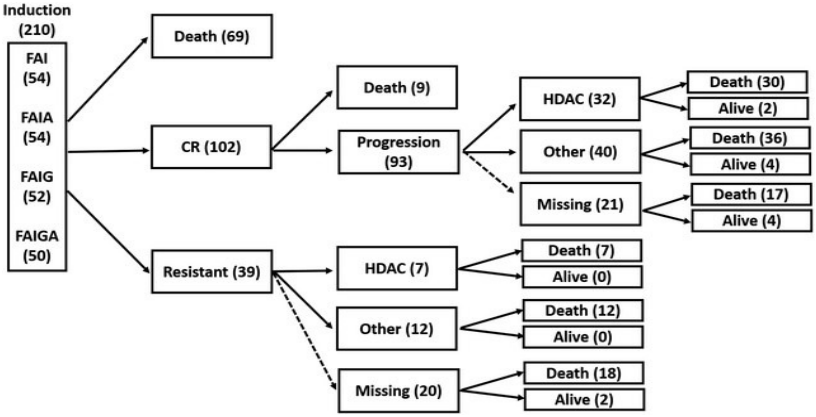
Imputation-based Q-learning for optimizing dynamic treatment regimes with right-censored survival outcome

Lingyun Lyu, Yu Cheng, Abdus S. Wahed

Apr 19, 2024

Presented by Qin Weng

Introduction



Treatment pathways for the AML study

Introduction

Dynamic Treatment regimes (DTR):

Sequentially adaptive medical decision-making algorithms to determine the optimal DTR that yields the best expected outcome

Q-Learning is a conceptually straightforward implementable method for optimizing DTRs

This paper propose an **imputation-based Q-learning (IQ-learning)** to identify the optimal DTR for patients in AML dataset by maximizing their expected overall survival time while accounting for **missing treatment data** and **right-censored data**.

Notation setup

For the j^{th} stage, j = 1,2,…J
A_j: treatment received at Stage *j*
R_j: indicator of treatment been observed
X_j: covariates be observed between Stage *j-1* and *j*
η_j: indicator of entering Stage *j*
T_j: survival time within Stage *j*

T: overall survival time
C: censoring time
U: observed time, U=min(T,C)
δ: event indicator, I(T≤C)

Framework

Overall survival time $T = \sum_{j=1}^{J} \eta_j T_j$

Cumulative information for past and future $\bar{\boldsymbol{A}}_j = (A_1, \ldots, A_j)$ and $\underline{\boldsymbol{A}}_j = (A_j, \ldots, A_J)$

Potential overall survival time $T^*(\bar{\boldsymbol{a}}_J) = \sum_{j=1}^{J} \eta_j T_j^*(\bar{\boldsymbol{a}}_j)$

A DTR is a set of decision rules, to map the historical information space to the treatment space

$$d = \{d_1(\boldsymbol{h}_1), \ldots, d_J(\boldsymbol{h}_J)\} \in \mathscr{D}, \text{ where } d_j(\boldsymbol{h}_j) : \boldsymbol{H}_j \to \boldsymbol{A}_j$$

An optimal DTR is the set of decision rules yielding maximal expected overall survival time

$$d^{opt} = \{d_1^{opt}(\boldsymbol{h}_1), \ldots, d_J^{opt}(\boldsymbol{h}_J)\}$$

Basic Q-learning for DTR Optimization

Cumulative counterfactual survival time at Stage j and onward

$$\tilde{T}_j(\bar{\boldsymbol{a}}_j, \underline{\boldsymbol{a}}_{j+1}^{opt}) = T_j^*(\bar{\boldsymbol{a}}_j) + \sum_{l=j+1}^{J} \eta_l T_l^*(\bar{\boldsymbol{a}}_{l-1}, a_l^{opt})$$

$$\tilde{T}_j = T_j + \sum_{l=j+1}^{J} \eta_l T_l^*(\bar{a}_{l-1}, a_l^{opt})$$

Q-function and estimation by accelerated failure time (AFT) model

$$Q_j(\boldsymbol{H}_j, A_j; \boldsymbol{\beta}_j) = E\left[f(\tilde{T}_j) \mid \boldsymbol{H}_j, A_j, \eta_j = 1\right] = \boldsymbol{\beta}_{j0}^T \boldsymbol{H}_{j0} + \left(\boldsymbol{\beta}_{j1}^T \boldsymbol{H}_{j1}\right) A_j,$$

