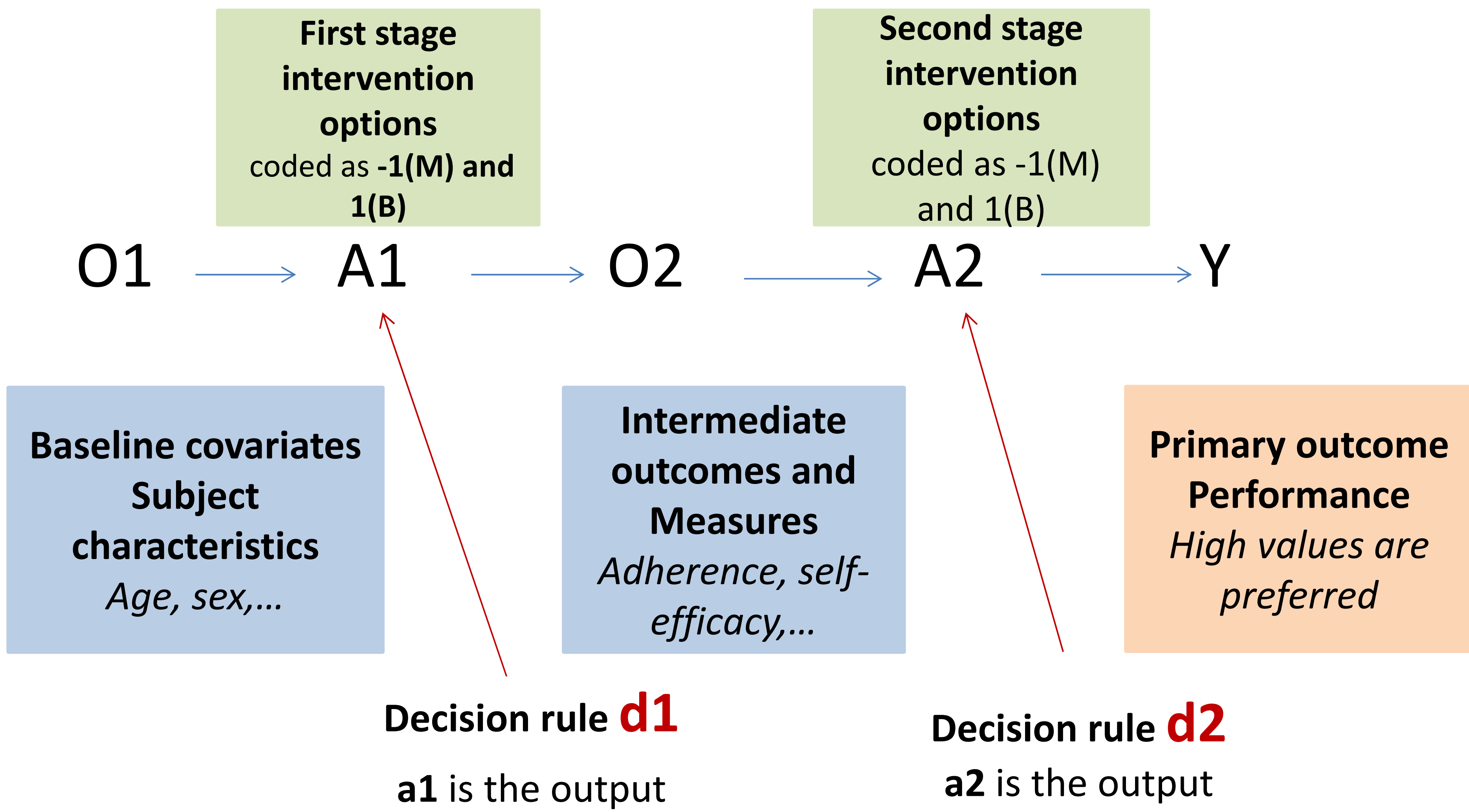
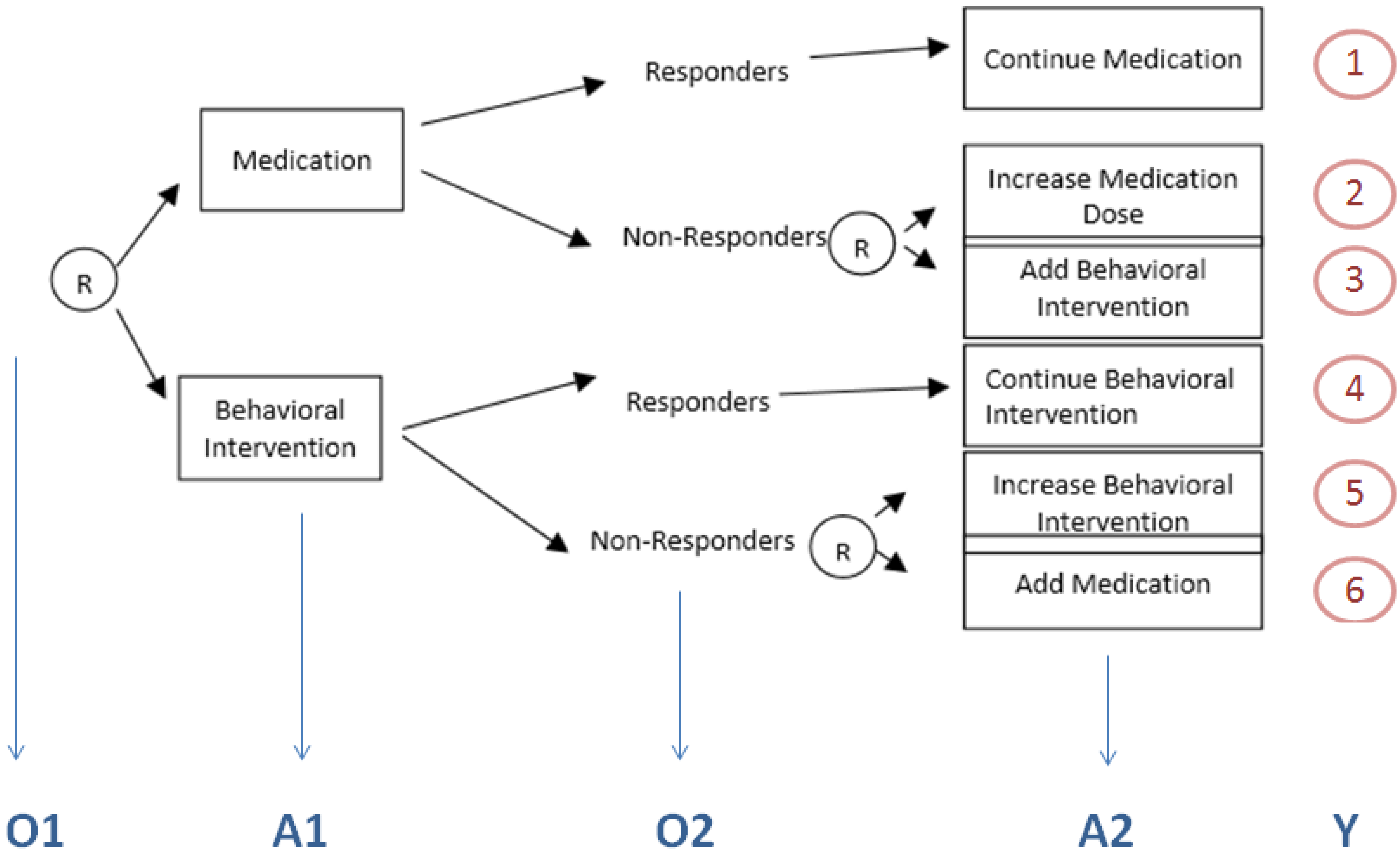
Q-learning

