A Bayesian Imputation Approach to Optimizing Dynamic Treatment Regimes

Thomas A. Murray

Division of Biostatistics
University of Minnesota

MBI Bayesian Causal Inference Workshop
June 3, 2019

Joint work with Ying Yuan and Peter Thall

Department of Biostatistics, The University of Texas MD Anderson Cancer Center

A Motivating Example

SMART to evaluate meal replacement for adolescent obesity reported by Berkowitz et al. (2010).

- Self-selected meal plans (CD, control) versus meal replacement (MR, active)
- 1:1 randomization at baseline
- 1:1 re-randomization of MR arm at 4 months to continue MR or switch to CD through 12 months
- Three regimes: MR+MR, MR+CD, CD+CD
- Outcome measures: BMI at 4 and 12 months
- Covariates: sex, race, parent BMI, baseline BMI, month 4 BMI

Aim: To identify a personalized dietary strategy that will minimize expected 12 month BMI

Dynamic Treatment Regime

Sequential decisions can be formalized as a dynamic treatment regime (DTR).

A DTR is a set of decision rules, one for each stage, that stipulate which treatment to assign (or dietary action to take) based on the patient's history at that stage.

Prior to the seminal papers by Murphy (2003) and Robins (2004), there was a dearth of statistical methods for evaluating DTRs.

In recent years, many approaches for defining, estimating and optimizing DTRs have been (and are still being) proposed.

Proposed Approach

The proposed approach bridges the gap between Bayesian inference and Q-learning (Watkins, 1989; Moodie et al., 2007).

- Provide another avenue for the use of hierarchical Bayesian modeling to optimize DTRs
- Attenuate inferential difficulties encountered by Q-learning and related methods

Some Notation

In the motivating setting we observe:

$$\boldsymbol{O}_1 \rightarrow A_1 \rightarrow O_2 \overset{A_1=MR}{\rightarrow} A_2 \rightarrow Y$$

- $\boldsymbol{O}_1 =$ sex, race, parent's BMI, baseline BMI
- $A_1 =$ CD or MR
- $O_2 =$ month 4 BMI
- for $A_1 =$ MR, $A_2 =$ continue MR or switch to CD
- $Y =$ month 12 BMI

where the sample data consists of n independent observations

Dynamic Treatment Regime

Let H_k denote a patient's history at stage k, e.g.,

- $H_2 = (\boldsymbol{O}_1, A_1, O_2)$
- $H_1 = \boldsymbol{O}_1$

Because H_k is observable at stage k, it can be used to select A_k.

A two-stage dynamic treatment regime (DTR) consists of two decision rules

$$d_k : \mathcal{H}_k \to \mathcal{A}_k, \ k = 1, 2$$

i.e., a person with history H_k gets A_k at stage k

Optimal Dynamic Treatment Regime

Assuming the aim is to maximize the expected payoff, following Bellman (1957), the optimal two-stage DTR is

$$d_2^{opt}(H_2) = \arg\max_{a_2 \in \mathcal{A}_2} E[Y \mid H_2, A_2 = a_2]$$

$$d_1^{opt}(H_1) = \arg\max_{a_1 \in \mathcal{A}_1} E\left[E\left[Y \,\middle|\, H_2(a_1), A_2 = d_2^{opt}(H_2(a_1))\right] \,\middle|\, H_1, A_1 = a_1\right]$$

Notice that d_1^{opt} depends on d_2^{opt}, but not conversely.

