Restricted Sub-Tree Learning (ReST-L) to Estimate an Optimal Dynamic Treatment Regime Using Observational Data - Supplemental Content

Kelly Speth, Lu Wang

24-APR-2021

1 Additional Simulation Results

1.1 Simulation Studies to Evaluate the Relative Performance of ReST-L When Using Incorrectly-Specified Propensity Models

In order to demonstrate the sensitivity of our findings, when utilizing ReST-L we first evaluate ReST-L performance in estimation of a two-stage optimal DTR when the propensity model is incorrectly specified. The two-stage data generation specifications for this simulation experiment are the same as those introduced in Section 4.3 of the primary manuscript. In contrast to the analyses presented in Table 3 (primary manuscript) in which we assumed that the variables determining treatment were known, here we consider all variables in $\mathbf{H}$ as variables that may be used to define the treatment assignment mechanism at both stages. While we understand that the AIPW estimator is consistent and doubly robust in large samples when either or both the propensity model and the conditional mean model are correctly specified, we believe that this supplemental simulation study in which neither the propensity model nor the conditional mean model are correctly specified will reflect a scenario that is likely to occur frequently in practice and will shed light on the use of ReST-L as an out-of-the-box solution for estimating optimal DTRs.

ReST-L and T-RL performance using incorrectly-specified propensity models is presented in Table S-1. Because Q-Learning methods do not rely on a propensity model, performance measures for Q-Learning methods are replicated from Table 3 (primary manuscript) for ease of comparison with ReST-L and T-RL. For tree-type DTRs, performance of ReST-L is slightly lower overall when the propensity models are incorrectly specified compared with correct specification. For example, with a sample size of $n = 350$, a covariate correlation of $\rho = 0.2$, and $|\mathbf{H}| = 20$, the percent of the test set classified to the correct treatment is 84.8% when the propensity models are incorrectly specified and 86.0% when correctly specified. Similarly, when the sample size and the number of variables in $\mathbf{H}$ are large (i.e., $n = 1000$, $\rho = 0.2$, and $|\mathbf{H}| = 100$), the performance is 93.2% and 95.2% for incorrectly- and correctly- specified propensity models, respectively. The percentage of correctly-treated observations in the test set remains reasonable across all sample sizes and variable settings, hovering above 85% correct classification on average, and maintains an improvement over T-RL across all settings. With an underlying nontree-type DTR, larger sample sizes are necessary to achieve reasonable performance, as was also observed in Table 3 (primary manuscript). While ReST-L is still likely to be favored when the assumptions of the method are fulfilled, i.e., that

the true DTR is defined only in terms of a subset of variables $\mathbf{H}_{\text{sub}}$, the improvement of ReST-L over restricted nonparametric Q-Learning decreases as sample size increases. With $n = 1000$ and a correlation of $\rho = 0.2$, for example, the percent of observations with correct treatment classification is 86.7% for restricted nonparametric Q-Learning compared with 87.0% for ReST-L.

Although either the propensity model or the conditional mean model must be correctly specified in order to ensure consistency of the expected counterfactual outcome, our results suggest that correct specification of the propensity model is perhaps less critical. Or it may be that our propensity model was not sufficiently complicated to induce larger performance differences.

Table 1: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR (top panel) or nontree-type DTR (bottom panel) and assuming incorrectly-specified propensity models. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	rho	ReST-L		T-RL		Q-L-R		Q-L		Q-NP-R		Q-NP	
			%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
Tree-type DTR														
1000	100/20	0.2	93.2 (7.0)	7.8 (0.2)	90.6 (8.8)	7.7 (0.3)	51.0 (2.1)	5.9 (0.1)	45.1 (2.6)	5.5 (0.2)	88.9 (4.8)	7.7 (0.2)	75.0 (8.6)	7.1 (0.4)
1000	100/20	0.6	92.4 (6.8)	7.8 (0.2)	89.7 (9.6)	7.7 (0.3)	50.9 (2.3)	5.9 (0.2)	44.8 (2.7)	5.6 (0.2)	86.9 (4.7)	7.6 (0.2)	73.1 (6.6)	7.1 (0.3)
600	100/20	0.2	79.8 (19.3)	7.4 (0.6)	67.9 (19.4)	6.8 (0.8)	49.4 (2.5)	5.8 (0.2)	40.0 (2.9)	5.2 (0.2)	79.4 (6.4)	7.3 (0.3)	60.1 (9.6)	6.4 (0.4)
600	100/20	0.6	81.2 (19.2)	7.4 (0.6)	68.7 (18.8)	6.9 (0.7)	49.6 (2.3)	5.9 (0.2)	39.9 (2.9)	5.2 (0.2)	76.2 (7.2)	7.2 (0.3)	57.6 (8.2)	6.3 (0.4)
500	50/35	0.2	82.0 (15.0)	7.4 (0.5)	79.9 (16.1)	7.3 (0.5)	46.9 (2.9)	5.6 (0.2)	44.8 (2.6)	5.5 (0.2)	69.2 (8.8)	6.9 (0.4)	60.5 (9.0)	6.4 (0.4)
500	50/35	0.6	82.3 (12.5)	7.4 (0.4)	78.8 (14.7)	7.3 (0.4)	47.1 (2.9)	5.7 (0.2)	45.1 (2.9)	5.6 (0.2)	67.0 (8.7)	6.9 (0.3)	57.6 (8.3)	6.4 (0.4)
500	50/10	0.2	86.5 (13.1)	7.6 (0.4)	78.2 (16.3)	7.3 (0.6)	50.8 (2.4)	5.9 (0.1)	44.8 (3.0)	5.5 (0.2)	79.8 (6.1)	7.4 (0.2)	60.9 (10.1)	6.5 (0.4)
500	50/10	0.6	86.5 (12.3)	7.6 (0.3)	78.0 (15.0)	7.3 (0.5)	50.8 (2.6)	5.9 (0.2)	44.7 (3.2)	5.6 (0.2)	76.2 (5.7)	7.2 (0.2)	57.9 (8.0)	6.4 (0.4)
350	20/7	0.2	84.8 (15.2)	7.5 (0.4)	79.2 (17.5)	7.3 (0.5)	51.0 (2.4)	5.9 (0.2)	48.4 (2.8)	5.7 (0.2)	75.0 (7.0)	7.2 (0.3)	59.6 (9.1)	6.4 (0.5)
350	20/7	0.6	84.3 (12.7)	7.5 (0.3)	78.8 (15.0)	7.4 (0.4)	50.9 (2.6)	5.9 (0.2)	48.1 (2.6)	5.8 (0.2)	72.0 (6.4)	7.1 (0.3)	57.5 (7.8)	6.4 (0.4)
Nontree-type DTR														
1000	100/20	0.2	87.0 (34.1)	7.7 (1.2)	78.2 (32.0)	7.4 (1.3)	24.3 (2.8)	4.9 (0.2)	18.3 (2.3)	4.4 (0.2)	86.7 (4.6)	7.7 (0.1)	69.8 (11.4)	7.0 (0.5)
1000	100/20	0.6	88.5 (32.2)	7.7 (1.2)	80.7 (29.8)	7.4 (1.2)	25.0 (2.6)	5.0 (0.2)	18.4 (2.0)	4.5 (0.1)	82.0 (5.4)	7.5 (0.2)	61.2 (9.3)	6.7 (0.4)
600	100/20	0.2	59.3 (33.4)	6.6 (1.1)	42.8 (21.6)	6.0 (0.9)	22.0 (3.1)	4.7 (0.2)	15.5 (2.4)	4.2 (0.2)	71.6 (7.4)	7.2 (0.3)	45.4 (10.9)	6.0 (0.5)
600	100/20	0.6	59.6 (31.5)	6.6 (1.2)	46.4 (20.6)	6.1 (0.9)	22.5 (3.3)	4.8 (0.2)	15.6 (2.3)	4.2 (0.2)	66.3 (8.1)	7.0 (0.3)	38.4 (10.5)	5.8 (0.5)
500	50/35	0.2	62.4 (32.8)	6.8 (1.1)	59.1 (30.8)	6.7 (1.1)	19.2 (2.8)	4.5 (0.2)	18.2 (2.7)	4.4 (0.2)	54.4 (9.8)	6.5 (0.4)	43.5 (10.9)	6.0 (0.5)
500	50/35	0.6	63.6 (28.3)	6.9 (1.1)	58.8 (29.2)	6.6 (1.1)	19.6 (2.8)	4.6 (0.2)	18.3 (2.7)	4.5 (0.2)	50.6 (10.3)	6.4 (0.4)	37.7 (10.3)	5.8 (0.5)
500	50/10	0.2	72.6 (33.2)	7.3 (1.2)	58.1 (31.4)	6.6 (1.1)	23.9 (3.4)	4.9 (0.2)	18.1 (2.8)	4.4 (0.2)	72.1 (7.6)	7.2 (0.3)	43.6 (10.5)	6.0 (0.5)
500	50/10	0.6	74.0 (30.6)	7.3 (1.1)	61.0 (26.4)	6.7 (1.0)	24.6 (3.3)	5.0 (0.2)	18.2 (2.7)	4.4 (0.2)	66.3 (7.7)	7.0 (0.3)	37.0 (9.4)	5.8 (0.4)
350	20/7	0.2	68.4 (33.6)	7.1 (1.1)	63.5 (31.0)	7.0 (1.1)	23.9 (4.1)	4.9 (0.3)	20.8 (3.7)	4.6 (0.2)	64.2 (8.1)	6.9 (0.3)	41.2 (9.7)	5.9 (0.4)
350	20/7	0.6	71.0 (30.0)	7.2 (1.1)	64.8 (29.3)	6.9 (1.1)	24.3 (4.2)	5.0 (0.3)	20.7 (3.6)	4.7 (0.2)	60.1 (8.3)	6.8 (0.3)	36.8 (8.7)	5.7 (0.4)