A data analysis method for constructing adaptive interventions

SMART


First stage intervention options
coded as -1(M) and 1(B)
Second stage intervention options
coded as -1(M) and 1(B)
O1
A1
O2
A2
Y
Baseline covariates
Subject characteristics
Age, sex,...
Intermediate outcomes and Measures
Adherence, self-efficacy,...
Primary outcome
Performance
High values are preferred
Decision rule d1
a1 is the output
Decision rule d2
a2 is the output

SMART: ADHD data example



Q-learning Algorithm

- Goal: to find optimal decision rules (d_1^*, d_2^*)
  - For example, SMART of ADHD data is aiming to develop an adaptive intervention for improving school performance
- Backwards induction:
  - start from the last intervention
  - Controlling for effects of both past and subsequent adaptive intervention options
- Assumption:
  - Multivariate distribution of O1,O2, and Y for every sequence of decisions a1,a2 is known
  - Larger primary outcome value is preferred

Algorithm- *framework*

- Optimal decision at second stage

$$d_2^*(O_1, a_1, O_2) = \arg\max_{a_2} Q_2(O_1, a_1, O_2, a_2),$$

$$\textit{where} \quad Q_2(O_1, a_1, O_2, a_2) = E[Y \mid O_1, a_1, O_2, a_2]$$

- Optimal decision at the first stage

$$d_1^*(O_1) = \arg\max_{a_1} Q_1(O_1, a_1),$$

$$\textit{where} \quad Q_1(O_1, a_1) = E[\max_{a_2} Q_2(O_1, a_1, O_2, a_2) \mid O_1, a_1]$$

- Q_1, Q_2 are the *Q-functions to be estimated*

Algorithm- *Q-function*

- Second stage Q-function(*linear regression*)

$$Q_2(O_1, A_1, O_2, A_2; \gamma_2, \alpha_2) = \gamma_{20} + \gamma_{21}O_1 + \gamma_{22}A_1 + \gamma_{23}O_1A_1 + \gamma_{24}O_2 + (\alpha_{21} + \alpha_{22}A_1 + \alpha_{23}O_2)A_2$$

$$parameters \quad vectors \quad \hat{\gamma}_2, \hat{\alpha}_2 \quad are \quad obtained \quad by:$$

$$Y \sim \gamma_{20} + \gamma_{21}O_1 + \gamma_{22}A_1 + \gamma_{23}O_1A_1 + \gamma_{24}O_2 + (\alpha_{21} + \alpha_{22}A_1 + \alpha_{23}O_2)A_2$$