- Motivates backward induction, i.e., identify d_2^{opt} then d_1^{opt}

Q-Learning

In our example, additive Q-learning is implemented as follows:

Let $A_k = -1$ for CD and $A_k = 1$ for MR:

1. For all $i : a_{1,i} = 1$, assume

$$y_i = \boldsymbol{x}'_{2,0}(h_{2,i})\boldsymbol{\beta}_{2,0} + a_{2,i}\{\boldsymbol{x}'_{2,1}(h_{2,i})\boldsymbol{\beta}_{2,1}\} + \epsilon_{2,i}$$

   and estimate $\boldsymbol{\beta}_2 = (\boldsymbol{\beta}_{2,0}, \boldsymbol{\beta}_{2,1})$

2. $\widetilde{y}_i = \begin{cases} \boldsymbol{x}'_{2,0}(h_{2,i})\widehat{\beta}_{2,0} + |\boldsymbol{x}'_{2,1}(h_{2,i})\widehat{\beta}_{2,1}|, & a_{1,i} = 1 \\ y_i, & a_{1,i} = -1 \end{cases}$

3. For all i, assume

$$\widetilde{y}_i = \boldsymbol{x}'_{1,0}(h_{1,i})\boldsymbol{\beta}_{1,0} + a_{1,i}\{\boldsymbol{x}'_{1,1}(h_{1,i})\boldsymbol{\beta}_{1,1}\} + \epsilon_{1,i}$$

   and estimate $\boldsymbol{\beta}_1 = (\boldsymbol{\beta}_{1,0}, \boldsymbol{\beta}_{1,1})$

The estimated optimal DTR consists of the rules:

$$\widehat{d}_k^{opt}(h_k) = \text{sign}\{\boldsymbol{x}'_{k,1}(h_k)\widehat{\boldsymbol{\beta}}_{k,1}\},\ k = 1, 2$$

Q-Learning Limitations

Estimating the sampling distribution of $\widehat{\boldsymbol{\beta}}_1$ is difficult due to the dependence of $\widetilde{y}$ on $|\boldsymbol{x}'_{2,1}\widehat{\beta}_{2,1}|$ when $\boldsymbol{x}'_{2,1}(h_2)\widehat{\beta}_{2,1} = 0$ for some h_2, i.e., stage 2 intervention has no effect for some people (Moodie et al., 2012).

Correctly specifying the interaction between A_1 and $\boldsymbol{O}_1$ in the stage 2 model is critical.

The support of $\widetilde{y}$ and y do not match when y is a binary, multinomial, or count variable making implementation with gams difficult.

Proposed Approach

Our proposed approach relies on potential outcomes:

- $Y_i(a_1, a_2) =$ i-th subject's month 12 BMI under (a_1, a_2).
- $H_{2,i}(a_1) =$ i-th subject's month 4 history under action a_1.

Requires one Bayesian regression model per stage in reverse order:

1. Stage 2 regression model for all $i : a_{1,i} = 1$
   - Response: $Y_i(a_{1,i}, a_{2,i}) = y_{2,i}$
   - Covariates: $H_{2,i}(a_{1,i}) = h_{2,i}, a_{2,i}$
   - Parameter: $\boldsymbol{\theta}_2$
2. Stage 1 regression model for all i
   - Response: $Y_i(a_{1,i}, a_{2,i}^{opt})$ where $a_{2,i}^{opt} = d_2^{opt}(h_{2,i})$
   - Covariates: $h_{1,i}, a_{1,i}$
   - Parameter: $\boldsymbol{\theta}_1$

Stage 2 Posterior Distribution an Its Uses

Sampling from the posterior for $\boldsymbol{\theta}_2$ is accomplished in the usual manner

Induces posterior samples for $d_2^{opt}(H_2)$ and thus for

$$\boldsymbol{a}_2^{opt} = \{a_{2,i}^{opt} : a_{1,i} = 1, i = 1 \ldots, n\}$$

- For $a_{2,i}^{opt} = a_{2,i}$, the stage 1 response is $y_i = Y_i(a_{1,i}, a_{2,i})$ and thus observed.
- For $a_{2,i}^{opt} \neq a_{2,i}$, the stage 1 response is missing.

Given $\boldsymbol{a}_2^{opt}$, upon assuming a relationship between y_i and $Y_i(a_{1,i}, a_{2,i}^{opt})$ such as additive local rank preservation, we can determine the full conditional posterior predictive distribution for $\{Y_i(a_{1,i}, a_{2,i}^{opt}) : a_{2,i}^{opt} \neq a_{2,i}, a_{1,i} = 1, i = 1 \ldots, n\}$.