1.2 Simulation Studies To Evaluate the Relative Performance of ReST-L in the Absence of Confounding by Variables in $\mathbf{H}^C_{\mathbf{sub}}$

In this supplemental simulation study we evaluate the relative performance of ReST-L in the absence of confounding by variables in $\mathbf{H}^C_{\text{sub}}$, i.e., the true data-generating model for treatment $\mathbf{A}$ and outcomes $\mathbf{Y}$ include only variables in $\mathbf{H}_{\text{sub}}$. This is in contrast to the simulation experiments presented in Section 4.3 (primary manuscript) in which confounding by variables in $\mathbf{H}^C_{\text{sub}}$ exist. Using the data generating mechanisms presented in Section 4.3 (primary manuscript), for this simulation experiment we replace variables X_{C1} and X_{C2}, the first two variables in $\mathbf{H}^C_{\text{sub}}$, with the final two variables in $\mathbf{H}_{\text{sub}}$. As in all previously-reported simulation experiments and by ReST-L assumption, the optimal regimes $\mathbf{g}^{\text{opt}}$ are defined using variables only in $\mathbf{H}_{\text{sub}}$.

As can be seen in Table S-2, performance for both ReST-L and T-RL are similar to those reported in Table 3 (primary manuscript). Performance for Q-Learning methods, with the exception of restricted nonparametric Q-Learning, are also similar. Only restricted nonparametric Q-Learning demonstrates a slightly lower performance in this setting. For a tree-type DTR, for example, we observe 88.9% correct treatment classification in Table 3 (primary manuscript) with a sample size $n = 1000$, $|\mathbf{H}| = 100$ variables and $\rho = 0.2$, compared with 85.5% correct classification when the true treatment allocation $\mathbf{A}$, outcomes $\mathbf{Y}$, and the optimal DTR $\mathbf{g}^{\text{opt}}$ are defined using variables only in $\mathbf{H}_{\text{sub}}$.

These results suggest that, when the clinical question substantiates a restricted set of covariates to consider as tailoring variables, the application of ReST-L over existing methods is warranted also when the level of confounding among variables not considered as candidate tailoring variables is low.

Table 2: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR (top panel) or nontree-type DTR (bottom panel) with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	rho	ReST-L		T-RL		Q-L-R		Q-L		Q-NP-R		Q-NP	
			%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
Tree-type DTR														
1000	100/20	0.2	94.9 (5.5)	7.8 (0.2)	92.5 (6.5)	7.7 (0.3)	50.9 (2.4)	5.9 (0.2)	45.3 (2.5)	5.5 (0.2)	85.5 (5.6)	7.5 (0.2)	73.6 (7.6)	7.1 (0.3)
1000	100/20	0.6	95.3 (5.6)	7.8 (0.2)	92.6 (6.8)	7.7 (0.3)	51.1 (2.3)	5.9 (0.1)	45.5 (2.6)	5.5 (0.2)	82.5 (6.1)	7.4 (0.2)	70.0 (8.0)	6.9 (0.3)
600	100/20	0.2	89.4 (10.1)	7.6 (0.3)	77.0 (18.4)	7.2 (0.7)	49.6 (2.5)	5.8 (0.2)	40.2 (2.8)	5.2 (0.2)	75.4 (7.6)	7.1 (0.3)	60.0 (10.8)	6.4 (0.5)
600	100/20	0.6	89.2 (10.7)	7.6 (0.3)	77.5 (17.5)	7.2 (0.6)	49.8 (2.5)	5.8 (0.2)	40.7 (3.1)	5.2 (0.2)	72.2 (7.7)	7.0 (0.4)	54.8 (11.0)	6.2 (0.5)
500	50/35	0.2	84.9 (14.1)	7.5 (0.4)	83.4 (15.6)	7.5 (0.4)	47.2 (2.7)	5.6 (0.2)	45.1 (2.7)	5.5 (0.2)	65.3 (9.4)	6.7 (0.4)	61.2 (9.8)	6.5 (0.4)
500	50/35	0.6	85.6 (13.0)	7.5 (0.4)	82.7 (15.0)	7.4 (0.4)	47.6 (2.9)	5.7 (0.2)	45.4 (2.8)	5.5 (0.2)	61.5 (8.5)	6.5 (0.4)	56.8 (9.7)	6.3 (0.4)
500	50/10	0.2	89.9 (10.9)	7.7 (0.3)	83.4 (15.5)	7.4 (0.5)	50.8 (2.6)	5.9 (0.2)	45.1 (3.0)	5.5 (0.2)	78.3 (7.7)	7.3 (0.3)	61.2 (9.4)	6.4 (0.4)
500	50/10	0.6	89.9 (10.3)	7.7 (0.3)	83.7 (14.2)	7.5 (0.5)	50.7 (2.7)	5.9 (0.1)	45.3 (3.0)	5.5 (0.2)	74.6 (7.3)	7.1 (0.3)	57.0 (9.1)	6.3 (0.5)
350	20/7	0.2	85.3 (13.2)	7.5 (0.4)	81.3 (15.6)	7.4 (0.5)	50.7 (2.4)	5.9 (0.2)	48.2 (2.7)	5.7 (0.2)	74.5 (7.8)	7.1 (0.3)	60.6 (10.2)	6.5 (0.5)
350	20/7	0.6	86.7 (13.0)	7.6 (0.3)	82.2 (15.5)	7.4 (0.4)	51.2 (2.7)	5.9 (0.2)	48.8 (3.0)	5.7 (0.2)	72.0 (6.7)	7.0 (0.3)	58.6 (8.3)	6.4 (0.4)
Nontree-type DTR														
1000	100/20	0.2	93.3 (33.5)	7.8 (1.2)	85.7 (35.7)	7.6 (1.2)	25.2 (2.7)	4.9 (0.2)	18.4 (2.0)	4.4 (0.1)	81.6 (7.3)	7.5 (0.3)	69.4 (12.2)	7.0 (0.5)
1000	100/20	0.6	93.6 (35.0)	7.8 (1.2)	86.4 (35.3)	7.7 (1.2)	25.2 (2.6)	4.9 (0.2)	18.3 (2.1)	4.4 (0.2)	75.8 (7.5)	7.2 (0.3)	60.7 (8.9)	6.6 (0.4)
600	100/20	0.2	67.6 (42.6)	7.2 (1.2)	52.1 (26.3)	6.4 (1.0)	22.7 (3.1)	4.7 (0.2)	15.6 (2.4)	4.2 (0.2)	64.8 (9.8)	6.8 (0.4)	46.7 (12.2)	6.1 (0.6)
600	100/20	0.6	77.1 (38.8)	7.4 (1.3)	54.8 (28.8)	6.5 (1.0)	22.8 (3.0)	4.7 (0.2)	15.6 (2.1)	4.2 (0.2)	59.8 (8.5)	6.7 (0.4)	40.9 (11.4)	5.8 (0.6)
500	50/35	0.2	65.7 (37.3)	7.0 (1.1)	63.3 (34.8)	6.8 (1.1)	19.6 (3.0)	4.5 (0.2)	18.2 (2.7)	4.4 (0.2)	49.6 (10.5)	6.2 (0.5)	44.9 (11.3)	6.0 (0.5)
500	50/35	0.6	68.4 (32.3)	7.1 (1.1)	64.8 (33.8)	7.0 (1.1)	19.9 (2.9)	4.5 (0.2)	18.4 (2.9)	4.4 (0.2)	46.6 (10.5)	6.1 (0.4)	39.9 (10.8)	5.8 (0.5)
500	50/10	0.2	78.4 (37.1)	7.4 (1.2)	62.6 (35.8)	6.8 (1.1)	24.7 (3.6)	4.9 (0.3)	18.0 (2.8)	4.4 (0.2)	67.5 (7.8)	7.0 (0.3)	43.6 (12.3)	6.0 (0.5)
500	50/10	0.6	78.2 (33.0)	7.5 (1.2)	66.9 (32.5)	7.1 (1.1)	24.9 (3.3)	4.9 (0.2)	18.5 (2.7)	4.4 (0.2)	63.1 (7.3)	6.8 (0.3)	40.8 (10.1)	5.8 (0.5)
350	20/7	0.2	64.6 (40.0)	7.0 (1.2)	59.8 (34.7)	6.7 (1.2)	24.4 (3.9)	4.9 (0.3)	20.7 (3.2)	4.6 (0.2)	61.7 (9.3)	6.8 (0.3)	43.0 (9.7)	6.0 (0.4)
350	20/7	0.6	65.4 (38.1)	7.2 (1.3)	62.5 (33.3)	6.9 (1.2)	24.8 (3.9)	4.9 (0.3)	21.2 (3.2)	4.7 (0.2)	59.0 (8.8)	6.7 (0.3)	40.8 (8.7)	5.9 (0.4)