- First stage Q-function(*linear*)

$$Q_1(O_1, A_1; \gamma_1, \alpha_1) = \gamma_{10} + \gamma_{11}O_1 + (\alpha_{11} + \alpha_{12}O_1)A_1$$

$$parameters \quad vectors \quad \hat{\gamma}_1, \hat{\alpha}_1 \quad are \quad obtained \quad by:$$

$$\tilde{Y} \sim \gamma_{10} + \gamma_{11}O_1 + (\alpha_{11} + \alpha_{12}O_1)A_1, \quad where$$

$$\tilde{Y}_i = \hat{\gamma}_{20} + \hat{\gamma}_{21}O_{1i} + \hat{\gamma}_{22}A_{1i} + \hat{\gamma}_{23}O_{1i}A_{1i} + \hat{\gamma}_{24}O_{2i} + |\hat{\alpha}_{21} + \hat{\alpha}_{22}A_{1i} + \hat{\alpha}_{23}O_{2i}|$$

Algorithm- *estimated optimal rules*

- Optimal decision d_2^*

$$\hat{d}_{2i}^*(O_{1i}, A_{1i}, O_{2i}) = sign(\hat{\alpha}_{21} + \hat{\alpha}_{22}A_{1i} + \hat{\alpha}_{23}O_{2i}), \text{which} \quad leads \quad to$$

$$\hat{Y}_i = \hat{\gamma}_{20} + \hat{\gamma}_{21}O_{1i} + \hat{\gamma}_{22}A_{1i} + \hat{\gamma}_{23}O_{1i}A_{1i} + \hat{\gamma}_{24}O_{2i} + |\hat{\alpha}_{21} + \hat{\alpha}_{22}A_{1i} + \hat{\alpha}_{23}O_{2i}|,$$

the expected mean outcome obtained by choosing the optimal decision d_2^*

- Optimal decision d_1^*

$$\hat{d}_{1i}^*(O_{1i}) = sign(\hat{\alpha}_{11} + \hat{\alpha}_{12}O_{1i})$$

CI of regression coefficients

max Q2

Y model → estimated $\hat{\alpha}_2$

Best decision: $\hat{d}_2^*$

Ordinary linear regression is used to estimate coefficients in Q2.
Bootstrap: std errors, CIs and hypothesis

max Q1

$\tilde{Y}$ Model → estimated $\hat{\alpha}_1$

Best decision: $\hat{d}_1^*$

$\tilde{Y}$ model contains an absolute value function which is nondifferentiable at 0.
Soft-thresholding with percentile bootstrap: for each bootstrap sample,

$$|\hat{\alpha}_{21}+\hat{\alpha}_{22}A_1+\hat{\alpha}_{23}O_2|$$

Replaced by

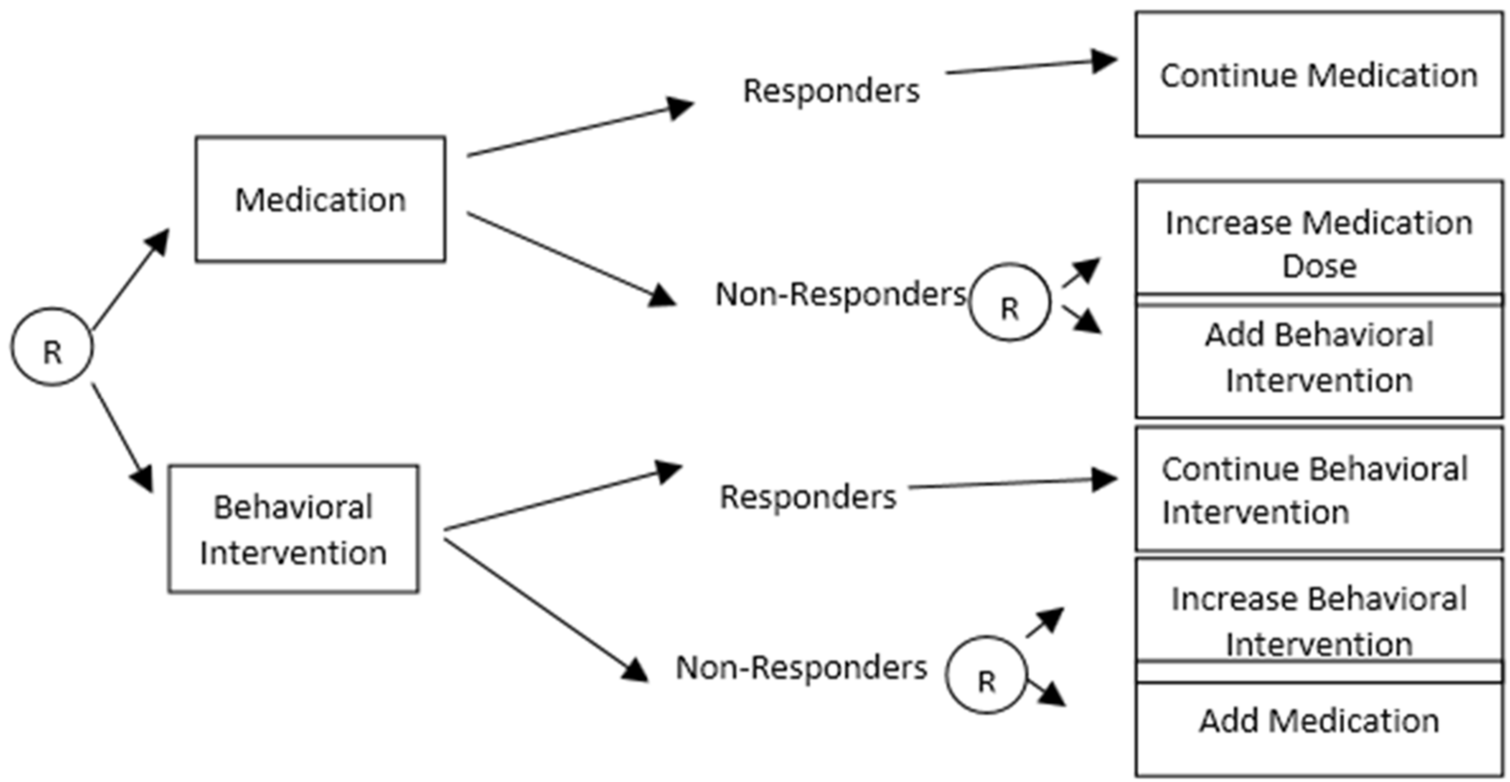
$$|\hat{\alpha}_{21}+\hat{\alpha}_{22}A_1+\hat{\alpha}_{23}O_2|\left(1-\frac{\lambda}{|\hat{\alpha}_{21}+\hat{\alpha}_{22}A_1+\hat{\alpha}_{23}O_2|}\right)$$