Stage 1 Posterior Distribution

Sampling from the posterior distribution for $\boldsymbol{\theta}_1$ is accomplished using Bayesian data augmentation

1. Draw $\boldsymbol{\theta}_2$ for its posterior distribution and determine $\boldsymbol{a}_2^{opt}$
2. For $a_{2,i} = a_{2,i}^{opt}$, set $y_{2,i}^{opt} = y_{2,i}$, whereas for $a_{2,i} \neq a_{2,i}^{opt}$
   - Draw $\{y_{2,i}^{opt} : a_{2,i}^{opt} \neq a_{2,i}, a_{1,i} = 1, i = 1 \ldots, n\}$ from its full conditional posterior predictive distribution
3. Draw $\boldsymbol{\theta}_1$ from its full conditional posterior distribution

Iterate the above steps to sample from the stage 1 posterior distribution.

Motivating Example Data Analysis

Implemented the proposed approach using Bayesian Additive Regression Trees (BART)

BART assumes a nonparametric mean function, and thus can identify higher-order interactions and non-linear associations.

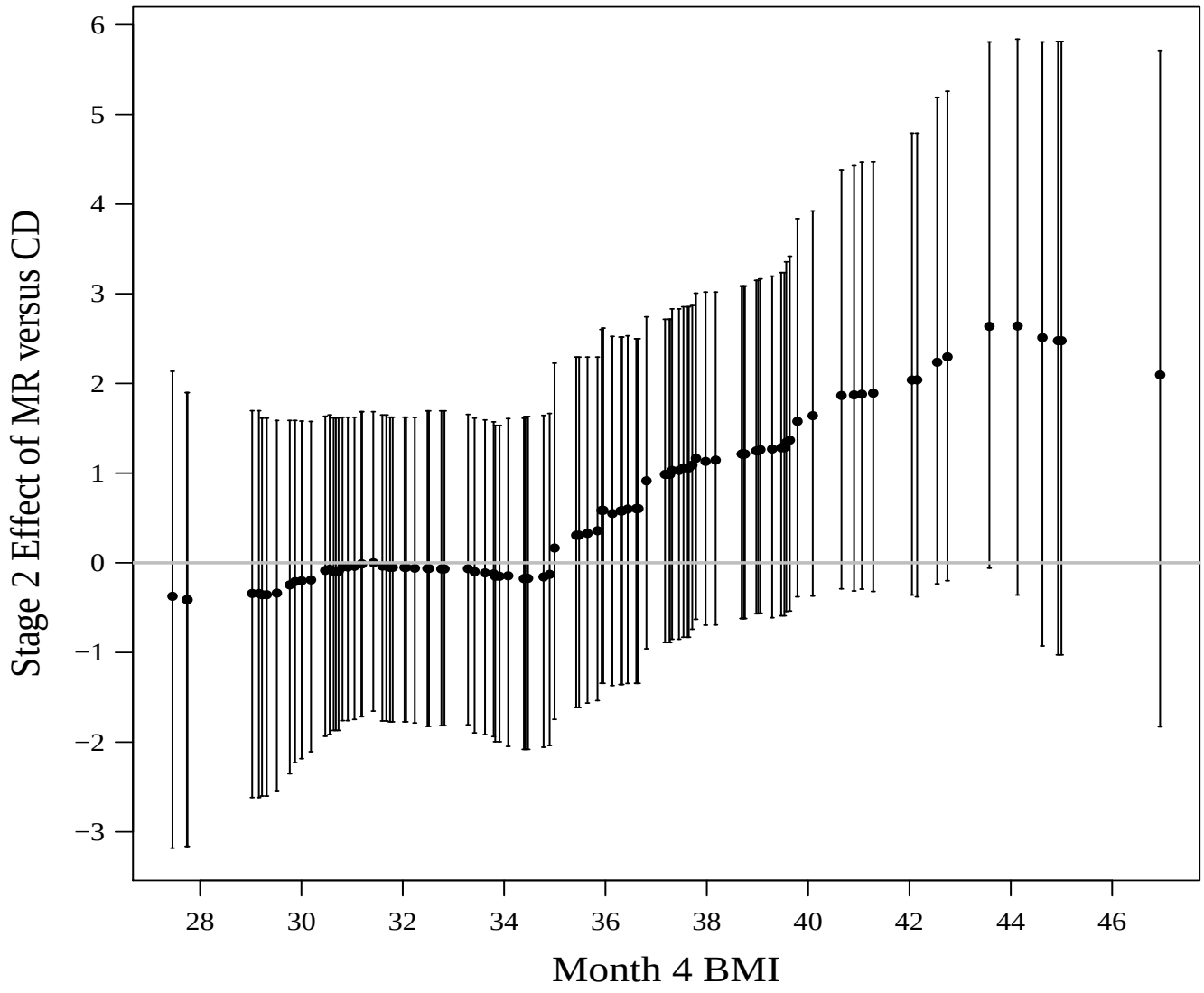
$$Y = \sum_{j=1}^{m} g(x; T_j, M_j) + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma^2),$$