1.3 Simulation Studies to Evaluate the Relative Performance of ReST-L Under Violations of ReST-L Assumptions

ReST-L is an analytic solution that can be used when the investigators are relatively certain about the set of covariates that can be considered in the optimal dynamic treatment regime. For scenarios in which the optimal DTR is actually defined, at least in part, by variables in $\mathbf{H}_{\text{sub}}^{C}$, we would expect the performance of ReST-L to be lower than its unrestricted counterpart, T-RL. Although this will be the case overall, here we present an illustration of performance differences that may be observed under violations of the ReST-L assumption that only variables in $\mathbf{H}_{\text{sub}}$ are involved in an optimal DTR. Specifically, using the data generating mechanisms presented in Section 4.3 (primary manuscript) for both tree- and nontree-type DTRs, we modify g_1^{opt} to be defined using X_{C1} rather than X_1. Additionally, g_2^{opt} is now defined using the final variable in $\mathbf{H}_{\text{sub}}^{C}$ rather than on X_3. (Recall that X_{C1} refers to the first variable in the set of $\mathbf{H}_{\text{sub}}^{C}$.) Otherwise, similar to the data generating mechanisms used in Section 4.3 (primary manuscript), confounding is introduced using variables in both $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^{C}$.

As can be seen in Table S-3 for a tree-type DTR, performance of ReST-L is lower than that of T-RL across all data generation settings. With larger sample sizes, performance of ReST-L reaches about 75% of observations correctly classified to their optimal treatment. T-RL, on the other hand, achieves more than 90% correct treatment classification, which is similar to that reported in Table 3 (primary manuscript). With a nontree-type DTR structure, we observe that performance of ReST-L is poor across all data settings. Performance for T-RL is mediocre for lower sample sizes, as well, but is similar to the performance presented in Table 3 (primary manuscript).

The significance of these simulation results are two-fold. First, all analysis methods rely upon some set of assumptions. One of the foundational assumptions of ReST-L is that the investigator has sufficient justification to exclude certain variables from consideration as candidate tailoring variables. If this assumption is not met, an alternative analysis method should be considered. Secondly, as we have seen in simulation studies, ReST-L can provide additional power to correctly estimate the optimal multi-stage DTR with a smaller sample size, when the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ is lower, or when the level of confounding of the outcome is high (refer to Appendix 1.4). If the sample size obtained to address a specific research question is large, then the benefits of using ReST-L instead of T-RL may be negligible.

Table 3: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR (top panel) or nontree-type DTR (bottom panel) with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$ and $\mathbf{H}_{\text{sub}}^C$. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Q-L-R = Restricted Linear Q-Learning; Q-L = Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; Q-NP = Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	rho	ReST-L		T-RL		Q-L-R		Q-L		Q-NP-R		Q-NP	
			%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
Tree-type DTR														
1000	100/20	0.2	73.6 (3.2)	7.2 (0.2)	92.2 (4.7)	7.8 (0.2)	47.6 (2.2)	5.7 (0.2)	44.8 (2.4)	5.4 (0.2)	67.7 (4.3)	7.0 (0.2)	64.6 (11.0)	6.7 (0.5)
1000	100/20	0.6	73.4 (2.9)	7.2 (0.2)	92.0 (6.1)	7.8 (0.2)	48.5 (2.3)	5.8 (0.1)	44.8 (2.6)	5.5 (0.2)	67.3 (4.3)	6.9 (0.2)	64.7 (8.0)	6.7 (0.3)
600	100/20	0.2	69.8 (8.7)	7.1 (0.3)	81.1 (10.7)	7.4 (0.5)	46.1 (2.5)	5.6 (0.2)	39.8 (2.9)	5.1 (0.2)	61.8 (6.9)	6.7 (0.3)	50.1 (13.9)	6.0 (0.6)
600	100/20	0.6	69.1 (7.4)	7.0 (0.3)	79.8 (11.3)	7.3 (0.5)	47.0 (2.7)	5.7 (0.2)	39.3 (3.0)	5.2 (0.2)	60.9 (6.3)	6.6 (0.3)	50.1 (9.9)	5.9 (0.5)
500	50/35	0.2	65.5 (9.3)	6.9 (0.3)	85.0 (8.9)	7.5 (0.3)	43.2 (2.9)	5.4 (0.2)	44.7 (3.1)	5.4 (0.2)	53.4 (10.2)	6.2 (0.5)	53.5 (10.7)	6.1 (0.5)
500	50/35	0.6	64.8 (10.1)	6.9 (0.3)	84.3 (9.1)	7.5 (0.4)	44.0 (3.0)	5.5 (0.2)	44.6 (3.0)	5.5 (0.2)	54.1 (8.2)	6.2 (0.4)	52.2 (9.8)	6.0 (0.5)
500	50/10	0.2	69.6 (8.8)	7.1 (0.3)	84.7 (8.6)	7.5 (0.3)	47.2 (2.6)	5.7 (0.2)	44.7 (3.0)	5.4 (0.2)	63.1 (5.9)	6.7 (0.3)	52.6 (12.0)	6.1 (0.5)
500	50/10	0.6	69.2 (9.1)	7.0 (0.2)	84.2 (9.6)	7.5 (0.3)	48.5 (2.8)	5.8 (0.2)	44.8 (3.2)	5.5 (0.2)	62.4 (5.5)	6.7 (0.3)	51.9 (8.8)	6.0 (0.4)
350	20/7	0.2	66.7 (10.9)	7.0 (0.3)	83.5 (9.5)	7.5 (0.4)	47.5 (2.6)	5.7 (0.2)	48.0 (2.7)	5.6 (0.2)	60.7 (5.7)	6.6 (0.3)	56.8 (8.5)	6.3 (0.5)
350	20/7	0.6	66.1 (9.9)	7.0 (0.3)	83.1 (10.0)	7.5 (0.3)	48.6 (2.9)	5.8 (0.2)	48.2 (2.9)	5.7 (0.2)	60.1 (6.0)	6.6 (0.3)	54.5 (7.6)	6.2 (0.4)
Nontree-type DTR														
1000	100/20	0.2	50.4 (10.6)	6.7 (0.3)	64.7 (44.6)	6.9 (1.3)	24.8 (2.7)	5.0 (0.2)	18.6 (2.2)	4.5 (0.2)	46.9 (3.5)	6.6 (0.2)	58.3 (17.9)	6.6 (0.6)
1000	100/20	0.6	48.2 (10.8)	6.6 (0.4)	90.0 (33.9)	7.6 (1.3)	25.3 (2.7)	5.0 (0.2)	18.7 (2.4)	4.5 (0.2)	44.7 (3.4)	6.4 (0.2)	48.1 (12.6)	6.2 (0.5)
600	100/20	0.2	44.3 (14.2)	6.4 (0.9)	43.1 (20.2)	6.1 (0.8)	22.3 (3.1)	4.8 (0.2)	15.8 (2.3)	4.2 (0.2)	40.2 (4.3)	6.2 (0.3)	32.1 (11.7)	5.5 (0.6)
600	100/20	0.6	43.3 (14.2)	6.4 (0.9)	47.8 (24.4)	6.2 (0.8)	22.8 (3.4)	4.9 (0.2)	15.8 (2.4)	4.2 (0.2)	38.3 (5.6)	6.1 (0.3)	27.9 (9.3)	5.3 (0.5)
500	50/35	0.2	43.0 (14.0)	6.4 (0.9)	53.0 (37.3)	6.5 (1.1)	19.8 (2.8)	4.6 (0.2)	18.8 (2.5)	4.5 (0.2)	32.6 (6.0)	5.7 (0.3)	32.4 (10.8)	5.6 (0.5)
500	50/35	0.6	42.5 (13.7)	6.3 (0.9)	62.2 (37.3)	6.6 (1.1)	20.1 (3.1)	4.6 (0.2)	18.6 (2.8)	4.5 (0.2)	31.9 (5.7)	5.7 (0.3)	29.2 (9.8)	5.4 (0.5)
500	50/10	0.2	45.3 (12.1)	6.5 (0.5)	52.4 (37.4)	6.5 (1.1)	24.5 (3.3)	5.0 (0.2)	18.4 (2.6)	4.5 (0.2)	41.1 (4.0)	6.2 (0.2)	32.4 (11.7)	5.6 (0.5)
500	50/10	0.6	44.0 (13.7)	6.5 (0.9)	60.5 (35.9)	6.6 (1.1)	25.1 (3.4)	5.0 (0.2)	18.4 (2.5)	4.5 (0.2)	39.0 (4.3)	6.1 (0.2)	28.6 (8.9)	5.4 (0.4)
350	20/7	0.2	44.5 (12.5)	6.4 (0.7)	54.9 (36.6)	6.6 (1.1)	24.6 (4.0)	5.0 (0.3)	21.2 (3.3)	4.7 (0.2)	38.8 (5.1)	6.1 (0.3)	34.2 (10.2)	5.6 (0.4)
350	20/7	0.6	43.4 (14.3)	6.4 (0.9)	58.8 (36.3)	6.6 (1.2)	24.9 (3.7)	5.0 (0.3)	21.0 (3.4)	4.7 (0.3)	37.2 (4.9)	6.0 (0.2)	30.5 (7.7)	5.5 (0.3)

1.4 Simulation Studies to Evaluate the Relative Performance of ReST-L Under a Data Generation Mechanism with a High Degree of Confounding

Here we present comprehensive results for the two-stage simulation experiment introduced in Section 4.1 (primary manuscript), which demonstrated the bias of Naive T-RL in estimating the counterfactual mean outcome if all patients were to receive treatment according to their estimated optimal DTR.