$$\lambda = 3(1, A_1, O_2)^T\hat{\Sigma}_2(1, A_1, O_2)/N, \hat{\text{cov}}(\hat{\alpha}_2) = \hat{\Sigma}_2/N$$

(shrink the term to zero if it's small)

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome(O2)*



Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- **Data**
  - Primary outcome

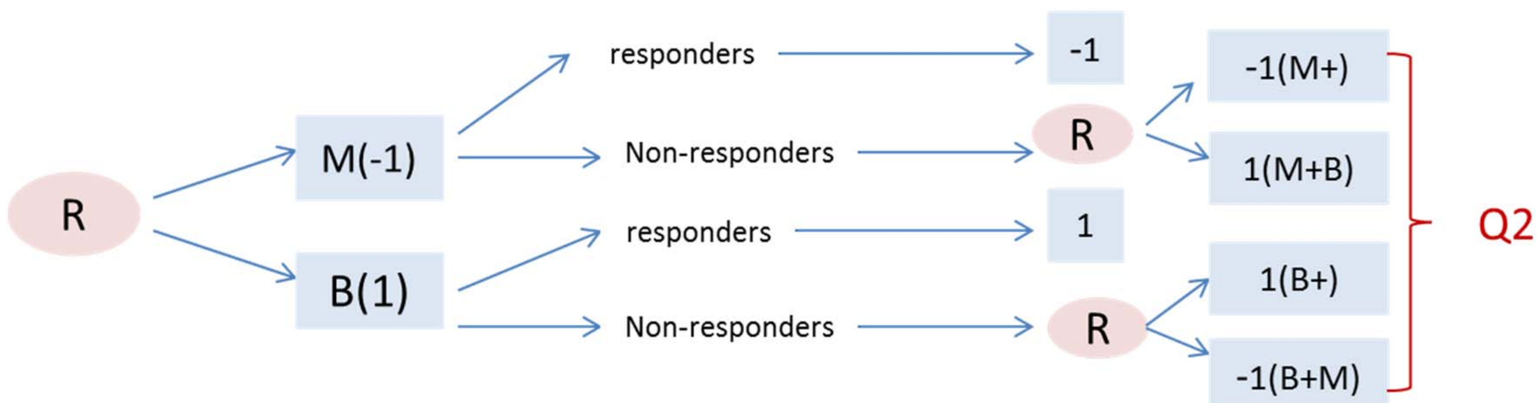
    Y: level of children's classroom performance based on the IRS after an 8-month period is our primary outcome. This outcome ranges from 1 to 5, with higher values reflecting better classroom performance.
  - vectors $(\underline{O_1}, A_1, \underline{O_2}.A_2)$

O11	(0,1)	medication prior to first-stage intervention, "yes"=1
O12	0~3	attention-deficit/hyperactivity disorder symptoms, fewest=3(better)
O13	(0,1)	oppositional defiant disorder diagnosis, "has ODD"=1
A1	(-1,1)	second-stage intervention options, "medicine"=-1
O21	**integer**	month of nonresponse
O22	(0,1)	adherence to first-stage intervention, "high"=1
A2	(-1,1)	second-stage intervention options

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- **Model Q2** (only for "non-responders")
  - Obtain estimated coefficients: $\hat{\gamma}_2, \hat{\alpha}_2$
  - Best decision: $\hat{d}^*_{2i} = sign(\hat{\alpha}_{21} + \hat{\alpha}_{22} A_{1i} + \hat{\alpha}_{23} O_{2i})$
  - Calculate $\tilde{Y}$ with optimal decision above
  - For "responders", let $\tilde{Y} = Y$



