Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Controlling Cumulative Adverse Risk in Learning Optimal Dynamic Treatment Regimens

Mochuan Liu, Yuanjia Wang, Haoda Fu & Donglin Zeng

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2023.2270637

Taylor & Francis
Taylor & Francis Group

Controlling Cumulative Adverse Risk in Learning Optimal Dynamic Treatment Regimens

Mochuan Liu[a], Yuanjia Wang[b], Haoda Fu[c], and Donglin Zeng[d]

[a]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC; [b]Department of Biostatistics, Columbia University, New York, NY; [c]Eli Lilly and Company, Indianapolis, IN; [d]dDepartment of Biostatistics, University of Michigan, Ann Arbor, MI

ABSTRACT

Dynamic treatment regimen (DTR) is one of the most important tools to tailor treatment in personalized medicine. For many diseases such as cancer and type 2 diabetes mellitus (T2D), more aggressive treatments can lead to a higher efficacy but may also increase risk. However, few methods for estimating DTRs can take into account both cumulative benefit and risk. In this work, we propose a general statistical learning framework to learn optimal DTRs that maximize the reward outcome while controlling the cumulative adverse risk to be below a pre-specified threshold. We convert this constrained optimization problem into an unconstrained optimization using a Lagrange function. We then solve the latter using either backward learning algorithms or simultaneously over all stages based on constructing a novel multistage ramp loss. Theoretically, we establish Fisher consistency of the proposed method and further obtain non-asymptotic convergence rates for both reward and risk outcomes under the estimated DTRs. The finite sample performance of the proposed method is demonstrated via simulation studies and through an application to a two-stage clinical trial for T2D patients. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2023
Accepted September 2023

KEYWORDS
Benefit-risk tradeoff;
Lagrange function;
Multistage decision-making;
Outcome weighted learning;
Q-learning

1. Introduction

Dynamic treatment regimen (DTR) is one of the most important tools to tailor treatment to a patient's evolving health status over multiple stages (Chakraborty and Moodie 2013). During the past decade, many methods have been developed to estimate the optimal DTRs, including regression methods such as A-learning (Murphy 2003), Q-learning (Qian and Murphy 2011), doubly robust regression (e.g., Zhang et al. 2012; Barrett, Henderson, and Rosthøj 2014), and machine-learning methods such as outcome weighted learning (O-learning) (Zhao et al. 2015; Liu et al. 2018) and tree-based methods (e.g., Laber and Zhao 2015; Qiu and Wang 2019).

For many diseases, treatments are usually multifaceted: more aggressive treatments may lead to a higher efficacy but are also more likely to induce elevated risk in the long term. For example, treatments of non-small cell lung cancer (NSCLC) usually consist of three lines of therapies (Socinski and Stinchcombe 2007); several studies suggest that adopting second- and third-line treatments may potentially elevate risks due to the toxicity of the treatments (Kumar and Wakelee 2006). Another example is the treatment of T2D, where the American Diabetes Association (ADA) recommends intensified insulin therapy when patients fail to reach a safe hemoglobin A1c level after receiving first- and second-line medications (American Diabetes Association 2022). However, several studies have shown that insulin therapies are commonly associated with long-term weight gain, which can potentially increase the risk of cardiovascular diseases, and a weight gain controlled under 5% is recommended for T2D patients (Park et al. 2022).

Most methods in personalized medicine literature that consider benefit-risk tradeoffs are restricted to a single-stage decision problem. One class of methods (Lee et al. 2015; Butler et al. 2018) prespecify a utility function to combine benefit and risk outcomes into a single composite outcome, and the optimal decision is obtained by maximizing the utility function. A major limitation of these methods is that reaching a consensus on prespecifying the composite outcome is often difficult, especially when the benefit and risk outcomes are measured on very different scales. Recent work in reinforcement learning (e.g., Mahdavi, Jin, and Yang 2012; Badanidiyuru, Kleinberg, and Slivkins 2018; Cayci, Eryilmaz, and Srikant 2020; Ding et al. 2021) have considered learning optimal policy under safety/budget constraints. However, these methods rely on the Markovian decision process assumption (MDP) and require parametric models for the unknown policy, which do not hold for general DTR problems.

When the cumulative risk needs to be considered in a DTR problem, the most important challenge is that due to delayed effects, treatments at one stage may affect both the benefit and risk outcomes in any of the future stages. Therefore, estimating the optimal treatment rule at any stage must consider its cumulative impact on future stages. However, commonly used backward algorithms such as Q-learning or O-learning require future stage rules to be estimated optimally. These methods are no longer applicable because the cumulative risk control

CONTACT Donglin Zeng ✉ dzeng@umich.edu Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

depends on the future stage rules and the treatment decision, which is yet to be estimated at the current stage.

In this work, we propose a new statistical learning framework, namely, a multistage cumulative benefit-risk (CBR) framework, to estimate the optimal DTRs that maximize the expected benefit (or reward) outcome but, at the same time, control the expected cumulative risk below a pre-specified threshold. We propose two methods to solve CBR. First, we introduce a Lagrange function and obtain its solution via solving an unconstrained DTR problem using a backward algorithm based on Q-learning or O-learning. Second and more interestingly, we propose a new procedure under multistage ramp loss (MRL) to estimate the DTRs simultaneously across all stages. The MRL can be viewed as an extension of the univariate ramp loss to a multivariate setting.

Our work contains several novel contributions. First, converting the constrained estimation for DTRs to the unconstrained problem enables us to adopt the backward algorithm from the existing methods to estimate the optimal DTRs, and we prove that the latter leads to the optimal DTRs that satisfy the cumulative risk control. Second, in addition to the backward induction algorithm, we also propose a simultaneous learning method based on MRL, for which the estimation of one decision function is contingent on other decisions at later stages so that we can estimate the treatment rules using all data simultaneously. Third, we show that the non-asymptotic convergence rates of the expected reward and risk under the estimated rules can be derived from the unconstrained DTRs associated with the Lagrange function, which provides the finite sample performance guarantee. We also show that using MRL is guaranteed to yield Fisher consistent rules for any unconstrained DTRs problem, and consequently, using the multistage ramp loss along with the proposed estimation procedure will yield the true optimal DTRs.

The remaining article is organized as follows. In Section 2, we formally introduce the CBR problem along with assumptions. We then describe a general framework to solve CBR by converting the problem to an unconstrained one. In the same section, we present a backward algorithm based on Q-learning and O-learning and the new MRL approach to obtain the solutions using empirical data. In Section 3, we obtain the non-asymptotic convergence rates for both the expected reward and risk under the estimated rules. In Section 4, we present simulation study results to examine the proposed approaches' performance. In Section 5, we apply the proposed methods to estimate the optimal DTRs using a two-stage trial for treating T2D patients, and future extensions are discussed in Section 6.

