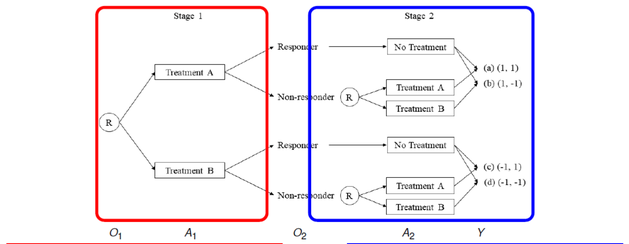
Modified Q-learning for Optimizing Dynamic Treatment Regime with Repeated-Measures Outcomes

Yuan Zhang, David M. Vock, Thomas A. Murray

Division of Biostatistics, School of Public Health, University of Minnesota

INTRODUCTION

- Optimal dynamic treatment regime: a sequence of decision making functions of a patient's covariate and treatment history that *maximizes* the outcome of interest
- Goals of statistical analysis of data collected from a sequential multiple assignment randomized trial (SMART): (1) Identify the stagewise optimal rule as a function of prior history; (2) Estimate the heterogeneous causal effect of treatment at a stage, assuming subjects follow the optimal rules at subsequent stages
- Q-learning: a backward induction algorithm
  - Work from stage 2, regress Y on H_2 and A_2, and choose the optimal rule $\hat{d}_2^{\text{opt}}$ so that the stage 2 regression form is maximized
  - Move to stage 1, regress the maximized parametric form on H_1 and A_1, and choose the optimal rule $\hat{d}_1^{\text{opt}}$ so that the stage 1 regression form is maximized



Let $H_1 = O_1$.

$$Q_1(H_1, A_1) = \mathbb{E}\left(\mu_2(H_2, A_2 = \hat{d}_2^{\text{opt}}; \hat{\psi}_2) \middle| H_1, A_1\right) = \mu_1(H_1, A_1; \psi_1)$$

$$\hat{d}_1^{\text{opt}} = \underset{a_1 \in \mathcal{A}_1}{\arg\max}\, \mu_1\left(H_1, A_1 = a_1; \hat{\psi}_1\right)$$

$$\hat{\tau}_1(H_1) = \mu_1\left(H_1, A_1 = 1; \hat{\psi}_1\right) - \mu_1\left(H_1, A_1 = -1; \hat{\psi}_1\right)$$

Let $H_2 = (O_1, A_1, O_2)$.

$$Q_2(H_2, A_2) = \mathbb{E}(Y|H_2, A_2) = \mu_2(H_2, A_2; \psi_2)$$

$$\hat{d}_2^{\text{opt}} = \underset{a_2 \in \mathcal{A}_2}{\arg\max}\, \mu_2\left(H_2, A_2 = a_2; \hat{\psi}_2\right)$$

$$\hat{\tau}_2(H_2) = \mu_2\left(H_2, A_2 = 1; \hat{\psi}_2\right) - \mu_2\left(H_2, A_2 = -1; \hat{\psi}_2\right)$$

MOTIVATION

- The DLD study continues to monitor participants' performance after stage 2 treatment
- Q-learning *collapses* repeated-measures outcome at the end of stage 2 using a weighted average
  - Not capturing the correlations between stagewise repeated-measures outcome
  - Not able to give estimates of treatment effect at all time points of interest

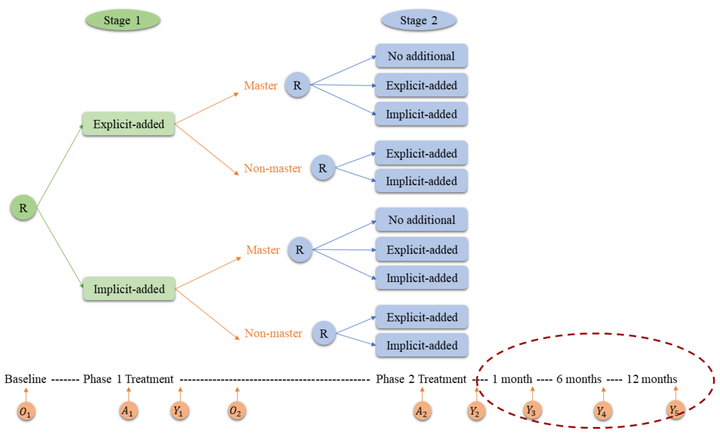


Figure: An ongoing SMART for children with developmental language disorder (DLD)