  - Model

$$Y \sim \gamma_{20} + \gamma_{21} O_{11} + \gamma_{22} O_{12} + \gamma_{23} O_{13} + \gamma_{24} A_1 + \gamma_{25} O_{11} A_1 + \gamma_{26} O_{21} + \gamma_{27} O_{22} + (\alpha_{21} + \alpha_{22} A_1 + \alpha_{23} O_{22}) A_2$$

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- **Estimated coefficients for Q2**

Estimated Coefficients for Q_2 (N = 81)

Effect	Estimate	SE	95% confidence interval Lower limit	Upper limit
Intercept	1.36	0.53		
O_{11} (medication prior to first-stage intervention)	−0.27	0.31		
O_{12} (baseline: attention-deficit/hyperactivity disorder symptoms)	0.94	0.26		
O_{13} (baseline: oppositional defiant disorder diagnosis)	0.93	0.28		
O_{21} (month of nonresponse)	0.02	0.10		
O_{22} (adherence to first-stage intervention)	0.18	0.27		
A_1 (first-stage intervention options)	0.03	0.14		
A_2 (second-stage intervention options)	−0.72	0.22	−1.15	−0.29
$O_{22} \times A_2$ (Adherence to First-Stage Intervention × Second-Stage Intervention Options)	0.97	0.28	0.41	1.53
$A_1 \times A_2$ (First-Stage Intervention Options × Second-Stage Intervention Options)	0.05	0.13	−0.22	0.32

Estimates of $(\hat{\alpha}_{21} + \hat{\alpha}_{22}A_1 + \hat{\alpha}_{23}O_{22})$ for Every Combination of A_1 and O_{22} (N = 81)

A_1	O_{22}	Estimated $(\hat{\alpha}_{21} + \hat{\alpha}_{22}A_1 + \hat{\alpha}_{23}O_{22})$	SE	95% confidence interval Lower limit	Upper limit
−1 (medication)	1 (high adherence)	0.20	0.23	−0.26	0.67 ★
−1 (medication)	0 (low adherence)	−0.77	0.27	−1.30	−0.23
1 (behavioral intervention)	1 (high adherence)	0.30	0.22	−0.13	0.74 ★
1 (behavioral intervention)	0 (low adherence)	−0.67	0.24	−1.14	−0.19

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- **Model Q1**
  - $\tilde{Y} = Y$ for responders
  - Estimated quality of the optimal second-stage intervention option for non-responders

$$\tilde{Y} = \hat{\gamma}_{20} + \hat{\gamma}_{21} O_{11} + \hat{\gamma}_{22} O_{12} + \hat{\gamma}_{23} O_{13} + \hat{\gamma}_{24} A_1 + \hat{\gamma}_{25} O_{11} A_1 + \hat{\gamma}_{26} O_{21} + \hat{\gamma}_{27} O_{22} + | \hat{\alpha}_{21} + \hat{\alpha}_{22} A_1 + \hat{\alpha}_{23} O_{22} |$$

  - estimated coefficients $\hat{\gamma}_1, \hat{\alpha}_1$ for Q1 is obtained by

$$\tilde{Y} \sim \gamma_{10} + \gamma_{11} O_{11} + \gamma_{12} O_{12} + \gamma_{13} O_{13} + (\alpha_{11} + \alpha_{12} O_{11}) A_1$$

  - Best decision: $\hat{d}_{1i}^{*} = sign(\hat{\alpha}_{11} + \hat{\alpha}_{12} O_{1i})$

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- **Estimated coefficients for Q1**

Estimated Coefficients and Soft-Threshold Confidence Intervals for Q_1 (N = 138)

Effect	Estimate	*SE*	90% confidence interval	
			Lower limit	Upper limit
Intercept	2.61	0.16		
O_{11} (medication prior to first-stage intervention)	−0.37	0.14		
O_{12} (baseline: ADHD symptoms)	0.73	0.11		
O_{13} (baseline: ODD diagnosis)	0.75	0.13		
A_1 (first-stage intervention options)	0.17	0.07	−0.01	0.34
$O_{11} \times A_1$ (Medication Prior to First-Stage Intervention × First-Stage Intervention Options)	−0.32	0.14	−0.59	−0.06

Note. ADHD = attention-deficit/hyperactive disorder; ODD = oppositional defiant disorder.

Estimates of $(\hat{\alpha}_{11} + \hat{\alpha}_{12}O_{11})$ for Each Level of O_{11}

O_{11}	Estimated $(\hat{\alpha}_{11} + \hat{\alpha}_{12}O_{11})$	*SE*	90% confidence interval	
			Lower limit	Upper limit
1 (medication prior to first-stage intervention)	−0.15	0.12	−0.44	0.11
0 (no medication prior to first-stage intervention)	0.17	0.07	−0.01	0.34

Example: ADHD data

- *whether to randomize or not depends on an intermediate outcome*

- Overall, optimal sequence of decision rules are:

IF the child received medication prior to the first stage of the intervention,

THEN offer low dose of medication or low-intensity behavioral intervention.

ELSE IF the child did not receive medication prior to the first stage of the intervention,

THEN offer low-intensity behavioral intervention.