2. Method

2.1. Problem Setup and Assumptions

Consider a T-stage decision problem, where T is finite and often small in clinical settings. We use Y to denote the total reward and R to denote the cumulative risk by the end of stage T, both assumed to be bounded by a constant M. We consider a sequence of dichotomous treatments over T stages and let $A_t \in \{-1, +1\}$ denote the observed treatment at stage t. Additionally, we let H_t denote all observed feature variables prior to stage t, including the treatments or any immediate outcomes in previous stages. Thus, $H_1 \subset H_2 \subset \cdots \subset H_T$. We assume that data are from a sequential multiple assignment randomized trial (SMART) (Murphy 2005), so the observed data for n independent subjects consist of $H_{i1}, A_{i1}, \ldots, H_{iT}, A_{iT}, Y_i$ and R_i for $i = 1, \ldots, n$. A DTR is defined as a function from the space:

$$\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_T \to \{-1, +1\}^T, \text{ where } \mathcal{D}_t : \mathcal{H}_t \to \{-1, +1\}.$$

To control the cumulative risk, we formulate the CBR problem as seeking the optimal rule $\mathcal{D}^* = (\mathcal{D}_1^*, \ldots, \mathcal{D}_T^*)$ that solves the optimization problem

$$\max_{\mathcal{D}} E^{\mathcal{D}}[Y], \quad \text{subject to } E^{\mathcal{D}}[R] \leq \tau$$

for a prespecified risk constraint τ. Here, $E^{\mathcal{D}}[\cdot]$ denotes the expectation when A_t are forced to be $\mathcal{D}_t(H_t)$ for all t. In other words, the optimal treatment rule yields the maximal reward at stage T among all feasible rules whose cumulative risk is no greater than threshold τ.

To ensure that $E^{\mathcal{D}}[\cdot]$ is estimable given the observed data, we require several assumptions. Let $\bar{a}_t = (a_1, \ldots, a_t) \in \{-1, +1\}^t$ denote a determined treatment sequence and $\bar{A}_t = (A_1, \ldots, A_t)$ denote the observed treatments prior to stage t. For a random variable X, we use $X(\bar{a}_t)$ to denote the potential outcome when $\bar{A}_t = \bar{a}_t$.

Assumption 1. Stable Unit Treatment Value Assumption (SUTVA): A subject's cumulative potential outcome is not influenced by other subjects' treatment allocation, that is, $(Y, R) = (Y(\bar{a}_T), R(\bar{a}_T))$ if $\bar{A}_T = \bar{a}_T$.

Assumption 2. No Unmeasured Confounders (NUC): For any $t \in \{1, \ldots, T\}$ and $\bar{a}_T \in \{-1, +1\}^T$, A_t is independent of $(H_{t+1}(\bar{a}_t), \ldots, H_T(\bar{a}_{T-1}), Y(\bar{a}_T), R(\bar{a}_T))$ conditioning on H_t.

Assumption 3. Positivity: Let $p(A_t = a|H_t)$ denote the treatment assignment probability of assigning treatment $A_t = a$ given H_t at stage t. For any $t \in \{1, \ldots, T\}$, there exist universal constants $0 < c_1 \leq c_2 < 1$ such that the treatment assignment probability at stage t satisfies $c_1 \leq p(A_t = 1|H_t) \leq c_2$ almost surely.

Assumptions 1 to 3 are standard assumptions for causal inference (Chakraborty and Moodie 2013). Since we consider data from SMART, Assumptions 2 and 3 hold automatically when any treatment assignment probability, known in a SMART, is bounded away from 0. Under Assumptions 1–3, Qian and Murphy (2011) showed that the original problem can be reformulated as

$$\begin{aligned} &\max_{\mathcal{D}} E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right], \\ &\text{subject to } E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^T p(A_t|H_t)}\right] \leq \tau. \end{aligned} \tag{1}$$

Finally, assuming that the decision rules are determined as the signs of some decision functions $(f_1, \ldots, f_T)$, that is, $\mathcal{D}_t(H_t) = \text{sign}(f_t(H_t))$, then (1) becomes

$$\begin{aligned} &\max_{(f_1, \ldots, f_T) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_T} E\left[Y \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right], \\ &\text{subject to } E\left[R \frac{\prod_{t=1}^T \mathbb{I}(A_t f_t(H_t) > 0)}{\prod_{t=1}^T p(A_t|H_t)}\right] \leq \tau, \end{aligned} \tag{2}$$

where $\mathcal{F}_t$ denotes the set of all measurable functions from $\mathcal{H}_t$ to $\mathbb{R}$.

2.2. A General Procedure for Solving CBR Problem

To solve CBR problems, we consider the Lagrange function of (1), or equivalently, (2). For any $\kappa \in [0, \infty]$, the Lagrange function of (1) with multiplier κ is given by

$$E\left[\{Y - \kappa(R - \tau)\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right].$$

Letting $\gamma = \kappa/(1+\kappa) \in [0, 1]$, we aim to solve the following problem for each γ:

$$\begin{aligned}\mathcal{D}^*_\gamma &= (\mathcal{D}^*_{1,\gamma}, \ldots, \mathcal{D}^*_{T,\gamma}) \\ &= \arg\max_{\mathcal{D}} E\left[\{(1-\gamma)Y - \gamma R\} \frac{\prod_{t=1}^{T} \mathbb{I}(A_t = \mathcal{D}_t(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right]\end{aligned} \quad (3)$$

where we omit the constant τ which will not affect the solution when γ is fixed.

Let $\mathfrak{Y}(\gamma)$ and $\mathfrak{R}(\gamma)$ denote the expected reward and risk associated with the optimal decision rules of (3), that is,

$$\mathfrak{Y}(\gamma) = E\left[Y \frac{\prod_{t=1}^{T} \mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma}(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right],$$

$$\mathfrak{R}(\gamma) = E\left[R \frac{\prod_{t=1}^{T} \mathbb{I}(A_t = \mathcal{D}^*_{t,\gamma}(H_t))}{\prod_{t=1}^{T} p(A_t|H_t)}\right].$$

To ensure that there exists a nontrivial solution to the above problem, we also require the following regular assumption:

Assumption 4. $\mathfrak{R}(\gamma)$ is a continuous function for $\gamma \in [0, 1]$ and $\mathfrak{R}(1) < \tau \leq \mathfrak{R}(0)$.

As a note, the restriction $\mathfrak{R}(1) < \tau$ in Assumption 4 ensures that there exists at least one feasible DTR that satisfies the risk constraint, and $\tau \leq \mathfrak{R}(0)$ is to exclude the trivial case when the cumulative risk for the optimal DTR without the constraint is not larger than τ. The continuity assumption in Assumption 4 implies that there exists some γ^*, which may not be unique, satisfying $\mathfrak{R}(\gamma^*) = \tau$. For any such γ^*, our following lemma shows that $\mathcal{D}_{\gamma^*}$ is the optimal DTR.

Lemma 1. Under Assumptions 1–4, both $\mathfrak{Y}(\gamma)$ and $\mathfrak{R}(\gamma)$ are nonincreasing function of γ. Furthermore, $E^{\mathcal{D}_{\gamma^*}}[Y] \geq E^{\mathcal{D}}[Y]$ for any DTRs, $\mathcal{D}$, satisfying $E^{\mathcal{D}}[R] \leq \tau$.

Lemma 1 indicates that to solve the CBR problem, we only need to solve the unconstrained problem (3) for any γ and then find γ^* such that the risk associated with γ^* satisfies the constraint. The proof of Lemma 1 is given in the supplemental material. In addition, the continuity of $\mathfrak{R}(\gamma)$ implies that searching for γ^* can be carried out using the bisection procedure starting from $\gamma_{\min} = 0$ and $\gamma_{\max} = 1$ until reaching the termination condition $|\gamma_{\min} - \gamma_{\max}| \leq \epsilon$ for some convergence threshold ϵ. The complete numerical algorithm based on bisection search is provided in the supplemental material.

As an important remark, although the lemma implies that the optimal DTRs are associated with a linear combination of Y and R, it should be noted that the coefficient in this linear combination, that is, γ^*, is data-driven and depends on the DTRs. Therefore, this problem fundamentally differs from learning the optimal DTRs based on a utility function where the linear combination needs to be pre-specified.

