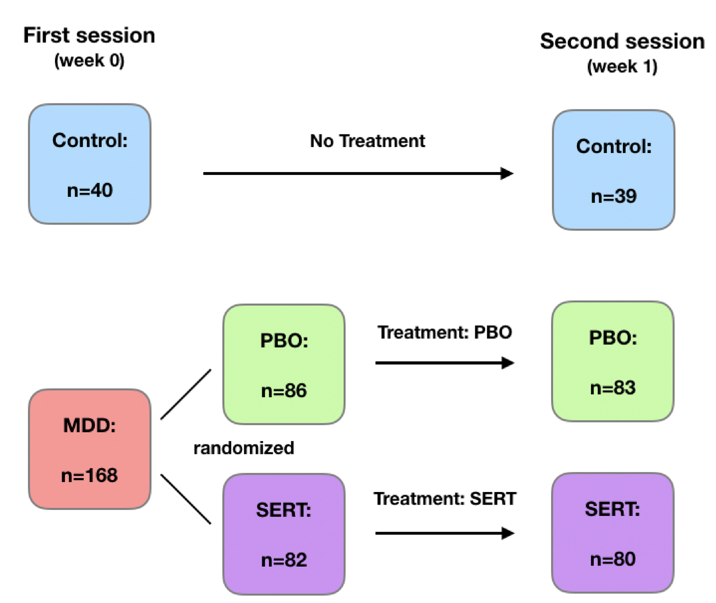
Probabilistic reward task (PRT) is a computer-based behavioral experiment that measures the subject's ability to modify behavior in response to rewards.

- Two states: Subjects see a cartoon face with a **short** or **long** mouth on each trial. **Difference in size** between the short and long mouths is **small**.
- Task: Subjects indicate whether a short or long mouth was presented. Correct responses have chances to be rewarded **(not always rewarded)**.
- Imbalanced rewards: The correct response **(rich reward state)** to a short mouth was rewarded **more frequently** than the correct response to a long mouth **(lean reward state)**.
- Response bias: Subjects tend to **prioritize** states with **higher rewards**.

EMBARC study is a randomized trial for patients with **Major depressive disorder (MDD)**.



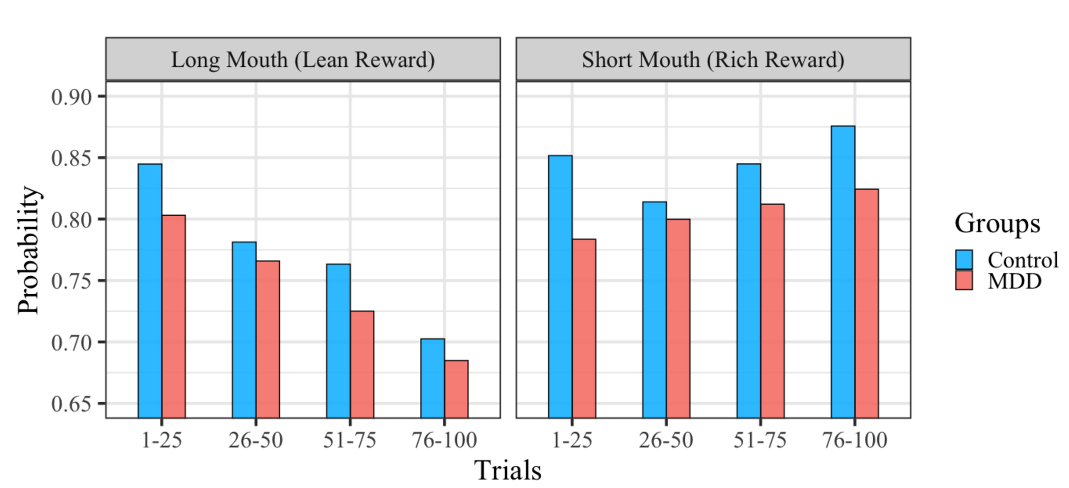
(a) PRT Schematic Diagram



(b) EMBARC Design Diagram

Exploratory analysis

We estimated $P(\text{Action} = \text{State}|\text{State} = j),\ j = 0, 1$ for the MDD and Control group. We observed an **increase** of **response bias** as the trial progresses.