Parameters varied across this simulation study include the sample size, with fixed sample sizes $n = 500, n = 1000$, and $n = 2000$, the number of covariates in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$ (|H|/|Hsub| = $20/7, 50/10, 50/35, 100/20$), the correlation used to generate the correlation matrix for covariates in $\mathbf{H}$ ($\rho = 0, 0.2, 0.6$), and the true, underlying structure of the DTR (i.e., tree or nontree-type). As discussed in the primary manuscript, the data generating mechanism includes a binary variable Z that has a strong confounding relationship with both the treatment assignment mechanisms for A_1 and A_2. This differs from the simulation experiments presented in Sections 4.2 and 4.3 (primary manuscript) in which the confounding relationship is defined using only continuous covariates.

The specific data generating mechanisms are as follows: Covariate data with dimension n x $|\mathbf{H}|$, where $|\mathbf{H}|$ refers to the cardinality of the vector of covariates collected for each individual, are generated using the multivariate normal distribution with a mean of $\mathbf{0}_{|\mathbf{H}|}$ and an autoregressive (AR1) correlation structure with specified ρ, but with the following modifications: Pairwise correlation between the first variable in $\mathbf{H}_{\text{sub}}$ and the first three variables in $\mathbf{H}^C_{\text{sub}}$ is equal to ρ and pairwise correlations between the first three variables in $\mathbf{H}_{\text{sub}} = 0$. This modification was intended to reflect quasi-real world complexities among covariates but specifying a moderate or high degree of correlation between the variables involved in the optimal DTR and confounding variables. An additional covariate, $Z \in \mathbf{H}^C_{\text{sub}}$, was generated for each observation using a Bernoulli distribution with $p = 0.4$. The actual treatment received A_1 is randomly generated from the multinomial distribution with probabilities $\pi_{10}, \pi_{11}, \pi_{12}$ where,

- $\pi_{10} = 1 - \pi_{11} - \pi_{12}$
- $\pi_{11} = \exp(0.5X_{C1} - 0.5X_1 + Z - 0.5)/[1 + \exp(0.5X_{C1} - 0.5X_1 + Z - 0.5) + \exp(0.5X_{C2} + 0.5X_1 - Z)]$
- $\pi_{12} = \exp(0.5X_{C2} + 0.5X_1 - Z)/[1 + \exp(0.5X_{C1} - 0.5X_1 + Z - 0.5) + \exp(0.5X_{C2} + 0.5X_1 - Z)]$

X_{C1}, X_{C2} represent the first two covariates in $\mathbf{H}^C_{\text{sub}}$, i.e., confounding variables not considered as candidate tailoring variables. The intermediate outcome following the first stage is defined as $Y_1 = \exp\{1.5 + 0.3X_{C1} - 1.5Z - |1.5X_1 - 2| \cdot (A_1 - g_1^{\text{opt}})^2\} + \epsilon$, where $\epsilon \sim N(0, 1)$. This reflects an unequal penalty dependent on the value of X_1–a variable used in the true optimal treatment regime–if the patient was not treated according to their optimal therapy, which adds an additional degree of complexity reflective of a real world setting into the data generating scenario. The optimal decision rules for the first stage are as follows:

- Tree-type DTR: If $X_1 > -1$ & $X_2 > 0.25$, then $g_1^{\text{opt}} = 2$; if $X_1 > -1$ & $-0.5 < X_2 \leq 0.25$, then $g_1^{\text{opt}} = 1$; otherwise, $g_1^{\text{opt}} = 0$
- Nontree-type DTR: $g_1^{\text{opt}} = \mathcal{I}[\{\log_2(|X_1| + 1) \leq 2\}$ & $(X_2 < -0.25)] + \mathcal{I}(X_2^2 > 0.35)$

Data for the treatment assignment for the second stage, A_2, are generated randomly also using the multinomial distribution, but with probabilities $\pi_{20}, \pi_{21}, \pi_{22}$, where:

- $\pi_{20} = 1 - \pi_{21} - \pi_{22}$
- $\pi_{21} = \exp(0.2Y_1 + 0.5 - Z)/[1.5 + \exp(0.2Y_1 + 0.5 - Z) + \exp(0.5X_{C2} + Z)]$
- $\pi_{12} = \exp(0.5X_{C2} + Z)/[1.5 + \exp(0.2Y_1 + 0.5 - Z) + \exp(0.5X_{C2} + Z)]$

We define the intermediate outcome following the second stage as $Y_2 = \exp\{1.18 + 0.2X_{C2} - 2Z - |1.5X_3 + 2| \cdot (A_2 - g_2^{\text{opt}})^2\} + \epsilon$, and the final outcome Y is defined as the sum of the stage-specific intermediate outcomes, i.e., $Y = Y_1 + Y_2$. The true, second-stage optimal decision rules are defined as follows:

- Tree-type: If $Y_1 > 0.5$ & $X_3 > 0$, then $g_2^{\text{opt}} = 2$; if $Y_1 > 0.5$ & $-1 < X_3 \leq 0$, then $g_2^{\text{opt}} = 1$; otherwise, $g_2^{\text{opt}} = 0$
- Nontree-type: $g_2^{\text{opt}} = \mathcal{I}\{(|X_3| > 0.6)\ \&\ (Y_1 > 1)\} + \mathcal{I}(Y_1^2 > 3)$

For the analysis we assume that there is an additive linear relationship between the intermediate outcomes and covariate or treatment history; for ReST-L the assumed model includes all observed covariates in $\mathbf{H}$ and a treatment interaction with the subset of candidate tailoring variables in $\mathbf{H}_{\text{sub}}$ whereas for Naive T-RL the assumed model includes only those observed covariates in $\mathbf{H}_{\text{sub}}$ with a corresponding treatment interaction. As in the previously-described simulations and consistent with the assumptions required by this method, we assume that only variables in $\mathbf{H}_{\text{sub}}$ may be included in an optimal DTR; variables from either $\mathbf{H}_{\text{sub}}$ or $\mathbf{H}_{\text{sub}}^C$, however, may define the intermediate outcomes and the treatment assignment mechanisms. We further assume that the propensity models used in ReST-L and T-RL are correctly specified. Under optimal treatment allocation and assuming independence across covariates, $\hat{E}[Y^*\{g^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 5.4$.

Results for the simulation studies are presented in Table S-4 for a tree-type DTR and in Table S-5 for a nontree-type DTR. With a tree-type DTR and a strong confounding relationship of covariate Z with the treatment assignment and outcomes, a larger sample size than that needed in the analyses presented in Sections 4.2 and 4.3 (primary manuscript) is required to achieve reasonable performance. With a correlation of $\rho = 0.2$ and $|\mathbf{H}| = 20$, a sample size of $n = 1000$ is required to achieve similar levels of performance to a sample size of $n = 350$ in Table 3 (primary manuscript). With a sample size of $n = 2000$, performance across all data generating settings reaches about 90% correct classification and the ReST-L results for $n = 2000$ are comparable to those of T-RL. For smaller sample sizes, however, we observe that ReST-L improves upon T-RL across all data settings, although the improvement is most apparent with a larger number of covariates and when the proportion of variables in $\mathbf{H}_{\text{sub}}$ relative to $\mathbf{H}$ is lower. Consider, for example, a sample size of $n = 1000$ with $|\mathbf{H}| = 100$ and $\rho = 0.2$ in which we report an estimated 82% correct classification for ReST-L compared with 73% for T-RL. With a nontree-type DTR structure, performance is slightly lower across all settings for both ReST-L and T-RL than with a tree-type DTR, but both ReST-L and T-RL achieve about 88-90% correct classification with $n = 2000$.

These results suggest that, with a high degree of confounding – particularly the confounding that may occur with a categorical variable – a larger sample size is needed to achieve the same estimated performance as with a lower degree of confounding. However, when the assumptions of ReST-L are met, ReST-L delivers improved performance with a smaller sample size over competing methods.