2.3. Backward Algorithm

Since (3) is an unconstrained problem for estimating DTRs for fixed γ, many existing methods such as Q-learning and O-learning can be used to learn the optimal DTRs using a backward procedure, after treating $(1-\gamma)Y - \gamma R$ as the reward outcome. Specifically, we define Q-function in turn for $t = T, T-1, \ldots, 1$ as

$$\begin{aligned}Q_{t,\gamma}(h_t, a_t) = E[&\arg\max_{a_{t+1}\in\{-1,+1\}} Q_{t+1,\gamma}(H_{t+1}, a_{t+1})| \\ &H_t = h_t, A_t = a_t]\end{aligned}$$

with $Q_{T+1,\gamma} = (1-\gamma)Y - \gamma R$. Then the optimal solution for $\mathcal{D}^*_\gamma$ is

$$\mathcal{D}^*_{t,\gamma}(h_t) = \text{sign}(Q_{t,\gamma}(h_t, 1) - Q_{t,\gamma}(h_t, -1)), \quad t = 1, \ldots, T.$$

A backward Q-learning estimates the conditional expectation in the definitions of $Q_{t,\gamma}$ using regression models, in turn from $t = T$ to $t = 1$, then the estimated DTRs are obtained by plugging the estimated Q-functions into the above expression (Qian and Murphy 2011).

A more robust procedure without fitting regression models, namely backward OWL, uses weighted support vector machines to directly optimize the objective function at each stage (Zhao et al. 2015). Specifically, let $(g^*_{1,\gamma}, \ldots, g^*_{T,\gamma})$ denotes the optimal decision functions corresponding to the outcome $O_\gamma = (1-\gamma)Y - \gamma R$, then Zhao et al. (2015) indicates that $\{g^*_{t,\gamma}\}_{t=1}^T$ can be sequentially estimated via

$$g^*_{t,\gamma} = \arg\max_{f\in\mathcal{F}_t} E\left[O_\gamma \frac{\mathbb{I}(A_t f_t(H_t) > 0) \prod_{s=t+1}^{T} \mathbb{I}(A_s g^*_{s,\gamma}(H_s) > 0)}{\prod_{s=t}^{T} p(A_s|H_s)}\right] \quad (4)$$

for $t = T, \ldots, 1$ in a backward order. In other words, the optimal decision function at stage t can be obtained by maximizing the expected cumulative reward up to stage t among patients whose future observed treatments follow the optimal treatments. Theoretically, Zhao et al. (2015) proves that the OWL produces Fisher consistent rules of the original problem. Consequently, Lemma 1 indicates that the proposed procedure and OWL yield a consistent estimation of the original CBR problem. Using empirical data, an estimator of $g^*_{t,\gamma}$ can be obtained by solving the empirical version of (4) in a backward order and by replacing the 0-1 function, $\mathbb{I}(A_t f_t(H_t) > 0)$, with some surrogate function. In particular, Zhao et al. (2015) adopted the hinge loss defined as $\phi(x) = (1-x)^+$ and sequentially solved the following problem

$$\begin{aligned}\widehat{f}_{t,\gamma} = \arg\min_{f_t\in\mathcal{G}_t} \frac{1}{n}\sum_{i=1}^{n}\{(1-\gamma)Y_i - \gamma R_i\} \\ \frac{\prod_{s=t+1}^{T} \mathbb{I}(A_{is}\widehat{f}_{s,\gamma}(H_{is}) > 0)}{\prod_{s=t}^{T} p(A_{is}|H_{is})} \phi(A_{it} f_t(H_{it})) + \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t},\end{aligned} \quad (5)$$

where $\mathcal{G}_t$ is a subspace of $\mathcal{F}_t$. The last term, $\lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t}$, is a regularization term to mitigate overfitting. When $\{\mathcal{G}_t\}_{t=1}^T$ are reproducing kernel Hilbert space (RKHS), the optimization problem (5) can be reformulated as a weighted support vector machine problem (SVM) (Cortes and Vapnik 1995), which can be efficiently solved using standard optimization software. Typical choices of RKHS include the space generated by a linear kernel or a Gaussian kernel with inner product $\langle H_{it}, H_{jt}\rangle = e^{-\sigma^2\|H_{it}-H_{jt}\|_2^2}$ for bandwidth σ^{-1}. Given observed data, the tuning parameter $\{\lambda_{n,t}\}_{t=1}^T$ and $\{\sigma_{n,t}\}_{t=1}^T$ can be selected via cross-validation.

To further improve the performance of OWL, Liu et al. (2018) proposed the augmented OWL (AOWL) by predicting the Q-function of subjects whose observed treatments do not follow the optimal estimated rules and incorporating such predictions to calculate pseudo-outcomes through a doubly robust construction. The detailed implementation of AOWL is presented in the supplemental material. In our subsequent numerical studies, we use both OWL and AOWL for this backward algorithm to solve the Lagrange function in (3) and use O-learning to refer to either OWL or AOWL when the context is clear.

2.4. Simultaneous Algorithm

One disadvantage of O-learning is that the estimation of the early stage can only utilize the information from patients whose observed treatment assignments follow the optimal rules as shown in (4). Moreover, for backward induction methods such as Q-learning and O-learning, the estimation error from later stages due to either model misspecification or overfitting will be accumulated and always present in early-stage estimation. To overcome these disadvantages, we propose a simultaneous algorithm based on multi-stage ramp loss (MRL) described in this section to overcome these disadvantages.

Our key idea is to replace the multivariate 0-1 indicator function in (3) with a continuous surrogate function to be directly optimized without any backward algorithm. Specifically, define $\psi(\cdot)$ as a piecewise linear function given by $\psi(x) = \max(\min(x,1),0)$ then we consider solving the following surrogate problem to substitute (3):

$$\max_{(f_1,\ldots,f_T)\in\mathcal{F}_1\times\cdots\times\mathcal{F}_T} E\left[O_\gamma^+ \frac{\min(\psi(A_1f_1(H_1)),\ldots,\psi(A_Tf_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right] + E\left[\sum_{a_t\in\{-1,+1\},\, a_t\neq A_t} O_\gamma^- \frac{\min(\psi(a_1f_1(H_1)),\ldots,\psi(a_Tf_T(H_T)))}{\prod_{t=1}^T p(A_t|H_t)}\right], \tag{6}$$

Here, O_γ^+ and O_γ^- denote the positive and negative part of O_γ, respectively, that is, $O_\gamma^+ = \max(O_\gamma, 0)$ and $O_\gamma^- = \max(-O_\gamma, 0)$. When O_γ is nonnegative, the optimization problem (6) can be viewed as a minimization problem with loss function $L(f) = -E[\min(\psi(A_1f_1(H_1)),\ldots,\psi(A_Tf_T(H_T)))] + 1$. Figure 1 presents a three-dimensional visualization of this loss function (i.e., $T = 2$). In other words, L can be considered a multivariate extension of the univariate ramp loss function proposed by Huang, Shi, and Suykens (2014). Numerically, MRL can be more robust against extreme errors in f_t's than O-learning because MRL is bounded between 0 and 1 and is closer to the 0-1 loss than the hinge loss used in O-learning. Note that the expression (6) does not require the decision function f_{t_1} to be estimated before or after another decision function f_{t_2}. This implies that MRL solves the optimal decision rules simultaneously so that all patients' information will be used during the estimation, and updating the decision functions in early stages will also update the decision functions of later stages. The second augmentation term of (6) changes the negative response variable to a positive value by reverting the observed treatments to any other treatment sequences. This expression ensures that the weights in each term are always nonnegative even if O_γ is negative. The following lemma ensures that MRL is a valid surrogate problem for (3).