METHODOLOGY

- **Q-learning with GEE** to capture the correlation between stagewise repeated-measures outcomes
  - Treat repeated-measures as a vector $\boldsymbol{Y}_2 = (Y_2, Y_3, Y_4, Y_5)^T$
  - Consider time-varying coefficients in the parametric form of regression
  - Estimate coefficients using generalized estimating equations (GEE)
- **Modified Q-learning with GEE** to capture the correlation between vector outcomes across stages
  - Use observed $\boldsymbol{Y}_2$ and add back Murphy's regret function if the subject does not follow the optimal rule at stage 2

$$Q_2(H_{2i}, A_{2i}, \boldsymbol{t}_2) = \mathbb{E}\left(\boldsymbol{Y}_{2i} | H_{2i}, A_{2i}\right) = \boldsymbol{x}_{20,i}^T \beta_{2,\boldsymbol{t}_2} + A_{2i} \boldsymbol{x}_{21,i}^T \psi_{2,\boldsymbol{t}_2}$$

$$Q_1(H_{1i}, A_{1i}, \boldsymbol{t}_1) = \mathbb{E}\left\{ \left(\begin{matrix} \boldsymbol{Y}_{1i} \\ Y_{2i} + 2\mathbb{1}\left\{A_{2i} \neq \hat{d}_{2i}^{\text{opt}}\right\} \left| \boldsymbol{x}_{21,i}^T \hat{\psi}_{2,\boldsymbol{t}_2} \right| \end{matrix} \right) \middle| H_{1i}, A_{1i} \right\}$$

  - This semiparametric approach reinstates the correlation between Y_1 and $\boldsymbol{Y}_2$ to some degree and is robust to misspecification of $\boldsymbol{x}_{20,i}^T \beta_{2,\boldsymbol{t}_2}$

SIMULATION

- Conduct a simulation study to compare the two methods illustrated above
- Simulation scenarios
  1) Correlation structure between Y_1 and $\boldsymbol{Y}_2$: positive, independent, and negative
  2) Model misspecification of $\boldsymbol{x}_{20,i}^T \beta_{2,\boldsymbol{t}_2}$
- Results
  - Modified Q-learning with GEE has a more stable performance in correctly identifying the optimal rule, it performs especially well when the main model is misspecified and correlation structure between Y_1 and $\boldsymbol{Y}_2$ is negative
  - The predicted heterogeneous causal effects by modified Q-learning with GEE are closer to the true values
  - Modified Q-learning with GEE universally gives a lower standard deviation of prediction errors

Table: Probability of correct identification (PCI) of stage 1 optimal rules and root mean square error (RMSE, mean (SD)) of estimated heterogeneous causal effects at time 2 and 3, based on stage 1 Q-function.