Table 4: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments based on an underlying, tree-type DTR and assuming a large confounding effect of a binary variable. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; Q-Linear-R = Restricted Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	ρ	ReST-L		T-RL		Q-Linear-R		Q-Linear		Q-NP-R		Q-NP	
		%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
n = 500													
20/7	0	71.4 (18.7)	4.8 (0.5)	63.0 (18.4)	4.7 (0.5)	41.4 (3.8)	4.0 (0.2)	36.2 (3.2)	3.8 (0.2)	50.0 (5.7)	4.2 (0.2)	45.6 (5.2)	4.3 (0.3)
20/7	0.2	71.6 (18.2)	4.9 (0.5)	62.2 (17.3)	4.7 (0.5)	41.6 (3.6)	4.1 (0.2)	36.4 (3.4)	3.8 (0.2)	49.1 (6.3)	4.2 (0.3)	45.4 (5.2)	4.3 (0.2)
20/7	0.6	70.7 (16.9)	4.8 (0.4)	63.3 (16.8)	4.7 (0.5)	41.6 (4.0)	4.1 (0.2)	36.5 (3.6)	3.9 (0.2)	47.6 (6.1)	4.2 (0.3)	45.0 (5.1)	4.3 (0.2)
50/10	0	68.0 (17.7)	4.7 (0.5)	54.6 (16.0)	4.4 (0.5)	39.6 (3.4)	4.0 (0.2)	28.8 (2.7)	3.5 (0.2)	47.4 (6.3)	4.1 (0.3)	40.0 (6.5)	4.0 (0.3)
50/10	0.2	67.1 (18.7)	4.7 (0.5)	55.0 (16.7)	4.4 (0.5)	39.8 (3.6)	4.0 (0.2)	28.9 (3.0)	3.5 (0.2)	45.4 (6.8)	4.1 (0.3)	39.8 (5.7)	4.0 (0.3)
50/10	0.6	67.9 (17.3)	4.8 (0.5)	54.8 (15.2)	4.5 (0.5)	40.2 (3.6)	4.0 (0.2)	29.5 (3.0)	3.6 (0.2)	44.8 (6.6)	4.1 (0.3)	40.3 (5.0)	4.1 (0.3)
50/35	0	58.1 (17.6)	4.5 (0.5)	53.7 (17.7)	4.4 (0.5)	32.0 (4.0)	3.6 (0.2)	28.6 (2.9)	3.5 (0.2)	35.6 (7.1)	3.6 (0.4)	40.0 (5.5)	4.0 (0.3)
50/35	0.2	57.7 (16.8)	4.5 (0.5)	52.8 (17.3)	4.4 (0.5)	32.2 (3.4)	3.7 (0.2)	29.0 (3.2)	3.5 (0.2)	35.0 (8.1)	3.6 (0.4)	39.7 (5.0)	4.0 (0.3)
50/35	0.6	59.1 (16.0)	4.5 (0.5)	54.5 (16.1)	4.4 (0.5)	32.8 (2.9)	3.7 (0.2)	29.2 (3.0)	3.6 (0.2)	36.0 (7.0)	3.7 (0.3)	39.7 (5.2)	4.1 (0.3)
100/20	0	58.6 (17.9)	4.5 (0.6)	37.0 (12.6)	3.8 (0.5)	35.0 (3.3)	3.8 (0.2)	20.5 (2.7)	3.0 (0.2)	40.7 (7.0)	3.8 (0.3)	34.7 (6.8)	3.8 (0.4)
100/20	0.2	58.4 (16.1)	4.5 (0.5)	37.6 (13.6)	3.9 (0.6)	35.6 (3.6)	3.8 (0.2)	20.5 (2.6)	3.1 (0.2)	40.8 (8.2)	3.8 (0.4)	35.1 (6.4)	3.8 (0.4)
100/20	0.6	59.5 (17.3)	4.6 (0.5)	40.2 (14.0)	4.0 (0.5)	36.1 (3.2)	3.9 (0.2)	21.1 (2.6)	3.1 (0.2)	40.0 (7.2)	3.9 (0.3)	36.5 (5.7)	3.9 (0.4)
n = 1000													
20/7	0	86.5 (10.6)	5.1 (0.3)	82.8 (12.3)	5.1 (0.3)	43.3 (3.5)	4.1 (0.2)	40.3 (2.8)	4.0 (0.2)	58.3 (5.2)	4.5 (0.2)	55.6 (5.5)	4.7 (0.2)
20/7	0.2	85.3 (10.2)	5.1 (0.3)	82.1 (12.9)	5.1 (0.3)	43.3 (3.9)	4.1 (0.2)	40.5 (2.8)	4.0 (0.2)	57.3 (4.8)	4.5 (0.2)	55.0 (4.8)	4.7 (0.2)
20/7	0.6	85.4 (10.5)	5.1 (0.3)	80.3 (14.6)	5.1 (0.3)	42.7 (3.9)	4.2 (0.2)	40.4 (2.8)	4.0 (0.2)	54.9 (5.2)	4.4 (0.2)	54.2 (4.8)	4.7 (0.2)
50/10	0	85.4 (11.4)	5.1 (0.3)	80.5 (16.8)	5.0 (0.3)	42.6 (3.2)	4.1 (0.2)	35.2 (2.8)	3.8 (0.2)	56.6 (5.0)	4.4 (0.2)	51.6 (4.8)	4.5 (0.2)
50/10	0.2	85.2 (12.0)	5.1 (0.3)	79.1 (15.6)	5.0 (0.3)	42.5 (3.1)	4.1 (0.2)	35.1 (2.5)	3.8 (0.2)	55.5 (4.9)	4.4 (0.2)	50.9 (4.4)	4.5 (0.2)
50/10	0.6	84.4 (12.1)	5.1 (0.3)	77.2 (15.3)	5.0 (0.4)	41.9 (3.4)	4.1 (0.2)	35.5 (2.8)	3.8 (0.2)	52.2 (4.9)	4.4 (0.2)	49.8 (4.3)	4.6 (0.2)
50/35	0	82.3 (13.6)	5.1 (0.3)	80.6 (16.0)	5.0 (0.3)	37.5 (2.6)	3.9 (0.2)	35.0 (2.7)	3.8 (0.2)	50.3 (5.5)	4.2 (0.2)	51.2 (4.2)	4.5 (0.2)
50/35	0.2	82.3 (13.2)	5.0 (0.3)	79.3 (14.4)	5.0 (0.3)	37.7 (2.6)	3.9 (0.2)	35.1 (2.6)	3.8 (0.2)	49.6 (5.9)	4.2 (0.2)	50.8 (4.4)	4.5 (0.2)
50/35	0.6	81.1 (11.9)	5.0 (0.3)	78.1 (12.3)	5.0 (0.3)	37.8 (2.6)	3.9 (0.2)	35.7 (2.6)	3.8 (0.2)	47.4 (6.1)	4.2 (0.2)	49.9 (4.5)	4.5 (0.2)
100/20	0	82.0 (13.5)	5.0 (0.3)	73.3 (16.3)	4.9 (0.4)	40.2 (2.8)	4.0 (0.2)	29.0 (2.3)	3.5 (0.2)	52.9 (4.9)	4.3 (0.2)	48.4 (4.1)	4.4 (0.2)
100/20	0.2	82.7 (11.6)	5.0 (0.3)	73.9 (16.0)	4.9 (0.4)	40.4 (2.8)	4.0 (0.2)	29.4 (2.5)	3.5 (0.2)	52.5 (5.5)	4.3 (0.2)	47.9 (4.5)	4.4 (0.2)
100/20	0.6	81.6 (12.9)	5.0 (0.3)	71.8 (15.1)	4.9 (0.4)	40.3 (2.7)	4.0 (0.2)	29.6 (3.0)	3.6 (0.2)	50.5 (4.8)	4.3 (0.2)	47.7 (4.0)	4.5 (0.2)
n = 2000													
20/7	0	92.7 (5.2)	5.2 (0.2)	92.0 (6.8)	5.2 (0.2)	43.6 (3.3)	4.2 (0.2)	42.7 (2.6)	4.1 (0.2)	68.5 (4.6)	4.8 (0.2)	67.3 (4.8)	5.0 (0.2)
20/7	0.2	91.8 (5.0)	5.2 (0.2)	91.3 (6.5)	5.2 (0.2)	43.8 (3.2)	4.2 (0.2)	42.7 (2.8)	4.1 (0.2)	66.3 (4.5)	4.8 (0.2)	66.0 (4.5)	5.0 (0.2)
20/7	0.6	91.8 (6.0)	5.2 (0.2)	90.6 (7.5)	5.2 (0.2)	43.2 (3.5)	4.2 (0.2)	42.8 (2.8)	4.1 (0.2)	61.9 (3.8)	4.7 (0.2)	64.4 (4.3)	5.0 (0.2)
50/10	0	91.9 (6.0)	5.2 (0.2)	90.8 (7.3)	5.2 (0.2)	43.6 (3.0)	4.1 (0.2)	39.6 (2.4)	3.9 (0.2)	67.6 (4.8)	4.8 (0.2)	63.7 (4.8)	4.9 (0.2)
50/10	0.2	91.7 (5.6)	5.2 (0.2)	90.6 (6.6)	5.2 (0.2)	43.3 (3.4)	4.1 (0.2)	39.6 (2.7)	4.0 (0.1)	65.6 (4.6)	4.7 (0.2)	62.1 (4.5)	4.9 (0.2)
50/10	0.6	91.0 (6.1)	5.2 (0.2)	89.7 (8.1)	5.2 (0.2)	43.2 (3.3)	4.2 (0.2)	39.7 (2.5)	4.0 (0.2)	60.4 (4.6)	4.6 (0.2)	60.6 (4.1)	4.9 (0.2)
50/35	0	91.9 (5.8)	5.2 (0.2)	91.5 (7.5)	5.2 (0.2)	41.0 (2.5)	4.0 (0.1)	39.5 (2.4)	4.0 (0.2)	64.7 (5.3)	4.7 (0.2)	63.5 (4.9)	4.9 (0.2)
50/35	0.2	91.4 (6.0)	5.2 (0.2)	90.8 (6.7)	5.2 (0.2)	41.2 (2.5)	4.0 (0.2)	39.6 (2.6)	4.0 (0.2)	63.0 (5.4)	4.7 (0.2)	62.2 (4.7)	4.9 (0.2)
50/35	0.6	90.6 (6.4)	5.2 (0.2)	88.9 (7.6)	5.2 (0.2)	41.2 (2.7)	4.1 (0.2)	39.9 (2.4)	4.0 (0.2)	58.3 (4.8)	4.6 (0.2)	60.5 (4.7)	4.9 (0.2)
100/20	0	91.5 (6.2)	5.2 (0.2)	90.2 (6.6)	5.2 (0.2)	42.5 (2.6)	4.1 (0.2)	35.3 (2.4)	3.8 (0.1)	67.1 (5.7)	4.8 (0.2)	61.1 (4.5)	4.8 (0.2)
100/20	0.2	91.0 (6.2)	5.2 (0.2)	89.5 (7.9)	5.2 (0.2)	42.8 (3.0)	4.1 (0.2)	35.5 (2.3)	3.8 (0.2)	65.0 (5.1)	4.7 (0.2)	59.7 (4.7)	4.8 (0.2)
100/20	0.6	91.1 (5.6)	5.2 (0.2)	89.1 (7.7)	5.2 (0.2)	42.5 (2.7)	4.1 (0.2)	35.8 (2.3)	3.8 (0.2)	60.0 (4.5)	4.6 (0.2)	58.0 (4.7)	4.8 (0.2)