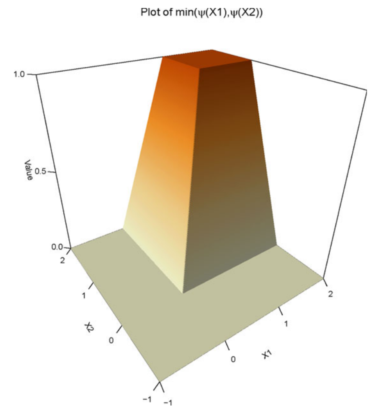

Figure 1. 3D plot of multivariate ramp loss for $T = 2$, $\min(\psi(x_1), \psi(x_2))$.

Lemma 2. If $(f_1^*,\ldots,f_T^*)$ is a solution to (6), then $(\mathrm{sign}(f_1^*),\ldots,\mathrm{sign}(f_T^*))$ maximizes (3).

As remarked after Lemma 1, Lemma 2 plus Lemma 1 indicates that the proposed procedure along with MRL is also guaranteed to yield consistent estimation of the original CBR problem once γ is chosen to satisfy the risk constraint. The proof of Lemma 2 is provided in the supplemental material. Using the empirical data, we propose to solve

$$\max_{(f_1,\ldots,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \frac{1}{n}\sum_{i=1}^n O_{i,\gamma}^+ \frac{\min(\psi(A_{i1}f_1(H_{i1})),\ldots,\psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^T p(A_{it}|H_{it})} + \frac{1}{n}\sum_{i=1}^n \sum_{a_t\in\{-1,1\},a_t\neq A_{it}} O_{i,\gamma}^- \frac{\min(\psi(a_1f_1(H_{i1})),\ldots,\psi(a_Tf_T(H_{iT})))}{\prod_{t=1}^T p(A_{it}|H_{it})} - \sum_{t=1}^T \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t}, \tag{7}$$

where $O_{i,\gamma} = (1-\gamma)Y_i - \gamma R_i$. Again, we introduce a regularization term $\sum_{t=1}^T \lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t}$ to prevent overfitting.

As a remark, note that in (3), the optimal solution is not affected after we subtract any function of H_1 from response variable O_γ. Similar to the augmentation technique used in AOWL, we can replace $O_{i,\gamma}$ by $\widehat{O}_{i,\gamma} = O_{i,\gamma} - \widehat{m}(H_{i1})$, where $\widehat{m}(H_1)$ is an estimator of the conditional expectation of O_γ given baseline feature variables H_1. The refined empirical problem then becomes

$$\max_{(f_1,\ldots,f_T)\in\mathcal{G}_1\times\cdots\times\mathcal{G}_T} \frac{1}{n}\sum_{i=1}^{n}\widehat{O}^{+}_{i,\gamma}\frac{\min(\psi(A_{i1}f_1(H_{i1})),\ldots,\psi(A_{iT}f_T(H_{iT})))}{\prod_{t=1}^{T}p(A_{it}|H_{it})}$$
$$+\frac{1}{n}\sum_{i=1}^{n}\sum_{a_t\in\{-1,1\},a_t\neq A_{it}}\widehat{O}^{-}_{i,\gamma}\frac{\min(\psi(a_1f_1(H_{i1})),\ldots,\psi(a_Tf_T(H_{iT})))}{\prod_{t=1}^{T}p(A_{it}|H_{it})}$$
$$-\sum_{t=1}^{T}\lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t}. \tag{8}$$

When context is clear, we will use $(\widehat{f}_{1,\gamma},\ldots,\widehat{f}_{T,\gamma})$ to denote the solution of (7) or (8).

Computationally, the objective function of (8) can be further decomposed as the difference between two convex functions. Therefore, one can adopt the difference of convex (DC) algorithm (Tao and An 1997) to solve (8) iteratively. When $\{\mathcal{G}_t\}_{t=1}^T$ are RKHS, in each iteration of the DC algorithm, the optimization problem can be further reduced to a quadratic programming problem so it can be efficiently solved using existing software. The details are given in the supplemental material.

2.5. Estimating γ^* Using the Risk Control

Finally, to determine the estimate for γ^*, since the empirical estimator of the risk, that is,

$$\frac{1}{n}\sum_{i=1}^{n}R_i\frac{\prod_{t=1}^{T}\mathbb{I}(A_{it}\widehat{f}_{t,\gamma}(H_{it})>0)}{\prod_{t=1}^{T}p(A_{it}|H_{it})},$$

is not continuous in γ, a small change of γ may lead to a significant risk control violation. Thus, we propose to estimate γ^* based on a smooth approximation to the above function. Specifically, we obtain γ^*'s estimator, denoted by $\widehat{\gamma}$, via bisection method to solve equation

$$\frac{1}{n}\sum_{i=1}^{n}R_i\frac{\min(\psi(A_{i1}\widehat{f}_{1,\widehat{\gamma}}(H_{i1})/\eta),\ldots,\psi(A_{iT}\widehat{f}_{T,\widehat{\gamma}}(H_{iT})/\eta))}{\prod_{t=1}^{T}p(A_{it}|H_{it})}=\tau. \tag{9}$$

Here, $\eta\in(0,1]$ is a small shifting parameter to be chosen data dependently.

3. Theoretical Results

This section presents the theoretical results for the expected reward and risk under the estimated DTRs. Recall that $(g^*_{1,\gamma},\ldots,g^*_{T,\gamma})$ are the optimal decision functions of unconstrained problem (3) and let $(g^*_1,\ldots,g^*_T)$ denote the optimal decision function of original CBR problem (1), then Lemma 1 indicates $(g^*_1,\ldots,g^*_T)$ can be selected as $(g^*_{1,\gamma^*},\ldots,g^*_{T,\gamma^*})$. We wish to obtain a non-asymptotic lower bound for

$$\mathcal{V}(\widehat{f}_{1,\widehat{\gamma}},\ldots,\widehat{f}_{T,\widehat{\gamma}})-\mathcal{V}(g^*_1,\ldots,g^*_T)=E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-E\left[Y\frac{\prod_{t=1}^{T}\mathbb{I}(A_tg^*_t(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right]$$

and an upper bound for

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t)>0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right]-\tau,$$

where $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^T$ are either from the O-learning algorithm or the simultaneous learning algorithm. We assume $\{\mathcal{G}_t\}_{t=1}^T$ to be Gaussian RKHS with bandwidth $\sigma_{n,t}^{-1}$.

We need additional assumptions to characterize the complexity of true optimal decision functions of each unconstrained DTR under different multipliers γ. For any given t and γ, we define

$$\mathcal{H}_{t,\gamma,1}=\{h_t\in\mathcal{H}_t|g^*_{t,\gamma}(h_t)>0\},$$
$$\mathcal{H}_{t,\gamma,-1}=\{h_t\in\mathcal{H}_t|g^*_{t,\gamma}(h_t)<0\},$$

and the Δ-function to be

$$\Delta_{t,\gamma}(h_t)=d(h_t,\mathcal{H}_{t,\gamma,1})I(h_t\in\mathcal{H}_{t,\gamma,-1})+d(h_t,\mathcal{H}_{t,\gamma,-1})I(h_t\in\mathcal{H}_{t,\gamma,1}),$$

where $d(x,\mathcal{S})$ denote the Euclidean distant from point x to set $\mathcal{S}$. We assume the following conditions for $t=1,\ldots,T$.

Assumption 5. For any $\gamma\in[0,1]$, there exist universal positive constants $\{\alpha_t\}_{t=1}^T$ and $K>0$ such that

$$\int_{\mathcal{H}_t}e^{-\frac{\Delta^2_{t,\gamma}(h)}{s}}P_t(dh)\le Ks^{\alpha_td_t/2}$$

holds for $t=1,\ldots,T$. Here, d_t denotes the dimension of $\mathcal{H}_t$ and P_t denotes the density function of the random variable H_t.