Then,

IF the child shows inadequate response to the first stage of the intervention,

THEN IF child's adherence to first stage of the intervention is low,

THEN augment the first-stage intervention option with the other type of intervention.

ELSE IF child's adherence to the first stage of the intervention is high,

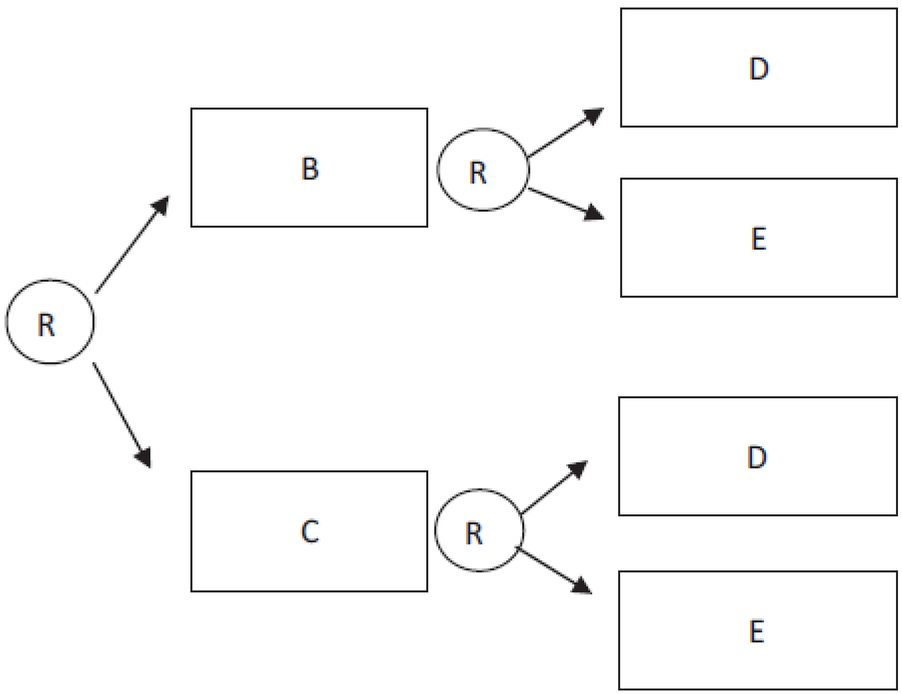
THEN augment the first-stage intervention option with the other type of intervention or intensify the first-stage intervention option.

ELSE IF the child shows adequate response to the first stage of the intervention,

THEN continue first-stage intervention.

Q-learning for other SMART(1)

-with no embedded tailoring variables



- Y of all the subjects are used for Q2 regression
- Optimal second-stage decision

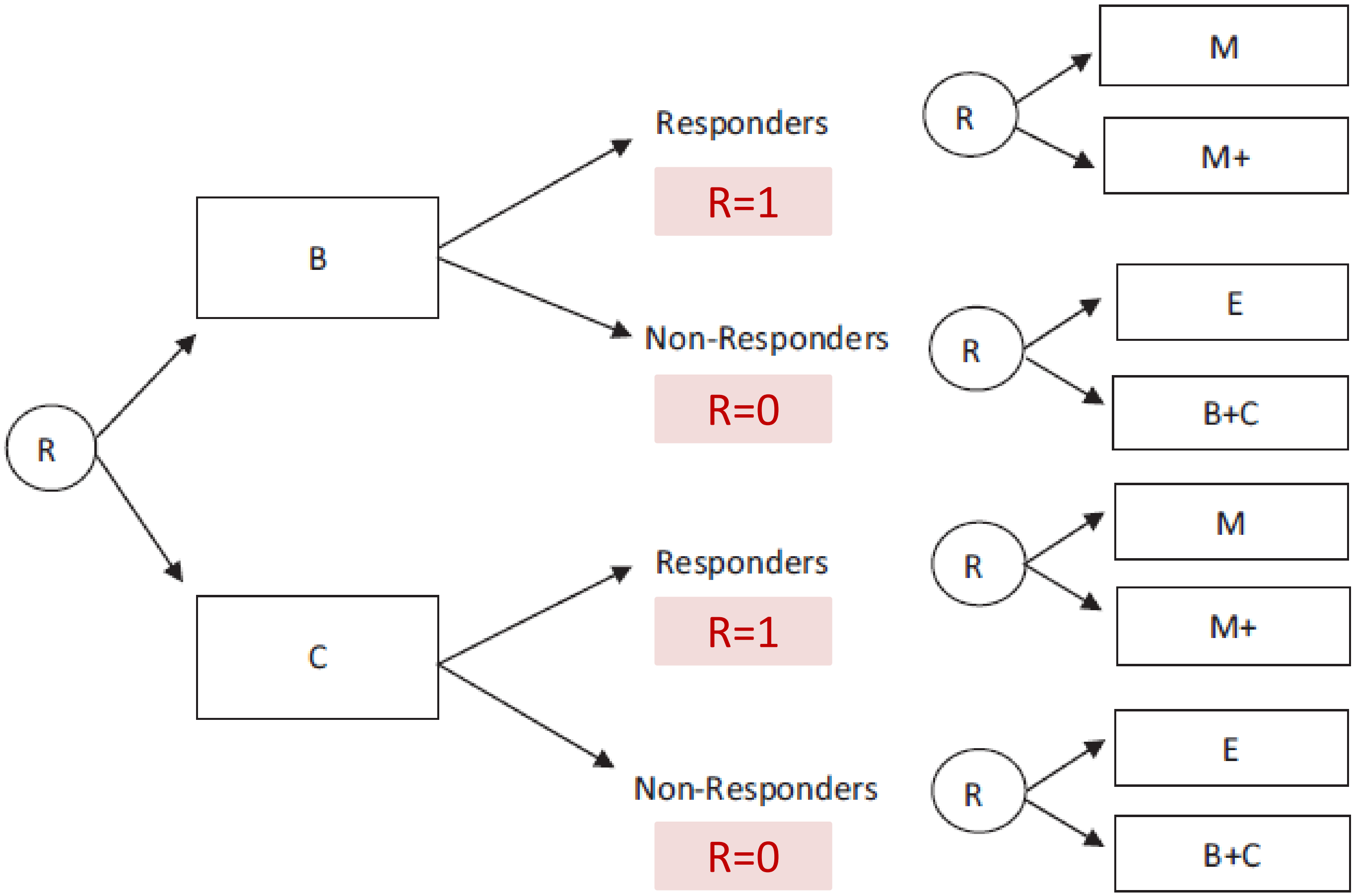
  $\hat{d}^*_{2i} = sign(\hat{\alpha}_{21} + \hat{\alpha}_{22} A_{1i} + \hat{\alpha}_{23} O_{2i})$
- Regress $\tilde{Y}$ on O1, A1
- Optimal first-stage decision