Table 5: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal two-stage dynamic treatment regime (DTR) with 3 possible treatments based on an underlying, nontree-type DTR and assuming a large confounding effect of a binary variable. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ρ = the correlation coefficient used to generate covariates in $\mathbf{H}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; Naive T-RL = Naive Tree-based Reinforcement Learning; Q-Linear-R = Restricted Linear Q-Learning; Q-NP-R = Restricted Nonparametric Q-Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment rule estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{g}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	ρ	ReST-L		T-RL		Q-Linear-R		Q-Linear		Q-NP-R		Q-NP	
		%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
n = 500													
20/7	0	66.5 (22.2)	4.8 (0.6)	58.2 (20.0)	4.6 (0.6)	32.4 (3.0)	3.7 (0.2)	25.0 (2.8)	3.5 (0.2)	41.0 (7.8)	3.9 (0.3)	33.1 (6.7)	3.9 (0.3)
20/7	0.2	65.2 (23.2)	4.7 (0.6)	58.0 (21.7)	4.5 (0.6)	32.4 (3.1)	3.7 (0.2)	25.0 (3.0)	3.5 (0.2)	39.8 (7.6)	3.9 (0.3)	32.9 (6.7)	3.9 (0.3)
20/7	0.6	68.4 (20.8)	4.8 (0.6)	58.9 (21.7)	4.6 (0.6)	32.1 (3.4)	3.7 (0.2)	25.0 (3.3)	3.5 (0.2)	38.6 (7.7)	3.9 (0.3)	32.2 (6.0)	3.9 (0.3)
50/10	0	60.7 (22.1)	4.6 (0.6)	49.6 (18.2)	4.3 (0.6)	31.1 (3.2)	3.6 (0.2)	20.8 (2.7)	3.3 (0.2)	36.1 (7.4)	3.8 (0.3)	25.9 (6.0)	3.6 (0.3)
50/10	0.2	61.5 (21.0)	4.6 (0.6)	49.6 (18.3)	4.3 (0.6)	30.9 (3.0)	3.6 (0.2)	20.9 (2.4)	3.3 (0.2)	35.2 (7.3)	3.8 (0.3)	26.3 (5.6)	3.6 (0.3)
50/10	0.6	60.7 (20.0)	4.6 (0.6)	49.2 (15.8)	4.3 (0.5)	30.7 (3.1)	3.7 (0.2)	20.5 (2.5)	3.3 (0.2)	34.2 (7.3)	3.7 (0.3)	26.2 (6.0)	3.6 (0.3)
50/35	0	53.4 (18.3)	4.4 (0.6)	48.8 (17.4)	4.3 (0.6)	25.0 (2.6)	3.3 (0.2)	20.7 (2.8)	3.2 (0.2)	24.8 (5.7)	3.3 (0.3)	25.8 (6.6)	3.6 (0.3)
50/35	0.2	52.9 (20.4)	4.3 (0.6)	49.2 (19.3)	4.3 (0.6)	24.9 (2.8)	3.3 (0.2)	20.9 (2.6)	3.3 (0.2)	25.3 (5.7)	3.3 (0.3)	26.3 (6.0)	3.6 (0.3)
50/35	0.6	53.2 (18.0)	4.4 (0.5)	48.4 (18.8)	4.3 (0.6)	24.4 (3.1)	3.3 (0.2)	20.5 (2.5)	3.3 (0.2)	25.0 (5.7)	3.4 (0.3)	25.7 (6.3)	3.6 (0.3)
100/20	0	53.2 (19.3)	4.3 (0.6)	31.7 (11.7)	3.7 (0.4)	27.5 (3.0)	3.5 (0.2)	16.1 (2.4)	2.9 (0.2)	29.5 (6.8)	3.5 (0.3)	22.0 (6.8)	3.3 (0.3)
100/20	0.2	52.1 (17.0)	4.3 (0.6)	32.4 (13.2)	3.7 (0.5)	27.3 (3.2)	3.5 (0.2)	16.2 (2.3)	2.9 (0.2)	28.4 (6.1)	3.5 (0.3)	22.1 (6.3)	3.4 (0.3)
100/20	0.6	54.1 (19.5)	4.4 (0.6)	33.1 (12.7)	3.7 (0.5)	26.9 (3.3)	3.5 (0.2)	16.2 (2.2)	2.9 (0.2)	28.6 (6.0)	3.5 (0.3)	22.4 (6.1)	3.4 (0.3)
n = 1000													
20/7	0	82.8 (17.2)	5.1 (0.4)	78.7 (20.8)	5.1 (0.5)	34.1 (3.1)	3.8 (0.2)	28.1 (2.6)	3.7 (0.2)	53.2 (7.3)	4.4 (0.3)	48.4 (6.0)	4.5 (0.2)
20/7	0.2	82.3 (15.9)	5.1 (0.4)	78.5 (19.3)	5.1 (0.4)	34.3 (3.4)	3.8 (0.2)	28.0 (2.5)	3.7 (0.2)	51.0 (7.4)	4.3 (0.3)	47.3 (6.4)	4.5 (0.2)
20/7	0.6	81.4 (15.2)	5.1 (0.4)	78.3 (17.5)	5.0 (0.4)	34.1 (3.8)	3.8 (0.2)	27.7 (2.7)	3.7 (0.2)	48.2 (6.4)	4.2 (0.3)	45.8 (6.1)	4.4 (0.2)
50/10	0	81.6 (20.4)	5.1 (0.4)	74.6 (21.5)	5.0 (0.5)	33.8 (2.5)	3.8 (0.2)	24.4 (2.4)	3.5 (0.1)	50.1 (8.3)	4.2 (0.3)	41.0 (6.4)	4.2 (0.3)
50/10	0.2	82.7 (18.4)	5.1 (0.4)	76.0 (20.8)	5.0 (0.5)	33.7 (3.2)	3.8 (0.2)	24.3 (2.4)	3.5 (0.1)	48.5 (7.6)	4.2 (0.3)	40.0 (5.3)	4.2 (0.3)
50/10	0.6	80.0 (21.6)	5.0 (0.5)	72.2 (20.5)	4.9 (0.5)	33.4 (2.8)	3.8 (0.2)	24.2 (2.3)	3.5 (0.2)	45.3 (6.6)	4.1 (0.3)	39.3 (5.2)	4.2 (0.2)
50/35	0	77.6 (20.5)	5.0 (0.5)	74.8 (20.9)	5.0 (0.5)	29.5 (2.6)	3.6 (0.1)	24.4 (2.4)	3.5 (0.1)	38.0 (6.7)	3.8 (0.3)	40.8 (5.2)	4.2 (0.2)
50/35	0.2	79.4 (19.4)	5.0 (0.5)	75.4 (21.4)	4.9 (0.5)	29.2 (2.6)	3.6 (0.1)	24.4 (2.4)	3.5 (0.2)	38.6 (6.4)	3.9 (0.3)	40.6 (5.6)	4.2 (0.3)
50/35	0.6	76.6 (22.6)	5.0 (0.6)	72.2 (21.7)	4.9 (0.6)	29.0 (2.3)	3.6 (0.2)	24.2 (2.5)	3.5 (0.2)	36.9 (6.6)	3.8 (0.3)	38.8 (4.9)	4.2 (0.2)
100/20	0	78.6 (22.7)	5.0 (0.6)	66.3 (22.5)	4.8 (0.7)	31.5 (2.6)	3.7 (0.2)	20.8 (2.2)	3.3 (0.1)	43.5 (7.1)	4.0 (0.3)	36.3 (5.7)	4.0 (0.3)
100/20	0.2	79.0 (22.4)	5.0 (0.5)	68.2 (21.6)	4.8 (0.6)	31.5 (2.8)	3.7 (0.2)	20.8 (2.0)	3.3 (0.2)	42.6 (7.2)	4.0 (0.3)	36.2 (5.3)	4.0 (0.3)
100/20	0.6	76.7 (21.3)	5.0 (0.5)	66.8 (19.9)	4.8 (0.6)	31.2 (2.8)	3.7 (0.2)	20.6 (2.2)	3.3 (0.2)	40.9 (6.1)	4.0 (0.2)	35.1 (5.1)	4.0 (0.2)
n = 2000													
20/7	0	90.5 (4.6)	5.3 (0.2)	90.1 (7.4)	5.3 (0.2)	34.6 (3.7)	3.8 (0.2)	30.2 (2.7)	3.7 (0.2)	64.1 (6.0)	4.7 (0.2)	64.8 (5.3)	5.0 (0.2)
20/7	0.2	89.6 (6.5)	5.3 (0.2)	89.0 (8.7)	5.3 (0.2)	35.0 (3.7)	3.8 (0.2)	30.2 (2.5)	3.8 (0.1)	62.4 (6.1)	4.6 (0.2)	63.3 (5.2)	4.9 (0.2)
20/7	0.6	89.2 (6.7)	5.3 (0.2)	88.0 (10.2)	5.2 (0.2)	34.8 (3.7)	3.9 (0.2)	29.9 (2.6)	3.8 (0.2)	57.1 (5.7)	4.5 (0.2)	59.9 (5.6)	4.8 (0.2)
50/10	0	89.7 (5.6)	5.3 (0.2)	88.8 (7.9)	5.2 (0.2)	34.7 (3.2)	3.8 (0.2)	27.7 (2.3)	3.6 (0.1)	62.9 (6.6)	4.6 (0.2)	58.3 (6.3)	4.8 (0.2)
50/10	0.2	89.9 (6.0)	5.3 (0.2)	88.6 (10.3)	5.2 (0.3)	34.9 (3.2)	3.8 (0.2)	27.6 (2.3)	3.6 (0.1)	60.9 (6.1)	4.6 (0.2)	57.0 (6.7)	4.7 (0.3)
50/10	0.6	89.5 (6.6)	5.2 (0.2)	86.9 (10.4)	5.2 (0.3)	34.7 (3.2)	3.8 (0.2)	27.2 (2.3)	3.6 (0.1)	55.5 (6.2)	4.4 (0.2)	54.5 (6.3)	4.7 (0.2)
50/35	0	89.4 (7.8)	5.2 (0.2)	88.7 (9.2)	5.2 (0.2)	32.9 (2.5)	3.7 (0.1)	27.7 (2.3)	3.6 (0.1)	57.8 (7.5)	4.5 (0.3)	58.5 (6.3)	4.8 (0.2)
50/35	0.2	88.7 (8.1)	5.2 (0.3)	87.4 (10.8)	5.2 (0.3)	32.8 (2.2)	3.7 (0.2)	27.6 (2.1)	3.6 (0.1)	55.6 (7.6)	4.4 (0.3)	57.1 (6.8)	4.7 (0.2)
50/35	0.6	88.6 (9.4)	5.2 (0.3)	86.6 (12.5)	5.2 (0.3)	32.2 (2.2)	3.7 (0.2)	27.4 (2.2)	3.6 (0.2)	50.4 (6.0)	4.3 (0.2)	54.6 (6.9)	4.7 (0.3)
100/20	0	89.9 (8.4)	5.2 (0.2)	87.3 (11.7)	5.2 (0.3)	34.0 (2.6)	3.8 (0.1)	24.5 (2.2)	3.5 (0.2)	60.8 (6.9)	4.6 (0.3)	53.2 (7.0)	4.6 (0.3)
100/20	0.2	89.2 (7.4)	5.2 (0.3)	87.0 (11.7)	5.2 (0.3)	34.0 (2.7)	3.8 (0.2)	24.5 (2.1)	3.5 (0.2)	58.7 (6.8)	4.5 (0.2)	52.4 (6.9)	4.6 (0.3)
100/20	0.6	88.0 (10.2)	5.2 (0.3)	84.8 (17.6)	5.2 (0.4)	33.9 (2.7)	3.8 (0.2)	24.4 (2.2)	3.5 (0.1)	53.1 (6.4)	4.4 (0.3)	50.4 (6.8)	4.5 (0.3)