As a note, Assumption 5 is a general version of the geometric noise exponent assumption first proposed in Steinwart and Scovel (2007) to establish a fast convergence rate for standard SVM problems. This assumption was later used by Zhao et al. (2015) to prove the convergence rate for O-learning. Our next assumption concerns the discrimination property of Q-function between the two treatments, which is sufficient to establish the convergence rate for the risk control.

Assumption 6. For any $\gamma\in[0,1]$, $t=1,\ldots,T$, set $D_t\subseteq\mathcal{H}_t$ and $\eta_1>0$, then

$$E[|Q_{t,\gamma}(H_t,A_t=1)-Q_{t,\gamma}(H_t,A_t=-1)|\mathbb{I}(H_t\in D_t)]\le\eta_1$$

implies that $P(H_t\in D_t)\le K_1\eta_1$ for some fixed positive constant K_1.

The following theorem gives the non-asymptotic convergence rates for the estimated DTRs using O-learning. To state the theorem, for $t=1,\ldots,T$, let θ_t and θ'_t be two parameters taking value in $(0,\infty)$, ν_t and ν'_t be two parameters taking value in $(0,2)$, $C_{1,t}$ be a positive constant depending on parameters $(\nu_t,\theta_t,d_t,c_1,M,K)$, $C_{2,t}$ be a positive constant depending on $(\nu'_t,\theta'_t,d_t,c_1,M)$, and c'_t be a positive constant depending on $(\nu_t,\theta_t,d_t,c_1,M)$. Furthermore, let C_3 and C_4 be two positive constants depending on (M,K_1) and τ, respectively, then we have the following theorem.

Theorem 1. Under Assumptions 1–6, assume that $\lambda_{n,t} \to 0$, $\sigma_{n,t} \to \infty$, $\lambda_{n,t}\sigma_{n,t}^{d_t} \to 0$, and $\mathcal{H}_t$ is compact for $t = 1,\ldots,T$. Let

$$\epsilon_n = \sum_{t=1}^{T} c_1^{-(1-t)} C_{1,t}\left(\begin{array}{l} \frac{1}{\sqrt{n}}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4} \\ +\lambda_{n,t}\sigma_{n,t}^{d_t} + \sigma_{n,t}^{-\alpha_t d_t} \end{array} \right),$$

$$\xi_n = Tc_1^{-2T}\sum_{t=1}^{T} C_{2,t}\frac{1}{\sqrt{n}}\sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2}\lambda_{n,t}^{-\nu_t'/4}\epsilon_n^{-\nu_t'/2}.$$

Then for $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^{T}$ estimated from the O-learning approaches and sufficient small $\delta > 0$, we have

$$\mathcal{V}(\widehat{f}_{1,\widehat{\gamma}},\ldots,\widehat{f}_{T,\widehat{\gamma}}) - \mathcal{V}(g_1^*,\ldots,g_T^*) \geq -C_4\{C_3 Tc_1^{-T}(\delta+\epsilon_n)+\xi_n\} \tag{10}$$

and

$$E\left[R\frac{\prod_{t=1}^{T}\mathbb{I}(A_t\widehat{f}_{t,\widehat{\gamma}}(H_t) > 0)}{\prod_{t=1}^{T}p(A_t|H_t)}\right] - \tau \leq C_4\{C_3 Tc_1^{-T}(\delta+\epsilon_n)+\xi_n\} \tag{11}$$

hold with probability at least $1-\sum_{t=1}^{T} e^{-c_t'\delta^2 n} - e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$, where $\widehat{\gamma}$ is determined via (9) with $\eta = \epsilon_n/M$.

Let $\nu_t \to 0$, $\nu_t' \to 0$, $\theta_t \to 0$, $\theta_t' \to 0$ and assume that parameter α_t in Assumption 5 can be arbitrarily large for any $t = 1, ..., T$, then Assumption 1 indicates that the right-hand side of (10) and (11) can be both lower and upper bounded by a term of order as close as $O(n^{-\frac{1}{2}})$. Hence, Theorem 1 shows that under the ideal case, the beneficial reward under the estimated rules will be expected as high as the reward under optimal decision rules up to a small loss of order $O(n^{-\frac{1}{2}})$, with an induced adverse risk no exceeding than τ plus an error term also up to order $O(n^{-\frac{1}{2}})$. In terms of the dependency of the errors over time horizon T, when all other parameters are fixed and assume that $\{\nu_t\}_{t=1}^{T}$ and $\{\nu_t'\}_{t=1}^{T}$ are independent of t and go to 0 and $\{\alpha_t\}_{t=1}^{T}$ can be arbitrarily large, the right-hand sides of (10) and (11) are proportional to $Tc_1^{-T}(Tc_1^{-T}n^{-\frac{1}{2}}+\delta)$.

Similar to O-learning, we can obtain the non-asymptotic convergence rate for the estimated DTRs in the MRL approach using the Gaussian kernel. When $\widehat{\gamma}$ is determined via (9), a slightly different discrimination assumption is needed to quantify the impact of using $\widehat{\gamma}$ as an approximation or multiplier γ^* associated with τ. To this end, we assume

Assumption 7. For any $\gamma \in [0,1]$, $D_t \subseteq \mathcal{H}_t \times \{-1,+1\}$, $t \in \{1,\ldots,T\}$ and $\eta_2 > 0$, we assume that

$$E\left[\frac{\mathbb{I}((H_t,A_t)\in D_t)}{\prod_{s=1}^{t}p(A_s|H_s)}U_{t+1}(H_{t+1};g_{t+1,\gamma}^*,\ldots,g_{T,\gamma}^*;\gamma)\right] \leq \eta_2$$

implies $P((H_t,A_t) \in D_t) \leq K_2\eta_2$ for some fixed positive constant K_2. Here,

$$\begin{aligned} U_t(H_t;f_t,\ldots,f_T;\gamma) = E\Bigg[& O_\gamma^+\frac{\prod_{s=t}^{T}\mathbb{I}(A_s f_s(H_s) > 0)}{\prod_{s=t}^{T}p(A_s|H_s)} \\ & + \sum_{a_s\in\{-1,+1\},\, a_s\neq A_s} O_\gamma^-\frac{\prod_{s=t}^{T}\mathbb{I}(a_s f_s(H_s) > 0)}{\prod_{s=t}^{T}p(A_s|H_s)}\Bigg| H_t\Bigg]. \end{aligned}$$

Using the same notation definition as Theorem 1 with C_3 now being a constant depending on (M,K_2), the following theorem gives the theoretical results for the MRL approach.

Theorem 2. Under Assumptions 1–5 and 7, assume that $\lambda_{n,t} \to 0$, $\sigma_{n,t} \to \infty$, $\lambda_{n,t}\sigma_{n,t}^{d_t} \to 0$, and $\mathcal{H}_t$ is compact for $t = 1,\ldots,T$. Let

$$\begin{aligned}\epsilon_n = c_1^{-T}\sum_{t=1}^{T} C_{1,t}\Bigg(&\frac{1}{\sqrt{n}}\left(\sqrt{T}+(T^2c_1^{-3T})^{\nu_t/4}\sigma_{n,t}^{(1-\nu_t/2)(1+\theta_t)d_t/2}\lambda_{n,t}^{-\nu_t/4}\right) \\ &+\lambda_{n,t}\sigma_{n,t}^{d_t}+c_1^{-T}\sigma_{n,t}^{-\alpha_t d_t}\Bigg),\end{aligned}$$

$$\xi_n = Tc_1^{-2T}\sum_{t=1}^{T} C_{2,t}\frac{1}{\sqrt{n}}c_1^{-T\nu_t'/4}\sigma_{n,t}^{(1-\nu_t'/2)(1+\theta_t')d_t/2}\lambda_{n,t}^{-\nu_t'/4}\epsilon_n^{-\nu_t'/2},$$

then for sufficient $\delta \geq 0$ we have inequality (10) and (11) hold with probability at least $1-3e^{-\frac{1}{2}c_1^{2T}M^{-2}\delta^2 n}$, where $\widehat{\gamma}$ is determined via (9) with $\eta = \epsilon_n$.