Identify optimal treatment at Stage j $\hat{a}_j^{opt} = \hat{d}_j(\boldsymbol{h}_j) = \underset{a_j}{\operatorname{argmax}}\, Q_j(\boldsymbol{h}_j, a_j; \hat{\boldsymbol{\beta}}_j) = I\left(\hat{\boldsymbol{\beta}}_{j1}^T \boldsymbol{h}_{j1} > 0\right)$

Inversely calculate time $f^{-1}(Q_j(\boldsymbol{H}_{ji}, \hat{a}_j^{opt}; \hat{\boldsymbol{\beta}}_j))$

Basic Q-learning for DTR Optimization

Pseudo-outcome construction

Assume parametric models for the survival times to direct estimate counterfactual survival under optimal treatments

Limitation

- Q-function could be mis-specified
- Carry backward model misspecification to earlier stages
- May result in implausible values

IQ-Learning

Optimization

- Any class of flexible nonparametric or semiparametric survival models (COX-PH)

Prediction

- Hot-deck multiple imputation (MI)
- Only impute for those who did not receive optimal treatment from those who received optimal treatment

(Address missing treatment)

- Inverse-probability weighting (IPW) and MI

Optimization by Cox-PH model

$\lambda_j(t)$: hazard rate under the observed/counterfactual outcomes

Q-function and estimation by Cox-PH model

$$\lambda_J\left(t \mid \boldsymbol{H}_J, A_J, \eta_J = 1; \boldsymbol{\xi}_J\right) = \lambda_{J0}(t) \exp\left\{\boldsymbol{\beta}_{J0}^T \boldsymbol{H}_{J0} + \boldsymbol{\beta}_{J1}^T \boldsymbol{H}_{J1} A_J\right\}$$

$$Q_J\left(\boldsymbol{h}_J, a_J; \boldsymbol{\xi}_J\right) = \int_0^\infty f(t)\left\{-dS_J\left(t \mid \boldsymbol{h}_J, a_J; \boldsymbol{\xi}_J\right)\right\},$$

Identify optimal treatment at Stage J

$$\hat{a}_J^{opt} = I(\hat{\boldsymbol{\beta}}_{J1}^T \boldsymbol{h}_{J1} < 0)$$

*Account for **non-randomized treatments** by propensity score adjustment*

Prediction

Hot-deck imputation:

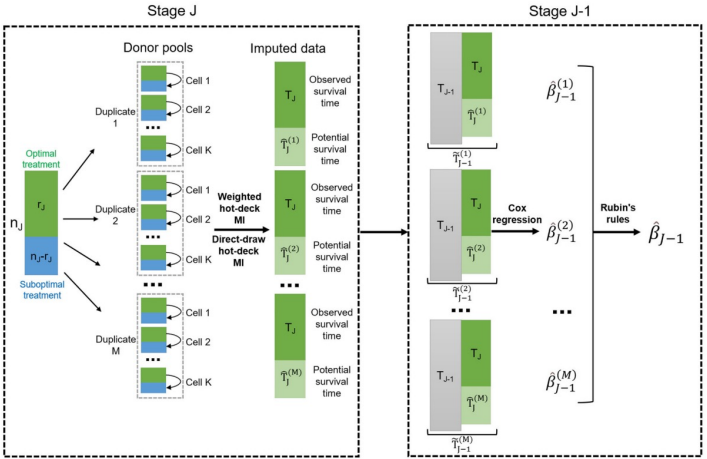
Replace missing values in a nonrespondent (the recipient) with observed values from a "similar" respondent (the donor), which is randomly selected from a group of "similar" units (donor pool)

Hot-deck MI:

- Creating donor pools
- Determining sampling weights
- Making estimation and inference based on the multiple imputed datasets

*Account for **right censoring** by weighted hot-deck (WHD) or direct-draw hot-deck (DHD)*

Optimal DTR estimation for complete data



...