  $\hat{d}^*_{1i} = sign(\hat{\alpha}_{11} + \hat{\alpha}_{12} O_{1i})$
- Example

O1=1	A1
$(\alpha_{11} + \alpha_{12}) > 0$	1
$(\alpha_{11} + \alpha_{12}) < 0$	-1

Q-learning for other SMART(2)

- randomization to different second stage interventions depends on an intermediate outcome(R)



Q-learning for other SMART(2)

- randomization to different second stage interventions depends on an intermediate outcome(R)

- Model Q2 by regression:

$$Y \approx \gamma_{20} + \gamma_{21} O_1 + \gamma_{22} A_1 + \gamma_{23} O_1 A_1 + \gamma_{24} O_{21} + \gamma_{25} O_{22}$$
$$+ [\alpha_{21} R + \alpha_{22}(1-R) + \alpha_{23} R A_1 + \alpha_{24}(1-R) A_1$$
$$+ \alpha_{25} R O_{21} + \alpha_{26}(1-R) O_{22}] A_2$$

- Optimal second-stage decision
  - For responders(R=1): $\hat{d}^*_2 = sign(\hat{\alpha}_{21} + \hat{\alpha}_{23} A_1 + \hat{\alpha}_{25} O_{21})$
  - For non-responders(R=0): $\hat{d}^*_2 = sign(\hat{\alpha}_{22} + \hat{\alpha}_{24} A_1 + \hat{\alpha}_{26} O_2)$

- $\tilde{Y}$ for first stage regression
  - For responders(R=1):

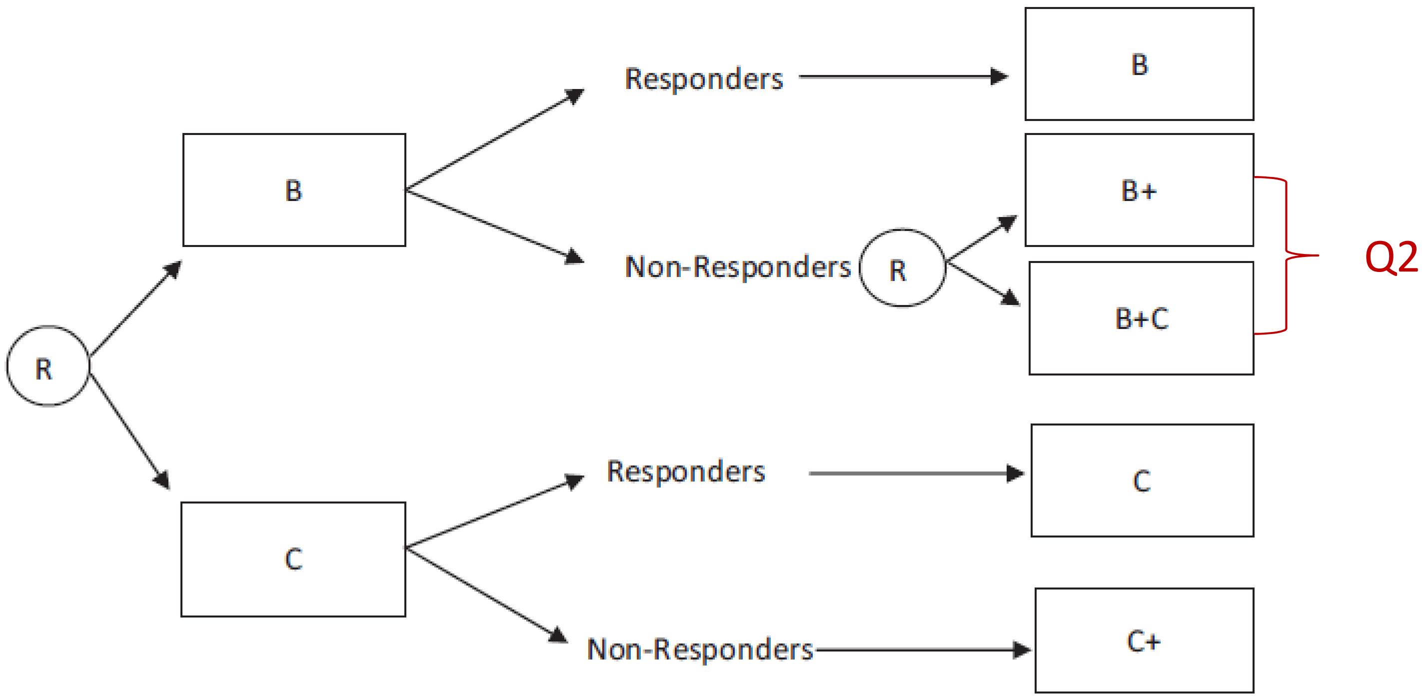
$$\tilde{Y} = \hat{\gamma}_{20} + \hat{\gamma}_{21} O_1 + \hat{\gamma}_{22} A_1 + \hat{\gamma}_{23} A_1 O_1 + \hat{\gamma}_{24} O_{21} + |\hat{\alpha}_{21} + \hat{\alpha}_{23} A_1 + \hat{\alpha}_{25} O_{21}|$$