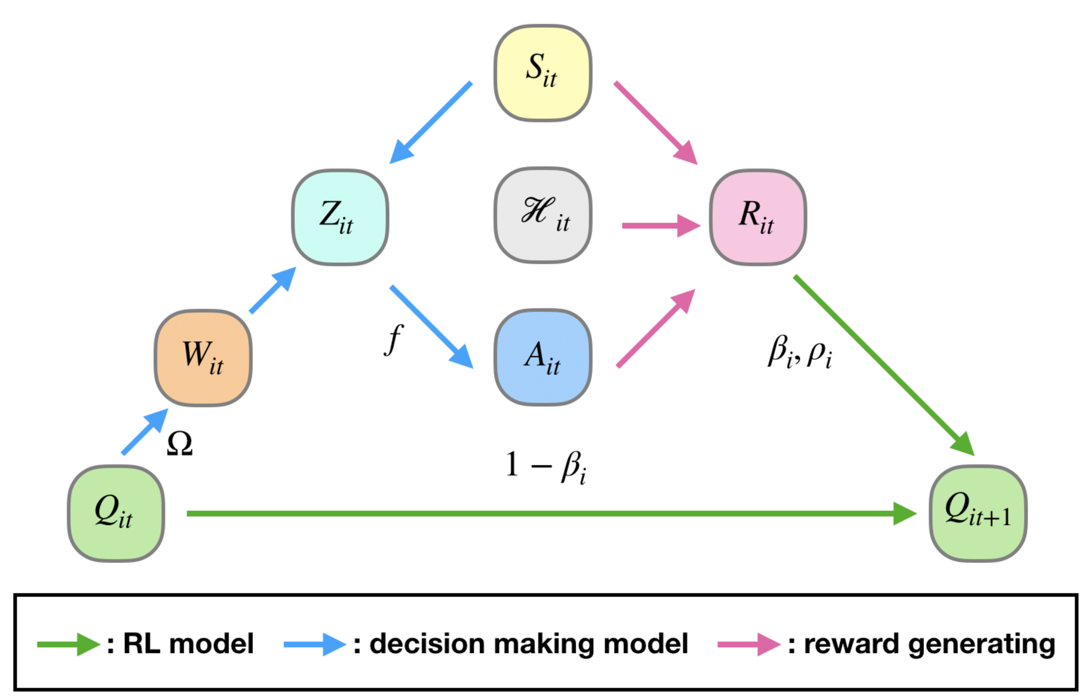
Our goal

- Characterize agent's **heterogeneous** decision making behavior **(Behavioral Cloning)**.
- Identify the **difference in reward learning abilities** between
  1. the patients with **MDD** vs the healthy **Control** group.
  2. **Antidepressant sertraline (SERT)** vs **placebo (PBO)** for the MDD patients.
- Investigate the **cause of abnormalities** in reward learning for the MDD patients.

Problem setups

- State and action space: $S_{it} \in \{0, \ldots, m-1\}$, $A_{it} \in \{0, 1\}$ (PRT is a special case for $m = 2$).
- Problem size: subjects $(i = 1, \ldots, n)$ from a subgroup, trials $(t = 1, \ldots, T)$ for each subject.
- Decision dynamics: $\ldots, \to S_{it-1} \to A_{it-1} \to R_{it-1} \to S_{it} \to A_{it} \to R_{it} \to \ldots$.
- State-generating: State S_{it} is generated **independent** of S_{it-1} and A_{it-1}.

Semiparametric inverse RL model



1. RL model

- $Q_{it}(a, s)$: the **expected reward** of taking action a at state s.
- The **reward prediction error** between obtained and expected reward:

$$\delta_{it} = \rho_i R_{it} - Q_{it}(a, s).$$

- $\rho_i > 0$: reward sensitivity.
- Expected reward **evolves** based on the **stochastic gradient descent**

$$Q_{i,t+1}(a, s) = Q_{it}(a, s) + \beta_i \delta_{it} I_{it}(a, s)$$

- $\beta_i \in (0, 1)$: learning rate.

2. Decision making model

- The **"belief" of the expected reward** characterizing the **uncertainty** of current state:

$$W_{it}(a, s) = \omega_{ss} Q_{it}(a, s) + \sum_{r \neq s} \omega_{sr} Q_{it}(a, r),$$

- The **contrast** between two actions:

$$Z_{it} = W_{it}(1, S_{it}) - W_{it}(0, S_{it}).$$

- The **probability** of $A_{it} = 1$ conditional on history $\mathcal{H}_{it}$ and S_{it}:

$$\text{logit}\, P(A_{it} = 1 \mid S_{it}, \mathcal{H}_{it}) = f(Z_{it}),$$

- $f(\cdot)$: an unknown **non-decreasing** reward sensitivity function satisfying $f(0) = 0$.

3. Subject specific heterogeneity

- **Transformed** (β_i, ρ_i):

$$(\nu_i, \gamma_i) \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{where} \quad \nu_i = \log(\beta_i/(1-\beta_i)) \quad \text{and} \quad \gamma_i = \log(\rho_i)$$

- Let $\mu_\gamma \equiv 1$ to ensure **identifiability**.
- γ_i: the relative sensitivity of the i-th subject.