where $g(x; T_j, M_j)$ is a regression tree with splitting rules (T_j) and terminal values (M_j).

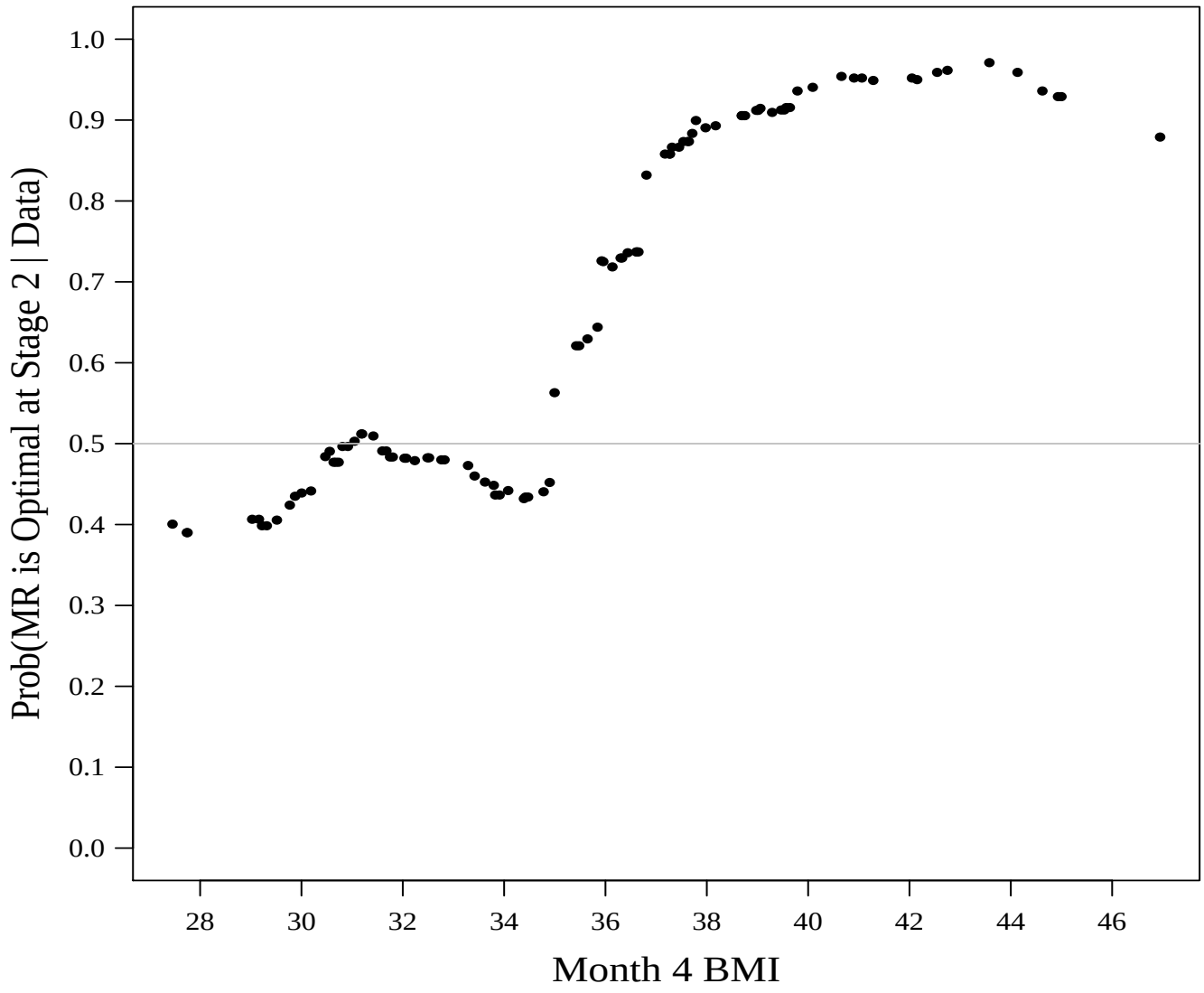
The mean of Y given x is the sum of the terminal values associated with x in the m trees.