Estimation for survival time at Stage *J-1*

Imputed potential survival time $\hat{T}_J^{(m)}$

- Not entered stage J

  Not censored: T_{J-1},

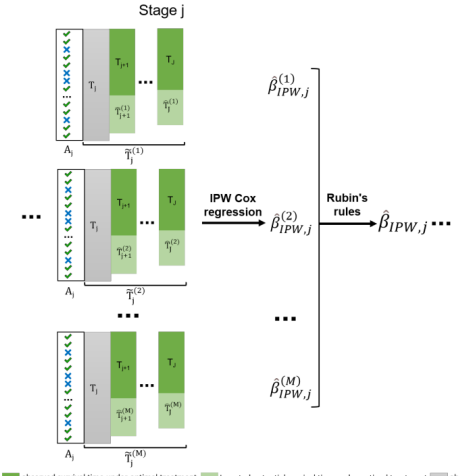
  Censored: $U - \sum_{l=1}^{J-2} T_l$

- Entered stage J

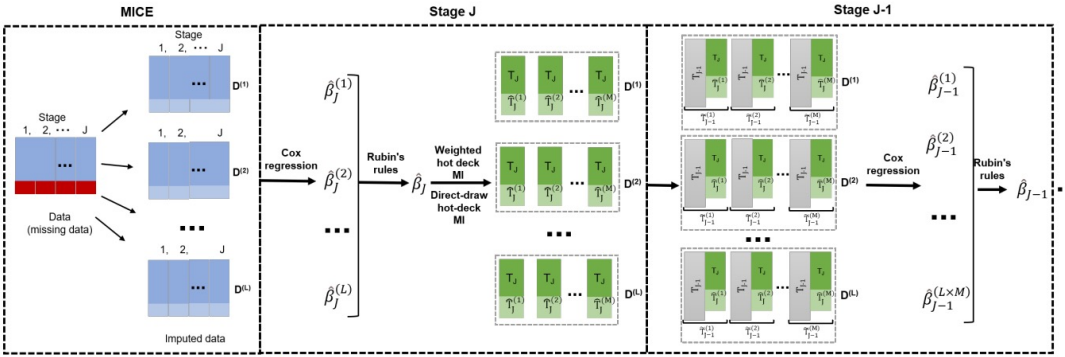
  Received optimal A_J: $T_{J-1} + T_J$

  Not received optimal A_J: $T_{J-1} + \hat{T}_J^{(m)}$

Optimal DTR estimation for incomplete data (IPW method)



Optimal DTR estimation for incomplete data (MICE method)



Simulation

Data generation:

Stage 1: X_{11}, X_{12}, X_{13}, X_{14}, A_1; Stage 2: η_2, X_{21}, X_{22}; Censoring: C

A_2 missing indicator R_2: $\text{logit}(P(R_{i2}=1)) = \gamma_0 + \gamma_1 X_{i12}$

Observed time T_2: $\text{Weibull}(\alpha_2, \exp(\psi_{20} + \psi_{21} X_{i11} + \psi_{22} X_{i22} + \psi_{23} A_{i2} + \psi_{24} A_{i2} X_{i11}))$

True optimal treatment: $A_{i2}^{opt} = I(\psi_{23} + \psi_{24} X_{i11} > 0)$

Counterfactual overall time $\widetilde{T}$i: $\text{Weibull}(\alpha_1, \exp(\psi_{10} + \psi_{11} X_{i11} + \psi_{12} X_{i12} + \psi_{13} A_{i1} + \psi_{14} A_{i1} X_{i12}))$

True optimal treatment: $A_{i1}^{opt} = I(\psi_{13} + \psi_{14} X_{i12} > 0)$

Observed Stage 1 time: $T_{i1} = \widetilde{T}_i - T_{i2}^{opt}$

Observed overall time: $T_i = T_{i1} + T_{i2}$

If not entering Stage 2: $T_i = T_{i1} = \widetilde{T}_i$

Simulation

Model correct specification vs. misspecification

Observed time T_2:

$$\text{log-logistic}\left(\alpha_2, \exp\left(\psi_{20} + \psi_{21} X_{i11} + \psi_{22} X_{i22} + \psi_{23} A_{i2} + \psi_{24} A_{i2} X_{i11}\right)\right)$$