Similar to before, let $\nu_t \to 0$, $\nu_t' \to 0$, $\theta_t \to 0$, $\theta_t' \to 0$ and assume that parameter α_t in Assumption 5 can be arbitrarily large for any $t = 1,\ldots,T$, then Theorem 2 implies that the right-hand side of (11) and (11) can also be lower and upper bounded by a term of order as close as $O(n^{-\frac{1}{2}})$, the same order as for the estimated DTRs using O-learning approach. In terms of the dependency over time horizon T, we again consider the special situation when $\{\nu_t\}_{t=1}^{T}$ and $\{\nu_t'\}_{t=1}^{T}$ are independent of t and both go to 0 and $\{\alpha_t\}_{t=1}^{T}$ can be arbitrarily large. Then the right-hand sides of (10) and (11) are proportional to Tc_1^{-T} $(Tc_1^{-2T}n^{-\frac{1}{2}}+\delta)$.

The main challenge to establish Theorems 1 and 2 is to show that the estimated multiplier $\widehat{\gamma}$ determined via (9) satisfies with a high probability that the expected risk under $(\widehat{f}_{1,\widehat{\gamma}},\ldots,\widehat{f}_{T,\widehat{\gamma}})$ is close to τ. This can be guaranteed by showing that each unconstrained DTRs problem (3) can be uniformly well estimated for fixed $\gamma \in [0,1]$ under Assumption 5. The proof of both theorems uses concentration inequalities for empirical processes, and detail is given in the supplemental material.

When the treatment assignment probabilities, $p(A_t|H_t)$, are unknown, they can be estimated by $\widehat{p}(A_t|H_t)$ via parametric models. Since $p(A_t|H_t) - \widehat{p}(A_t|H_t)$ has a parametric rate of $O(n^{-\frac{1}{2}})$ and this approximation will contribute an extra error term in the loss functions in (5), (7), and (9) through linearization at each stage t, the estimation of $p(A_t|H_t)$ will induce an additional error of order $O(n^{-\frac{1}{2}})$ to both the beneficial reward and adverse risk bounds. This error can be ignored compared to the error bounds given in the theorems. Thus, the results of both theorems remain to hold. More technical details are provided in Section S.4.4 of the supplemental materials.

As a note, the error bounds in Theorem 2 are obtained assuming $\{\widehat{f}_{t,\widehat{\gamma}}\}_{t=1}^{T}$ to be the global optimal rules of (7). In practice, since the DC algorithm may not guarantee this, one may want to consider different initial values, such as the solution to the problem without any risk constraints and gradually decreasing the risk constraint thresholds with the solution from the previous iteration as an initial value.

4. Simulation Studies

In the first simulation setting, we consider a 2-stage SMART. We first generate 7 baseline feature variables independently from Unif[−1,1], denoted as $(X_1, \ldots, X_7)$. To mimic the patient's evolving health status, we also generate a time-dependent covariate at the two stages, denoted as $(X_{8,1}, X_{8,2})$, using $X_{8,1} = \omega_0 + \omega_1, X_{8,2} = \omega_0 + \omega_2$, where ω_0, ω_1 and ω_2 are independently from Unif$[-0.5, 0.5]$. Treatments at the two stages, A_1 and A_2, take values 1 or −1 with equal probability. Finally, the cumulative reward variable Y and risk variable R are obtained using the following models:

$$Y = 1 - X_1 + X_2 + A_1(X_1 + 0.25X_{8,1} + 0.5) + A_2(X_{8,2} + A_1 + 0.25) + \epsilon_Y,$$

$$R = 2 + X_1 + X_2 + A_1(X_1 - X_2 + 0.5) + A_2(0.5X_1 + 0.5X_3 - X_{8,2} + 1) + \epsilon_R,$$

where $\epsilon_Y = \epsilon_0 + \epsilon_1, \epsilon_R = \epsilon_0 + \epsilon_2$ with ϵ_0 from $N(0, 1)$ truncated at ±0.25 and ϵ_1 and ϵ_2 both from $N(0, 1)$ truncated at ±0.5. In the second simulation setting, the feature variables are generated the same as before except that seven baseline feature variables are from independent Unif[0,1] and ω_0 is from unif$[0.5, 1]$. The cumulative reward Y and risk R are generated using the following nonlinear models:

$$Y = 1 + 2X_2 + A_1(X_{8,1}^2 + 1) + A_2(X_{8,2}^2 + X_1^2) + \epsilon_Y,$$

$$R = 2 - X_2 + A_1(X_1 + 1) + A_2(A_1X_{8,2} + 1) + \epsilon_R,$$

where ϵ_Y and ϵ_R are generated the same way as in the first simulation setting. Note that for both simulation settings, the feature variables at each stage are $H_1 = (X_1, \ldots, X_7, X_{8,1})$ and $H_2 = (H_1, A_1, X_{8,2})$, respectively. We choose the risk constraint $\tau = 1$ for the first simulation setting and $\tau = 1.5$ for the second simulation setting.

We randomly generate the training data for each simulation setting with sample sizes $n = 200$ and $n = 400$. Both linear kernel and Gaussian radial basis kernel are implemented for O-learning and MRL. When the Gaussian kernel is used, we follow Wu, Zhang, and Liu (2010) to choose $\sigma_{n,t}^{-1} = 1.25 * \text{median}_{A_{it} \neq A_{jt}} \|H_{it} - H_{jt}\|$. To choose the tuning parameters (λ_1, λ_2), we fix the tuning grid of $n\lambda_1$ and $n\lambda_2$ to be from $(2^{-8}, 2^{-6}, \ldots, 2^6, 2^8)$. The optimal tuning parameters are then determined via two-fold cross-validation, which yields the highest reward on the testing data.

Specifically to each algorithm in our proposed method, O-learning follows both original OWL from Zhao et al. (2015) and AOWL from Liu et al. (2018). For MRL, we replace the original response variable with its residual as described in Section 2.3, where we estimate the conditional mean via Lasso regression. For each stage, the initial values of MRL are set to be estimated from regression Y_i on the kernel basis functions. The quadratic optimization problem in the DC algorithm can be solved using standard R functions such as *solve_osqp()* from package *osqp* (*https://cran.r-project.org/web/packages/osqp/index.html*). For both O-learning and MRL we determine $\widehat{\gamma}$ via (9) where we set shifting parameter $\eta = 10^{-4}$ and bisection termination condition $\epsilon = 10^{-3}$. We also include Q-learning for comparison, where at each stage of the backward learning, the Q-function is estimated using linear regression with the kernel basis functions and their interactions with treatments as predictors. Finally, to examine the impact of imposing risk control when learning DTRs, we also estimate the unconstrained optimal DTRs by setting $\tau = \infty$.