Model implementation

- Maximizing the log-likelihood of actions **conditional** on states and rewards

$$\sum_{i=1}^{n} \log \left[\iint \left\{ \prod_{t=1}^{T} P(A_{it} \mid S_{it}, \mathcal{H}_{it}; \nu_i, \gamma_i) \right\} \phi(\nu_i, \gamma_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\nu_i d\gamma_i \right],$$

where $\phi(\cdot, \cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- The double integral is approximated by **bivariate Gauss–Hermite quadrature**.
- The non-decreasing function $f(\cdot)$ is approximated using monotone **I-splines**

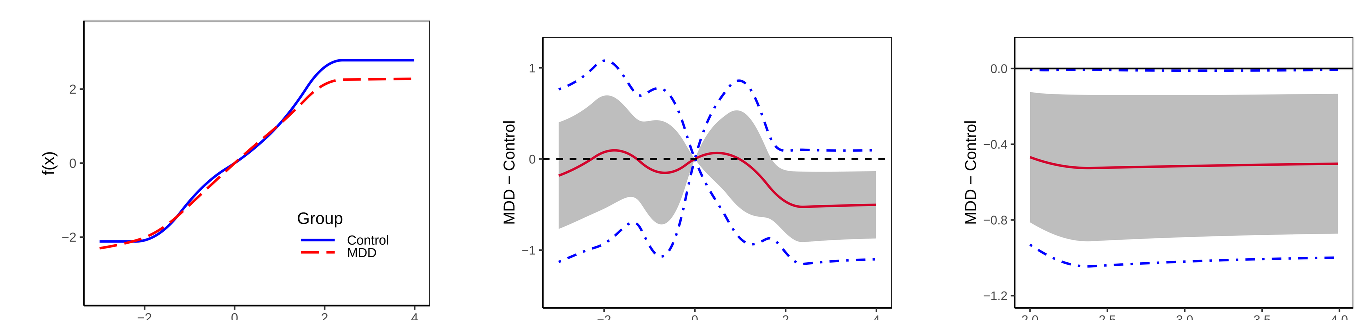
$$\tilde{f}(x) = \sum_{k=1}^{K} \{I_k(x) - I_k(0)\}\, b_k.$$

- **Nonparametric bootstrap** is used for inference.
- The 95% **pointwise confidence band (PCB)** and the 95% **simultaneous confidence band (SCB)** is constrcuted by bootstrap samples.

Application to EMBARC Study

MDD Group vs Control Group:

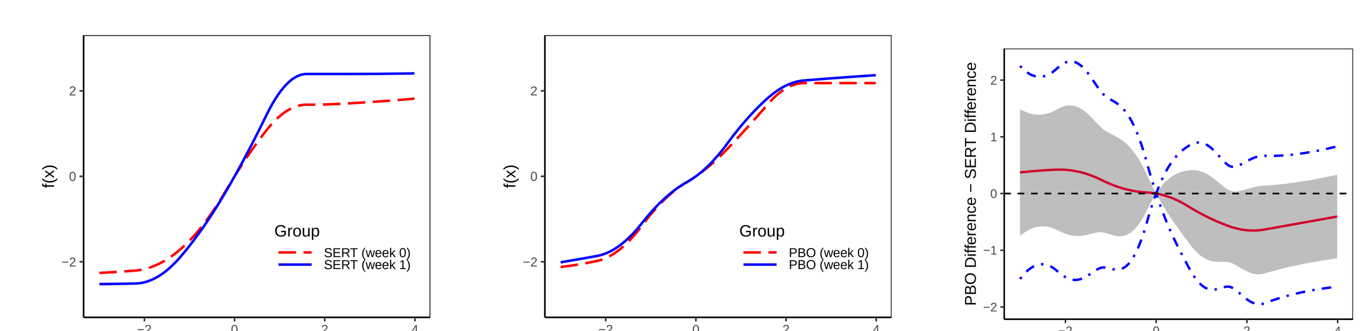
- The difference of **learning rate** between MDD group and control group is **not significant**.
- The Control group has a **larger probability** of taking **correct actions at rich reward states** than the MDD group when subjects in both groups **receive adequate rewards** in rich reward states.



(a) Estimates of f_m and f_c (b) PCB/SCB of $f_m - f_c$ (c) PCB/SCB of $f_m - f_c$ at $x > 2$

SERT Group vs PBO Group:

- The **one-week changes in learning rate** between PBO and SERT groups are **not significantly different**.
- There might be a **positive impact of sertraline** on MDD patients, potentially bringing their reward learning sensitivity **closer to** that of **healthy individuals** at the rich state.



(a) Estimates of f_{s0} and f_{s1} (b) Estimates of f_{p0} and f_{p1} (c) $(f_{p1} - f_{p0}) - (f_{s1} - f_{s0})$

In general:

- The **abnormalities in reward learning** for the MDD patients are more likely due to **reduced sensitivity to received rewards**.
- The fitted reward sensitivity functions are **nonlinear**.