Counterfactual overall time $\tilde{T}i$:

$$\text{log-logistic}\left(\alpha_1, \exp\left(\psi_{10} + \psi_{11} X_{i11} + \psi_{12} X_{i12} + \psi_{13} A_{i1} + \psi_{14} A_{i1} X_{i12}\right)\right)$$

Simulation

Scenarios:

Scenario 1 (original parameters)

- varying missingness (10%, 20%, 30%)
- censoring (0%, 30%, 50%)
- sample size (500, 1000)
- misspecification (Weibull, log-logistic)

Simulation

Estimation for Stage 2

- IPW outperformed MICE

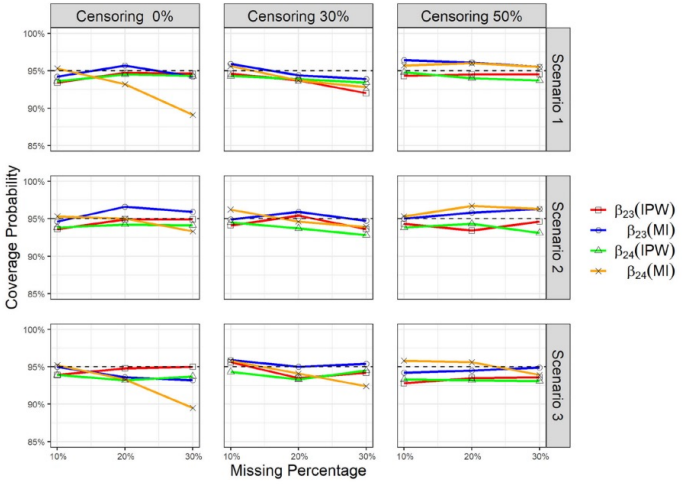


FIGURE 3 Coverage probabilities for 95% confidence intervals for the selected second-stage parameters for $n = 500$. Simulation scenarios are described in Web Table 1. IPW: Inverse-probability-weighting; MI: multiple imputation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Simulation

Estimation for Stage 1

- WHD and DHD are comparable

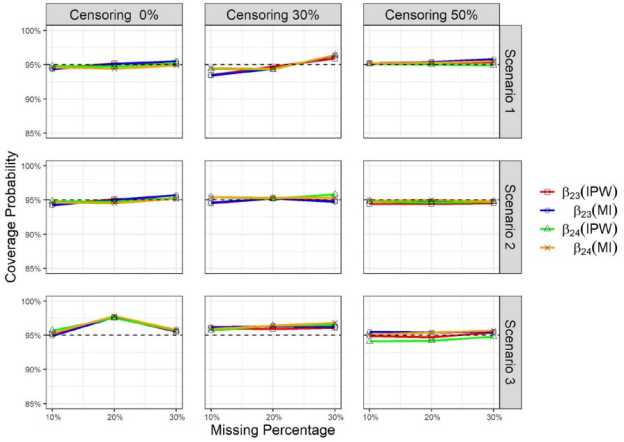


FIGURE 4 Coverage probabilities for 95% confidence intervals for the selected first-stage parameters for $n = 500$. Simulation scenarios are described in Web Table 1. IPW: Inverse-probability-weighting; MI: multiple imputation. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

Simulation

Misspecification

- IQ-learning is more robust

TABLE 1 Overall correct optimal DTR identification percentages, defined as the percentage (averaged over Monte Carlo samples) of participants for whom optimal treatment in both stages were correctly identified, with corresponding 95% confidence intervals.