All simulation studies are repeated 500 times for each setting. An independent testing dataset of sample size 5000 is generated to evaluate the performance and the estimated reward and risk on the independent testing data from each method are reported. To further quantify the benefit-risk tradeoff, we also report the efficacy ratio, one common measure to evaluate the benefit-risk tradeoff (Guo et al. 2010) defined as $r(\mathcal{D}) = (E^{\mathcal{D}}[Y] - E^{\mathcal{D}_0}[Y])/(E^{\mathcal{D}}[R] - E^{\mathcal{D}_0}[R])$, where $\mathcal{D}$ denotes the treatment rules being assessed and $\mathcal{D}_0$ represents the standard treatments. In our simulation study, the standard treatments are selected as the safest treatment rules, which induce the lowest cumulative risk among all four possible one-size-fits-all treatment rules. Since the standard comparison is set to be the treatment that yields the lowest expected risk, a higher efficacy ratio indicates that the treatment rule will gain more reward under the same risk increment than a treatment rule with a lower efficacy ratio. A treatment rule with a large efficacy ratio is preferable.

Table 1 presents the simulation results. For the first linear simulation setting, we note that when no risk constraint is imposed, the expected adverse risk under the unconstrained optimal rules is greater than 3.2, significantly higher than the prespecified risk constraint of $\tau = 1$. When the risk constraint is imposed, using either MRL, OWL, AOWL, or Q-learning yields the rules that give an expected risk below or close to the risk constraint on the independent testing data. This suggests that the proposed estimation procedure effectively finds the estimated rules that meet the risk restriction. Regarding the reward outcome, the theoretical maximum reward under the risk constraint $\tau = 1$ is approximately 2.17. As shown in Table 1, all four methods' testing rewards are close to the optimal value. This demonstrates that our proposed estimation procedure does find the treatment rules that improve the beneficial reward while preserving safety. For O-learning, using either OWL or AOWL yields a similar result, with OWL having better performance in risk control. Comparing different methods, under the linear kernel, MRL and OWL tend to yield more stable and safer rules with median testing risk strictly lower than the risk constraint and with a smaller variability. In contrast, AOWL and Q-learning tend to underestimate the expected risk, leading to a testing risk close to or slightly higher than the risk constraint with larger variability. The performance of the four methods under the Gaussian kernel is similar to that under the linear kernel, except that OWL also tends to underestimate the expected risk and produces higher risk under the Gaussian kernel. MRL generally has the best risk control compared to OWL, AOWL, and Q-learning. No significant difference is observed between different kernels.

For the second simulation setting, the risk under the unconstrained optimal rules can be as high as 4.6 when no risk constraint is imposed, which is also significantly higher than the risk constraint $\tau = 1.5$. When risk restriction is imposed, and the linear kernel is used, the result shows that all methods can control the risk below or close to the risk constraint on the testing data. In terms of reward, the theoretical maximum reward under $\tau = 1.5$ is approximately 3.29. The expected

Table 1. Complete results of simulation studies.

			Setting I			Setting II		
Kernel	n	Method	Testing reward	Testing risk	Efficacy ratio	Testing reward	Testing risk	Efficacy ratio
Linear	200	MRL	1.657(0.166)	0.835(0.169)	1.023(0.150)	2.681(0.046)	1.257(0.139)	1.585(0.064)
		OWL	1.792(0.118)	0.894(0.146)	1.289(0.181)	2.983(0.143)	1.376(0.203)	1.591(0.093)
		AOWL	1.875(0.122)	0.962(0.160)	1.290(0.187)	3.071(0.153)	1.477(0.229)	1.553(0.095)
		Q-Learning	1.844(0.107)	1.062(0.172)	1.043(0.149)	2.959(0.142)	1.516(0.208)	1.476(0.077)
		Unconstrained	2.726(0.055)	3.184(0.090)	0.555(0.015)	4.584(0.006)	4.674(0.028)	0.908(0.003)
	400	MRL	1.707(0.114)	0.863(0.125)	1.098(0.117)	2.681(0.000)	1.257(0.000)	1.528(0.000)
		OWL	1.866(0.092)	0.931(0.126)	1.350(0.157)	3.039(0.101)	1.385(0.144)	1.605(0.065)
		AOWL	1.926(0.089)	0.988(0.127)	1.329(0.134)	3.093(0.109)	1.451(0.153)	1.576(0.064)
		Q-Learning	1.837(0.074)	1.028(0.110)	1.089(0.107)	2.965(0.110)	1.506(0.153)	1.482(0.057)
		Unconstrained	2.757(0.045)	3.223(0.057)	0.560(0.013)	4.587(0.005)	4.685(0.017)	0.907(0.002)
Gaussian	200	MRL	1.460(0.168)	0.726(0.188)	0.845(0.385)	2.805(0.128)	1.257(0.170)	1.586(0.086)
		OWL	1.738(0.105)	1.082(0.170)	0.883(0.213)	2.826(0.171)	1.355(0.231)	1.530(0.107)
		AOWL	1.812(0.140)	1.021(0.192)	1.025(0.183)	3.141(0.177)	1.793(0.277)	1.378(0.087)
		Q-Learning	1.912(0.129)	1.011(0.178)	1.267(0.176)	3.122(0.152)	1.524(0.213)	1.544(0.085)
		Unconstrained	2.825(0.010)	3.452(0.037)	0.535(0.006)	4.587(0.000)	4.702(0.000)	0.905(0.000)
	400	MRL	1.609(0.099)	0.805(0.099)	1.076(0.168)	2.895(0.125)	1.289(0.147)	1.611(0.080)
		OWL	1.784(0.076)	1.042(0.128)	1.052(0.235)	2.891(0.112)	1.333(0.158)	1.567(0.093)
		AOWL	1.849(0.101)	0.996(0.119)	1.200(0.150)	3.168(0.137)	1.740(0.201)	1.419(0.074)
		Q-Learning	1.928(0.087)	0.986(0.122)	1.337(0.141)	3.115(0.099)	1.483(0.147)	1.568(0.063)
		Unconstrained	2.825(0.021)	3.382(0.077)	0.554(0.016)	4.587(0.000)	4.702(0.000)	0.905(0.000)

NOTE: Testing reward, testing risk, and efficacy ratio are reported in *median(dev)* format. *dev* denotes the median of the absolute difference between the estimated value and the median estimated value of 500 repeated simulations.

rewards on testing data using the linear kernel are all below but also close to the theoretical optimal reward, indicating that our proposed estimation procedure can maintain its performance and balance reward and adverse risk under a different simulation setting. Compared with AOWL and Q-learning, MRL and OWL still show better control of the adverse risk, with MRL having stricter risk control and more stability. When the Gaussian kernel is used, the performance of MRL, OWL, and Q-learning slightly improves, but AOWL starts to underestimate the risk with testing risk considerably exceeding $\tau = 1.5$. When the sample size is increased, all four methods using either linear or Gaussian kernel will improve, but AOWL under Gaussian kernel remains to underestimate the risk. The simulation results indicate that under this nonlinear setting, MRL, OWL, and Q-learning can still perform well, with MRL showing better control of the risk similar to the first simulation setting, while AOWL under the Gaussian kernel fails to strictly control the risk.

Regarding the efficacy ratio, in the first simulation setting, OWL has a higher efficacy ratio under the linear kernel, and Q-learning has a higher value under the Gaussian kernel. In the second simulation setting, MRL has roughly the same high efficacy ratio as OWL and a higher efficacy ratio than AOWL and Q-learning under the linear kernel. When the Gaussian kernel is used, MRL tends to achieve the highest efficacy ratio among the four methods. In summary, these simulation results show that MRL, OWL, AOWL, and Q-learning using our proposed estimation procedure can yield the rules that meet the risk restriction while maintaining a high beneficial reward for a CBR problem. MRL tends to have an overall better performance with stricter risk control. Additionally, we conduct a simulation setting with $T = 4$ and allow the treatment assignment probabilities to depend on the covariates at later stages. The results are reported in the supplemental materials and show a similar conclusion: MRL has better risk control and a higher efficacy ratio than the other competing methods.