We use the prior specification suggested by Chipman et al. (2010) for $(T_1, M_1), \ldots, (T_m, M_m)$ and σ, and carry out inference using the `R` package `BayesTree`.

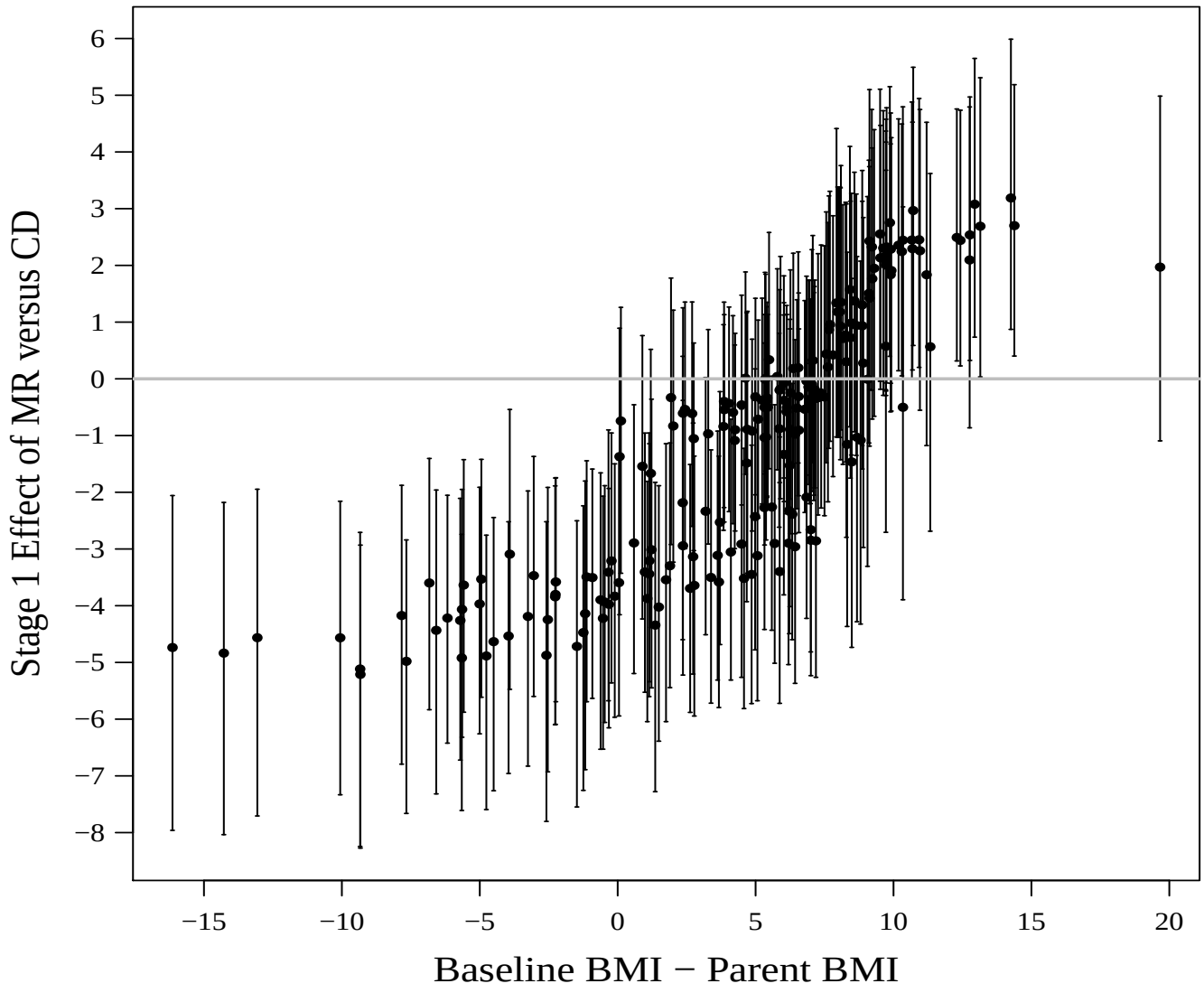
Stage 2 Treatment Effects



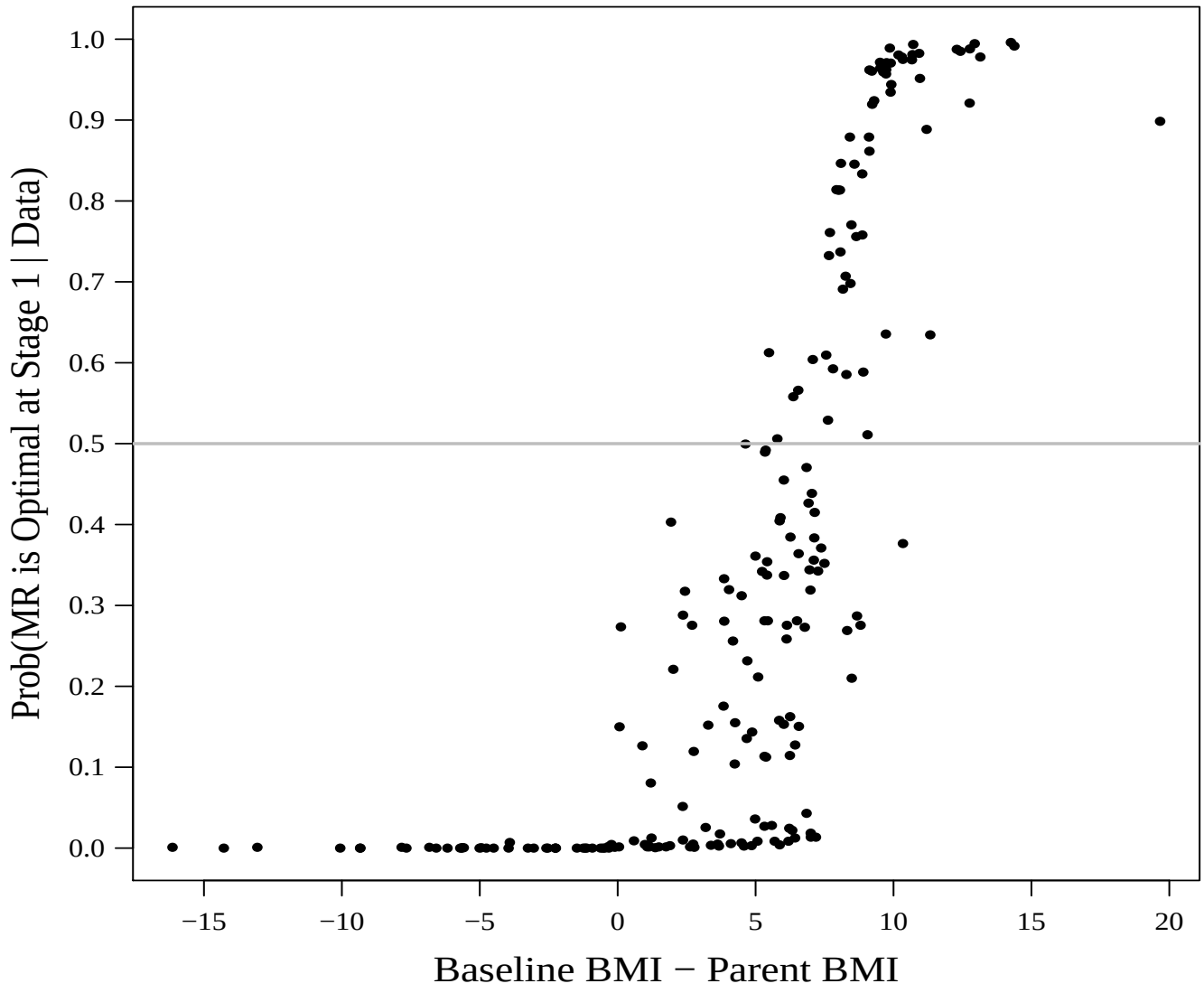
Stage 2 Posterior Optimality Probabilities