1.5 Simulation Studies to Evaluate the Performance of ReST-L in a Three-Stage Setting

In this simulation experiment we demostrate that ReST-L can be used also in an experimental setting with more than two stages. Specifically, we simulate a three-stage treatment setting with three possible treatments per stage.

Data generating mechanisms for the first two stages are the same as those described for a two-stage setting in the primary manuscript Section 4.3. In this supplemental simulation, the third stage treatment, A_3, is randomly generated using the multinomial distribution with probabilities π_{30}, π_{31}, π_{32}, where $\pi_{30} = 1 - \pi_{31} - \pi_{32}$, $\pi_{31} = \{0.5X_{C2} - 0.1Y_2\}/[1 + \{0.5X_{C2} - 0.1Y_2\} + \{0.2X_{C3}\}]$, and $\pi_{32} = \{0.2X_{C3}\}/[1 + \{0.5X_{C2} - 0.1Y_2\} + \{0.2X_{C3}\}]$, where C_m refers to the m-th variable in $\mathbf{H}^C_{\text{sub}}$. For an underlying tree-type structure, $g_3^{\text{opt}}(\mathbf{H}_{\text{sub}})$, is assiged as follows: If $X_5 > -1$ & $Y_2 > 1.8$, then $g_3^{\text{opt}} = 2$; if $X_5 > -1$ & $-0.5 < Y_2 < -1.8$, $g_3^{\text{opt}} = 1$; otherwise, $g_3^{\text{opt}} = 0$. When a nontree-type structure is assumed, $g_3^{\text{opt}}(\mathbf{H}_{\text{sub}}) = \mathcal{I}(X_5 + |Y_2| > 1.8) + \mathcal{I}(\sqrt{|10 - Y_2|} < 3.1)$. The intermediate outcome following the third stage is defined as: $Y_3 = \exp\{0.75 + 0.2X_{C3} - (A_3 - g_3^{\text{opt}})^2\} + \epsilon$, where $\epsilon \sim N(0, 1)$. The overall outcome $Y = Y_1 + Y_2 + Y_3$. Due to the increased computational demand of a three-stage setting, we estimate the percentage of subjects correctly-classified to their optimal treatments, as well as the estimated counterfactual outcome under the optimal regime, for $B = 100$ Monte Carlo iterations. Under optimal treatment allocation, $\hat{E}[Y^*\{\mathbf{g}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}] = 10.2$.

As can be seen in Table S-6, for a tree-type DTR with a correctly-specified propensity model, a sample size of $n = 800$, and 100 and 20 variables in $\mathbf{H}$ and $\mathbf{H}_{\text{sub}}$, respectively, ReST-L achieves nearly 90% correctly classification across all three treatment stages, exceeding the estimated 80% correct classification for T-RL. Similar to previous results, as the number of variables in $\mathbf{H}_{\text{sub}}$ increases to $\mathbf{H}$, performance estimates for ReST-L and T-RL are similar. When the propensity model is incorrectly-specified, we see a decrease in performance for both ReST-L and T-RL when all other specifications remain constant. ReST-L, however, maintains the advantage over T-RL in all cases, however. When the underlying DTR is nontree-type, more than twice the sample size is needed to achieve similar performance to the case with an underlying tree-type DTR. With a sample size of $n = 750$, $|\mathbf{H}| = 50$, and assuming a correctly-specified propensity model, for example, ReST-L achieves only about 60% correct classification irrespective of $|\mathbf{H}_{\text{sub}}|$. Although this is an improvement over T-RL, particularly when $|\mathbf{H}_{\text{sub}}| = 10$, which achieves only about 54% correct classification under the same settings, ReST-L with a sample size of $n = 1500$ achieves more than 80% correct classification. Interestingly, although we see some performance improvements for both ReST-L and T-RL with a correctly-specified propensity model, we see the largest improvement in performance when $|\mathbf{H}|$ is larger.