		Correct model specification		Model misspecification[c]	
$C\%$[a]	$M\%$[b]	IQ-learning	HNW	IQ-learning	HNW
$N = 500$					
0	10	93.0 (81.1, 100.0)	92.7 (79.9, 100.0)	83.4 (53.2, 100.0)	72.6 (31.5, 100.0)
	20	93.2 (81.0, 100.0)	92.3 (77.8, 100.0)	83.5 (51.5, 100.0)	74.2 (31.4, 100.0)
	30	93.0 (80.7, 100.0)	91.4 (76.5, 100.0)	82.2 (48.9, 100.0)	72.5 (27.0, 100.0)
30	10	91.1 (75.1, 100.0)	90.6 (71.7, 100.0)	79.7 (42.0, 100.0)	72.0 (26.6, 100.0)
	20	90.8 (73.8, 100.0)	89.7 (71.1, 100.0)	77.5 (36.9, 100.0)	69.3 (21.1, 100.0)
	30	91.3 (74.0, 100.0)	89.3 (68.3, 100.0)	76.6 (33.9, 100.0)	67.7 (17.4, 100.0)
50	10	89.1 (67.4, 100.0)	83.2 (45.9, 100.0)	75.7 (32.5, 100.0)	34.0 (23.0, 45.1)
	20	88.9 (65.8, 100.0)	80.2 (37.6, 100.0)	74.8 (30.6, 100.0)	34.1 (23.6, 44.7)
	30	88.1 (59.8, 100.0)	76.4 (30.3, 100.0)	73.5 (27.9, 100.0)	33.8 (23.0, 44.6)
$N = 1000$					
0	10	95.3 (87.7, 100.0)	95.1 (87.1, 100.0)	89.1 (68.0, 100.0)	80.9 (47.4, 100.0)
	20	95.4 (88.2, 100.0)	94.9 (86.8, 100.0)	89.4 (68.4, 100.0)	82.4 (51.6, 100.0)
	30	95.3 (87.5, 100.0)	94.5 (84.8, 100.0)	89.3 (69.8, 100.0)	80.2 (45.7, 100.0)
30	10	94.2 (84.5, 100.0)	93.9 (82.8, 100.0)	86.7 (61.7, 100.0)	81.2 (47.6, 100.0)
	20	94.6 (85.6, 100.0)	93.8 (82.9, 100.0)	87.0 (63.2, 100.0)	80.7 (45.8, 100.0)
	30	94.3 (84.7, 100.0)	93.1 (79.8, 100.0)	87.0 (61.2, 100.0)	80.4 (42.9, 100.0)
50	10	93.2 (81.8, 100.0)	92.3 (76.4, 100.0)	84.9 (56.2, 100.0)	35.3 (28.0, 42.6)
	20	93.1 (81.7, 100.0)	90.9 (69.5, 100.0)	84.1 (52.9, 100.0)	35.2 (27.5, 42.9)
	30	93.0 (78.9, 100.0)	87.9 (57.6, 100.0)	83.2 (50.2, 100.0)	34.9 (26.3, 43.5)

[a] Censoring percentage.
[b] Missing precentage.
[c] True survival times were generated from log-logistic distribution while the IQ-learning method fitted the Cox model and the HNW method fitted the Weibull model.

Simulation

Additional Scenarios:

- Scenario 2: reduced effect size of Stage 2
- Scenario 3: Decrease difference between overall time & Stage 2 time
- Scenario 4: Multilevel treatments (no covariate missing)
- Scenario 5: Time varying covariate dependent censoring (no covariate missing)
- Scenario 6: Mimicking the AML study data

Results:

- Scenario 2 < Scenario 1, only for small sample size
- Scenario 3 < Scenario 1
- Good correct identification rate for Scenario 4, 5, 6

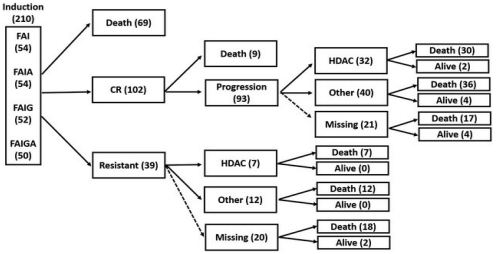
Conclusion

Salvage treatment (Stage 2)

- HDAC for patient with adverse cyto status
- Other therapies for patient with intermediate cyto status

Initial treatment (Stage 1)

- No optimal treatment



Treatment pathways for the AML study

Conclusion

Proposing an imputation-based Q-learning method for the DTR optimization with survival outcomes, and used it to estimate optimal DTR for AML study.

Significance:

- Less sensitive to model misspecification
- Imputed times are always plausible

Limitation:

- Imputation must base on categorical variables
- Number of matching donors cannot be too small
- Requires correct specification for weighting mechanism

Recommendation

Is it worth reading? Maybe

- Solid paper covers survival analysis, censoring, imputation, optimal DTR, Q-learning
- Intuitive methods without heavy theoretical reasoning

- Did not seem to emphasis on right-censoring
- Complicated problem framework and extensive definitions and notations

Thanks!