Stage 1 Treatment Effects



Stage 1 Posterior Optimality Probabilities



Non-regularity Simulation Study

$O_1 \to A_1 \to O_2 \to A_2 \to Y$ with $O_k, A_k \in \{-1, 1\}$ and $Y \in \mathbb{R}$.

Following Laber et al. (2014) and Chakraborty et al. (2013), we generated data as follows:

$$
\begin{aligned}
&\text{Prob}(O_1 = 1) = 0.5 \\
&\text{Prob}(A_1 = 1 \,|\, O_1) = 0.5 \\
&\text{Prob}(O_2 = 1 \,|\, \overline{A}_1) = \text{expit}\{\delta_1 O_1 + \delta_2 A_1\} \\
&\text{Prob}(A_2 = 1 \,|\, \overline{O}_2) = 0.5 \\
&Y = \alpha_0 + \alpha_1 O_1 + \alpha_2 A_1 + \alpha_3 O_1 \times A_1 + \alpha_4 O_2 + \\
&\qquad \alpha_5 A_2 + \alpha_6 A_2 \times O_1 + \alpha_7 A_2 \times A_1 + \alpha_8 A_2 \times O_2 + \epsilon,
\end{aligned}
$$

where $\epsilon \sim \text{Normal}(0, 1)$, and $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are specified in each case to exhibit varying degrees of non-regularity.

Continuous Payoff Implementations

To isolate the differences between the proposed approach and Q-learning, we implemented each method using the same linear stage 1 and stage 2 models.

- m-out-of-n bootstrap for variance estimation of stage 1 model parameters in Q-learning (Chakraborty et al., 2013)
- BIG sampler with $p(\boldsymbol{\beta}_k, \sigma_k) \propto 1/\sigma_k^2$
- We assume $Y_i(a_{1,i}, a_{2,i}^{opt})$ and Y_i have the same residual during imputation

Because direct sampling is feasible, the proposed method is 7 times faster than Q-learning method (when based on 2000 posterior samples vs 2000 bootstrap samples).

Partial Results for Stage 1 (1000 datasets with $n = 300$)

Proposed Approach

Case	Type	POA	Bias	RMSE	W95	C95
1	NR	1.000	0.062	0.135	0.611	0.968
2	NNR	1.000	0.052	0.131	0.611	0.972
6	R	0.999	-0.010	0.143	0.600	0.949
B	NR	0.984	0.030	0.141	0.606	0.957
C	NNR	0.982	0.026	0.139	0.606	0.961

Q-Learning

Case	Type	POA	Bias	RMSE	W95	C95
1	NR	1.000	0.089	0.148	0.603	0.963
2	NNR	1.000	0.080	0.143	0.603	0.965
6	R	1.000	-0.003	0.141	0.609	0.950
B	NR	0.979	0.044	0.147	0.608	0.955
C	NNR	0.975	0.040	0.144	0.609	0.957

Other Settings

Real-valued interim and subsequent payoffs:
$O_1 \rightarrow A_1 \rightarrow Y_1 \rightarrow A_2 \rightarrow Y_2$.