  - For non-responders(R=0):

$$\tilde{Y} = \hat{\gamma}_{20} + \hat{\gamma}_{21} O_1 + \hat{\gamma}_{22} A_1 + \hat{\gamma}_{23} A_1 O_1 + \hat{\gamma}_{25} O_{22} + |\hat{\alpha}_{22} + \hat{\alpha}_{24} A_1 + \hat{\alpha}_{26} O_{22}|$$

Q-learning for other SMART(3)

-whether to rerandomize or not depends on an intermediate outcome(R) and prior treatment(A1)



Q-learning for other SMART(3)

-whether to rerandomize or not depends on an intermediate outcome(R) and prior treatment(A1)

- *Model Q*2 only for non-responders to treatment A1=B by:

  $$Y \sim \gamma_{20} + \gamma_{21}O_1 + \gamma_{22}O_2 + (\alpha_{21} + \alpha_{22}O_2)A_2$$

- Obtained $\tilde{Y}$
  - Non-responders to A1=B

    $$\tilde{Y} = \hat{\gamma}_{20} + \hat{\gamma}_{21}O_1 + \hat{\gamma}_{22}O_2 + |\hat{\alpha}_{21} + \hat{\alpha}_{22}O_2|$$
  - others $\tilde{Y} = Y$

- $Q_1(O_1, A_1; \gamma_1, \alpha_1) = \gamma_{10} + \gamma_{11}O_1 + (\alpha_{11} + \alpha_{12}O_1)A_1$

Alternative to Q-learning

- Single regression for SMART study with no embedded tailoring variables

$$Y \sim \theta_0 + \theta_1 O_1 + \theta_2 A_1 + \theta_3 O_1 A_1 + \theta_4 O_2 + \theta_5 A_2 + \theta_6 A_1 A_2 + \theta_7 A_2 O_2;$$

- *O*2 can be a mediator in the relationship between *A*1 and *Y* . Adding O2 to a regression in which A1 is used to predict Y will reduce the effect of A1. In the presence of O2, the coefficient for A1 no longer expresses the total effect of the first-stage goal-setting options on the outcome

- Even if *O*2 is not a mediator, the coefficients of the *A*1 terms (main effects and interactions) can be impacted by unknown causes, i.e. unmeasured confounders affecting O2 and Y. of both *O2* and *Y* so that *A*1 might appear to be falsely less or more correlated with *Y*. This bias occurs when *A*1 affects *O*2 while *O*2 and *Y* are affected by the same unknown causes.

Discussion

- Advantages
  - Q-learning appropriately controls for the optimal second-stage intervention option when assessing the effect of the first-stage intervention;
  - The effects estimated by Q-learning incorporate both the direct and indirect effects of the first-stage intervention options, the combination of which is necessary for making intervention decision rules;
  - Q-learning reduces potential bias resulting from unmeasured causes of both the tailoring variables and the primary outcome;
  - Q-learning can be used for studies with more than two stages, and can be easily extended to continuous as well as categorical tailoring variables.

Discussion

- Challenges
  - Bias: In observational studies, direct implementation of this analysis might give biased results due to unmeasured confounding factors that predict the probability of being offered intervention options A1 or A2, given past intervention history. Q-learning should be implemented in combination with methodologies that adjust for confounding;

  - Non-differentiability: Inferential challenges caused by non-differentiability should be taken into consideration when applying Q-learning (non-differentiability arises because the formula for $\tilde{Y}$ (In the current analysis, we used the soft-threshold operation);

  - Tailoring variables selection: Studies often collect information on a large set of covariates. Methods for selecting tailoring variables in randomized settings are required.