2 Additional Detail for ReST-L Analysis of MIMIC-III data

We conduct this analysis using electronic medical record and administrative data from the Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson et al., 2016; Johnson et al., 2017), a retrospectively-collected and freely-available database accessible through PhysioNet (Goldberger et al., 2000) that contains de-identified and anonymized data for more than 45,000 patients treated in an ICU at Beth Israel Deaconness Medical Center in Boston, Massachusetts. Adult patients ($\geq$ 18 years old) were included in the analysis cohort if they received fluid resuscitation for suspected sepsis (Angus et al., 2001; Horng et al., 2017; Iwashyna et al., 2014) during at least one

Table 6: Performance summary [medians of % *opt* (IQR) and $\hat{E}[Y^*\{\hat{\mathbf{g}}^{\text{opt}}(\mathbf{H}_{\text{sub}})\}]$ (IQR)] for estimation of an optimal three-stage dynamic treatment regime (DTR) with 3 possible treatments per stage and based on an underlying, tree-type DTR (top panel) or nontree-type DTR (bottom panel) with outcomes, treatment assignment, and optimal dynamic treatment regime defined using variables in $\mathbf{H}_{\text{sub}}$. A correlation coefficient $\rho = 0.2$ was used for this simulation. n = sample size of the training dataset; $|\mathbf{H}|$ = number of variables in covariate history $\mathbf{H}$; $|\mathbf{H}_{\text{sub}}|$ = number of variables in subset of covariate history $\mathbf{H}_{\text{sub}}$; ReST-L = Restricted Sub-Tree Learning; T-RL = Tree-based Reinforcement Learning; % opt = percent of test set ($N_{\text{test}} = 1000$) classified to its optimal treatment using a treatment regime estimated using the applicable method; IQR = interquartile range; $\hat{E}(Y^*) = \hat{E}\{Y^*(\hat{\mathbf{g}}^{\text{opt}})\}$ represents the estimated counterfactual mean under the estimated optimal treatment assignment.

n	$\lvert\mathbf{H}\rvert/\lvert\mathbf{H}\text{sub}\rvert$	ReST-L		T-RL	
		%opt	$\hat{E}(Y^*)$	%opt	$\hat{E}(Y^*)$
Tree-type DTR					
Correctly-specified propensity model					
800	100/20	89.4 (8.4)	9.9 (0.3)	79.1 (13.9)	9.5 (0.4)
750	50/35	88.2 (11.4)	9.8 (0.3)	86.7 (14.7)	9.8 (0.4)
750	50/10	91.2 (7.8)	9.9 (0.3)	87.7 (11.3)	9.8 (0.3)
550	20/7	89.2 (8.0)	9.8 (0.3)	85.3 (13.8)	9.7 (0.4)
Incorrectly-specified propensity model					
800	100/20	83.0 (13.0)	9.7 (0.4)	70.8 (15.8)	9.3 (0.6)
750	50/35	85.3 (12.7)	9.7 (0.2)	83.4 (16.9)	9.6 (0.4)
750	50/10	88.9 (12.1)	9.8 (0.3)	84.1 (15.4)	9.7 (0.3)
550	20/7	87.9 (9.1)	9.8 (0.3)	83.8 (14.0)	9.7 (0.4)
Nontree-type DTR					
Correctly-specified propensity model					
1500	50/35	81.8 (29.3)	9.8 (1.2)	80.8 (29.2)	9.8 (1.2)
750	50/35	62.4 (35.6)	9.4 (1.3)	59.2 (34.6)	9.3 (1.3)
1500	50/10	81.8 (32.2)	9.8 (1.2)	79.8 (29.5)	9.8 (1.2)
750	50/10	62.9 (38.5)	9.4 (1.3)	54.5 (30.7)	8.9 (1.1)
Incorrectly-specified propensity model					
1500	50/35	80.8 (30.5)	9.8 (1.2)	80.8 (29.4)	9.7 (1.2)
750	50/35	56.5 (33.7)	9.2 (1.4)	57.2 (30.1)	9.0 (1.3)
1500	50/10	82.0 (29.0)	9.8 (1.2)	80.6 (28.7)	9.8 (1.2)
750	50/10	55.4 (32.4)	8.9 (1.2)	50.1 (27.6)	8.6 (1.2)

ICU admission between 2008 and 2012. For patients with more than one recorded ICU admission for sepsis during the study period, only the first ICU stay was included and analyzed. In order to yield a relatively homogeneous patient population, only those patients surviving less than 48 hours following admission were excluded (Figure S-1).

We impute missing height measurements for 250 individuals using multiple imputation with chained equations (`mice` package in `R`) to create five complete datasets. We then compute each patient's baseline BMI using the standard definition (weight in kilograms/(height in meters)^2). For the BMI values of ten patients that were identified as outliers (i.e., less than 14 or greater than 70) across all five imputed datasets, we set the BMI for these observations to either 14 or 70, respectively. We then conduct ReST-L analyses using each of the five imputed datasets.

The final outcome of interest is the SOFA score evaluated at 24 hours post-admission. Because SOFA scores typically exhibit a right-skewed distribution with values ranging from 0 to 18 and lower scores indicate better prognosis, values are log-transformed and then inverted (as $6 - x$) so that larger values represent better outcomes; the resulting transformed values are approximately symmetric and normally distributed.

All models needed to estimate the AIPW estimator of the counterfactual mean, including the propensity for assignment to a fluid liberal strategy and the conditional mean model for the transformed SOFA score, assume an additive linear relationship with the log odds of assignment or with the outcome, respectively. Selection of variables to include in the stage-specific propensity models and the conditional mean models estimated for each treatment within each stage was performed using stepwise variable selection with the `MASS` package in `R`, and two-way interactions were considered when reasonable. For implementation of tree-based learning, at both stages we specify the need for a 0.5% improvement in the mean counterfactual outcome across treatments in order to perform a covariate split; after accounting for the transformation of the outcome used in the regression models, this represents a SOFA improvement of roughly 0.1. Additionally, we specify a depth of 3 and a minimum of 20 observations per node.

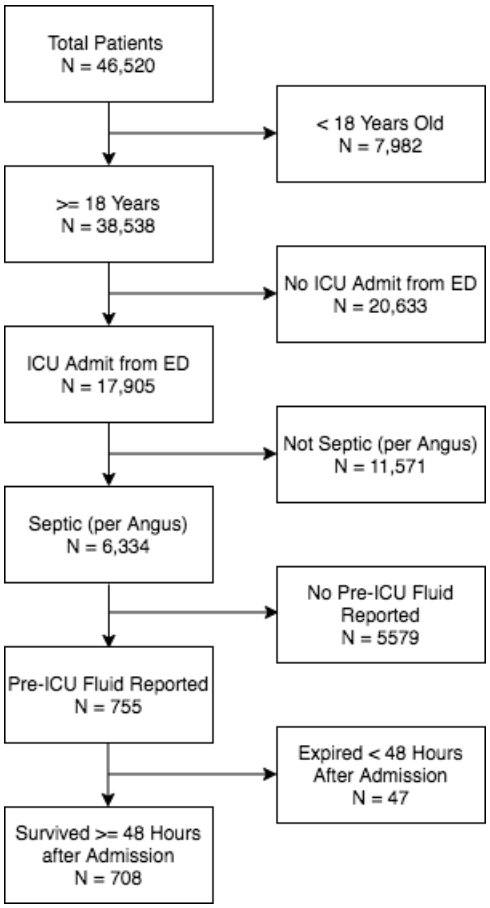

Figure 1: Analysis eligibility criteria and patient population. A total of 708 patients were included in this analysis. Inclusion criteria included being $\geq$ 18 years old at the time of intensive care unit (ICU) admission from the emergency department (ED), having a diagnosis of suspected sepsis (per Angus et al., 2001), receiving documented pre-ICU fluids, and surviving at least 48 hours after ICU admission.

3 References

Angus D.C., Linde-Zwirble W.T., Lidicker J., et al. (2001) Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. Critical Care Medicine, 29(7):1303–1310.

Horng S., Nathanson, L.A., Sontag, D.A., and Shapiro, N.I. (2017) Evaluation of the Angus ICD9-CM Sepsis Abstraction Criteria. bioRXiv. doi:10.1101/124289.

Goldberger A.L., Amaral L.A.N., Glass L., Hausdorff J.M., Ivanov P., Mark R.G., Mietus J.E., Moody G.B., Peng C., and Stanley H.E. (2000) Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. Circulation. 101(23), pe215–e220.

Iwashyna T.J., Odden, A., Rohde, J., Bonham, C., Kuhn, L., Malani, P., Chen, L., and Flanders, S. (2014) Identifying patients with severe sepsis using administrative claims: patient-level validation of the Angus implementation of the International Consensus Conference Definition of Severe Sepsis. Medical Care, 52(6):e39-43. doi: 10.1097/MLR.0b013e318268ac86.

Johnson A.E.W., Pollard T.J., Shen L., Lehman L., Feng M., Ghassemi M., Moody B., Szolovits P., Celi L.A., and Mark R.G. (2016) MIMIC-III, a freely accessible critical care database. Scientific Data. DOI: 10.1038/sdata.2016.35. Available at: "http://www.nature.com/articles/sdata201635"

Johnson, A.E.W., Stone, D.J, Celi L.A., and Pollard, T.J. (2017) The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association.