- Proposed approach based on Bayesian additive regression trees (Chipman et al., 2010)
- Q-learning based on generalized additive models

Binary payoff, initial responders do not continue:
$O_1 \rightarrow A_1 \rightarrow Y_1 \overset{Y_1=0}{\rightarrow} O_2 \rightarrow A_2 \rightarrow Y_2$.

- Subset of non-responders for stage 2 estimation
- Proposed approach based on probit BART
- $\widetilde{y} \in (0, 1)$, so Q-learning stage 1 estimation is based on a quasi-binomial regression model

Partial Results: Continuous Y_1 and Y_2

Linear associations (Stage 1)

Method	POA	Bias	RMSE	W95	C95
BML-GLM	0.943	0.000	0.181	0.743	0.951
QL-GLM	0.944	0.010	0.180	0.780	0.957
BML-BART	0.932	-0.039	0.283	1.484	0.988
QL-GAM	0.929	0.014	0.220	1.301	0.974

Nonlinear associations (Stage 1)

Method	POA	Bias	RMSE	W95	C95
BML-GLM	0.987	-0.333	0.417	0.354	0.319
QL-GLM	0.987	-0.320	0.406	0.366	0.307
BML-BART	0.989	-0.032	0.114	0.558	0.978
QL-GAM	0.992	-0.011	0.092	0.380	0.924

(Based on 1000 datasets with $n = 300$)

Partial Results: Binary Payoff

Linear associations (Stage 1)

Method	POA	Bias	RMSE	W95	C95
BML-GLM	0.858	0.001	0.050	0.197	0.933
QL-GLM	0.861	0.006	0.049	0.181	0.853
BML-BART	0.861	-0.010	0.056	0.315	0.993
QL-GAM	0.821	0.012	0.069	–	–

Nonlinear associations (Stage 1)

Method	POA	Bias	RMSE	W95	C95
BML-GLM	0.931	-0.084	0.105	0.267	0.788
QL-GLM	0.930	-0.075	0.097	0.239	0.674
BML-BART	0.924	-0.050	0.085	0.433	0.990
QL-GAM	0.891	-0.002	0.079	–	–

(Based on 1000 datasets with $n = 300$)

Conclusion

The proposed approach is a general framework that bridges the gap between Bayesian inference and Q-learning.

- Multiply imputes potential subsequent payoff under optimal actions at subsequent stages, as opposed to using a plug-in estimator.

BIG Sampler uses data augmentation to facilitate sampling from the stage 1 posterior.

Stage-wise Bayesian regression modeling for optimizing DTRs

- Minimizes modeling requirements
- Parametric models result in interpretable rules and parameters
- Characterizes uncertainty well in non-regular settings

Comments/Questions?

Thank You! **e-mail**: murra484@umn.edu

Selected References

- Murray T.A., Yuan Y., Thall P.F. (2017) "A Bayesian Machine Learning Approach for Optimizing Dynamic Treatment Regimes." *JASA* In Press.
- Berkowitz et al. (2010). "Meal replacements in the treatment of adolescent obesity: A randomized controlled trial." *Obesity* 19(6):11931199.
- Murphy, S.A. (2003) "Optimal dynamic treatment regimes." *JRSS-B* 65(2):331-355.
- Chakraborty, B., E.B. Laber, and Y. Zhao (2013) "Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme." *Biometrics* 69(3):714-723.
- Watkins, C. (1989) "Learning from delayed rewards." PhD Thesis, University of Cambridge, England.
- Moodie, E.E.M., T.S. Richardson, and D.A. Stephens (2007) "Demystifying optimal dynamic treatment regimes." *Biometrics* 63(2):447-455.