Correlation	Sample Size	Standard Q-learning with GEE			Modified Q-learning with GEE		
		PCI_1[2]	RMSE_{τ_2}[3]	RMSE_{τ_3}	PCI_1	RMSE_{τ_2}	RMSE_{τ_3}
Nonzero Correlation Only							
Positive	200	0.940	3.07 (0.69)	3.38 (0.69)	0.889	1.67 (0.44)	2.13 (0.40)
	400	0.957	3.02 (0.51)	3.31 (0.52)	0.924	1.44 (0.28)	1.94 (0.24)
Independent	200	0.959	1.32 (0.19)	1.85 (0.19)	0.958	1.36 (0.23)	1.88 (0.21)
	400	0.972	1.23 (0.11)	1.77 (0.11)	0.972	1.26 (0.13)	1.79 (0.12)
Negative	200	0.777	2.95 (0.66)	3.18 (0.60)	0.942	1.68 (0.45)	2.15 (0.40)
	400	0.836	2.84 (0.49)	3.08 (0.44)	0.961	1.45 (0.27)	1.93 (0.23)
Misspecification of Main Effect							
Positive	200	0.944	1.63 (0.38)	2.06 (0.34)	0.946	1.67 (0.44)	2.12 (0.40)
	400	0.959	1.44 (0.26)	1.91 (0.21)	0.960	1.44 (0.26)	1.94 (0.24)
Independent	200	0.965	2.58 (0.47)	2.77 (0.43)	0.978	1.38 (0.23)	1.89 (0.22)
	400	0.975	2.57 (0.32)	2.74 (0.31)	0.984	1.26 (0.14)	1.79 (0.13)
Negative	200	0.815	5.72 (0.74)	5.64 (0.75)	0.970	1.71 (0.46)	2.16 (0.41)
	400	0.862	5.74 (0.53)	5.66 (0.54)	0.979	1.47 (0.29)	1.96 (0.25)

[1] The figures are based on 1000 simulations.
[2] PCI_1 is not included because stage 2 model is the same for both methods.
[3] RMSE_{τ_1} is not included because both methods generate similar values for this metric.

APPLICATION

- DLD is an ongoing study and the data is unavailable, so we use a simulated data set ($n = 250$) from ENGAGE study to show how *versatile* modified Q-learning with GEE is

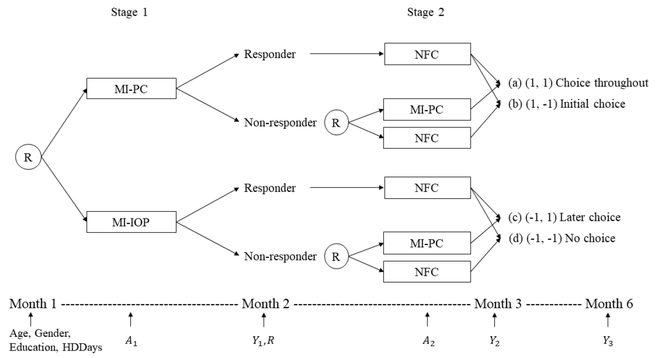
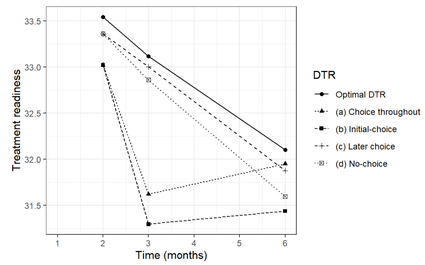
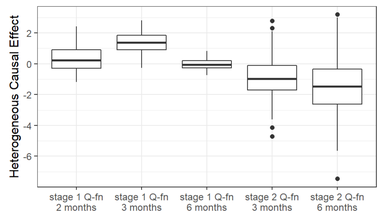


Figure: ENGAGE study for patients with relatively severe substance-use disorders. IOP = Intensive outpatient program; MI-IOP = Phone-based MI session focusing on engaging the individual in IOP; MI-PC = Phone-based MI session focusing on facilitating personal choice; NFC = No further contact. Outcome of interest (Y) = treatment readiness.

- Identification of optimal rules:
  - At stage 1, assuming subjects follow the optimal rule at stage 2, 15.6% of them should be assigned to MI-PC, and 84.4% should be assigned to MI-IOP
  - At stage 2, based on the history prior to stage 2, 70% of the subjects should not be randomized because they are responders, 23.2% should be assigned to MI-PC, and 6.8% should be assigned to NFC
- Estimation of the optimal trajectory:



- Distribution of estimated individual treatment effects at different time points of interest based on the Q-function at each stage:



DISCUSSION

- We proposed to use modified Q-learning with GEE to analyze SMART data with repeated-measures outcomes, and it is readily extended to analyze discrete outcomes
- Unstructured working correlation is recommended for the implementation of GEE
- Diagnostics of Q-learning with GEE can take advantage of existing goodness-of-fit techniques such as QIC