Computationally, it took around 2 min for Setting I with $n = 200$ for running MRL, which is slower than the other methods. The computing time increases quadratically with the sample size. More details are given in Table S.3 of the supplemental material. We discuss techniques to improve computational efficiency in Section 6.

5. Application to Durable Trial

DURABLE (Fahrbach et al. 2008) was a two-phase randomized trial that aimed to assess the safety and efficacy of insulin glargine therapy versus insulin lispro mix therapy in addition to oral antihyperglycemic agents among T2D patients. Patients qualified during the screening stage would enter the study's first phase and be randomly assigned to either the daily insulin glargine (G) group or the twice daily insulin lispro mix 75/25 (LMx2) group for 24 weeks. By the end of the first phase, patients' HbA1c level was measured to determine future treatment assignments. Patients who failed to reach a safe HbA1c level lower than 7% entered the second phase intensification study and were randomly reassigned to either basal-bolus therapy (BBT) or LMx2 for the insulin glargine group or basal-bolus therapy (BBT) or three times daily insulin lispro mix 50/50 (MMx3) therapy for the LMx2 group and followed for another 24 weeks. The remaining patients who reached HbA1c lower than 7% entered the maintenance study and continued receiving the initial treatment for another two years.

For T2D, the A1c level is the main efficacy outcome measuring the patient's health condition, and we choose HbA1c reduction by the end of week 48 compared to baseline level (week 0) as the cumulative reward outcome. Weight gain is one of the common long-term side effects of insulin therapy. In the analysis, we choose cumulative risk outcome to be BMI change by the end of week 48, with a lower BMI increment indicating better risk control. Due to the DURABLE study design, not all patients were re-randomized during the second phase. To implement CBR, we make a practical assumption that for patients who had reached HbA1c of 7% at the end of the first phase and entered the maintenance study, their treatments received at the second phase of the study were optimal. Hence, for the patients who entered the maintenance study, only their first-stage treatment needs to be evaluated and optimized. With this assumption, for MRL, we solve the modified empirical problem

$$
\begin{aligned}
&\max_{(f_1,f_2)\in\mathcal{G}_1\times\mathcal{G}_2} \frac{1}{n}\sum_{i\in I_1}\widehat{O}_i^{+}\frac{\min(\psi(A_{i1}f_1(H_{i1})),\psi(A_{i2}f_2(H_{i2})))}{p(A_{i1}|H_{i1})p(A_{i2}|H_{i2})}\\
&+\frac{1}{n}\sum_{i\in I_1}\sum_{a_t\in\{-1,1\},a_t\neq A_{it}}\widehat{O}_i^{-}\frac{\min(\psi(a_1f_1(H_{i1})),\psi(a_2f_2(H_{i2})))}{p(A_{i1}|H_{i1})p(A_{i2}|H_{i2})}\\
&+\frac{1}{n}\sum_{i\in I_2}\widehat{O}_i^{+}\frac{\psi(A_{i1}f_1(H_{i1}))}{p(A_{i1}|H_{i1})}\\
&+\frac{1}{n}\sum_{i\in I_2}\sum_{a_1\in\{-1,1\},a_1\neq A_{i1}}\widehat{O}_i^{-}\frac{\psi(a_1f_1(H_{i1}))}{p(A_{i1}|H_{i1})}-\sum_{t=1}^{2}\lambda_{n,t}\|f_t\|^2_{\mathcal{G}_t},
\end{aligned}
$$

where I_1 and I_2 denote the patients who entered the intensification and maintenance studies, respectively. For Q-learning and O-learning, in stage 2, only patients who entered the intensification study are used for estimation. In the first stage, we use all patients for estimation but update the outcomes by their estimated Q-functions for Q-learning, or inverse probability estimator $Y_i\mathbb{I}(A_{i2}\widehat{f_2}(H_{i2}) > 0)/p(A_{i2}|H_{i2})$ for the patients from the intensification study when applying O-learning.

To estimate the optimal decision rules, we extract 20 relevant feature variables as the baseline variables H_1. These variables include baseline HbA1c level, heart rate, systolic/diastolic blood pressure, body weight, body height, BMI, and 7 points self-monitored blood glucose measured at week 0, and demographic variables including patient's age, gender, along with the duration of T2D and three indicator variables of whether patients were receiving oral antihyperglycemic agent of metformin, thiazolidinedione, or sulfonylureas. The second stage feature variables H_2 include all variables in H_1, as well as the patient's stage 1 treatment assignment, heart rate, systolic/diastolic blood pressures, HbA1c, body weight, body height, BMI, and the same 7 points self-monitored blood glucose measured at the beginning of phase 2 study (24 weeks). All covariates are normalized to have mean zero and variance one.

Our analysis includes 573 patients from the intensification study and 771 from the maintenance study. To reduce the impact due to sampling variability, we repeatedly sample 30% of patients as training data and use the remaining 70% of data as testing data to evaluate the performance of estimated rules. The population average BMI change is approximately 1.5, and we repeat the analysis with τ from 1.5 to 1.65, increased by 0.05. We still implement both OWL and AOWL for O-learning, and the tuning grids of $n\lambda_1$ and $n\lambda_2$ for O-learning and MRL are set to be $(2^{-8}, 2^{-6}, \ldots, 2^{6}, 2^{8})$ with tuning pairs skipped when $\frac{\max(\lambda_1,\lambda_2)}{\min(\lambda_1,\lambda_2)} > 4$. The shifting parameter and the termination condition are set to be $\eta = 10^{-4}$ and $\epsilon = 10^{-3}$ similar to the simulation study. Preliminary exploratory analyses indicate that the optimal decision function is highly nonlinear; hence, we use the Gaussian kernel and select the bandwidth similar to the simulation studies. For each risk constraint, we repeat the analysis 100 times using MRL, OWL, AOWL, and Q-learning. For comparison, we also conduct MRL and set the risk constraint to be infinite to estimate the globally optimal decision rules with no risk control.

The estimated reward and risk on testing data are reported in Table 2. From the results, we first note that when no constraint is imposed, the unconstrained estimated optimal treatment rules will yield an overall increment of BMI approximately equal to 1.75 with a gain of 1.70% HbA1c reduction over a 48-week period, which is close to the expected BMI increment and HbA1c reduction induced by the most aggressive LMx2-MMx3 rules among all four one-size-fits-all rules shown in Table 3. In contrast, when the risk constraint is imposed, the expected increment of BMI can decrease from 1.60 to roughly 1.50, which is significantly lower than the unconstrained expected BMI increment at the price of a smaller HbA1c reduction (i.e., decreasing from 1.61% to roughly 1.56%). Comparing four different methods, both MRL, OWL, AOWL, and Q-learning can yield treatment rules with an expected BMI increment below or close to the prespecified constraint under different choices of τ. However, in terms of beneficial reward, MRL can always lead to an equal or higher HbA1c reduction than OWL, AOWL, and Q-learning under all choices of τ. The results indicate that all four proposed methods will still successfully yield treatment rules that meet the risk restriction in real data application, and MRL tends to have top performance with both ideal control over risk and higher gain in beneficial reward compared to OWL, AOWL, and Q-learning.

We also evaluate the capability of balancing the benefit-risk via efficacy ratio against the standard treatment rules. According to Table 3, assigning all patients with insulin glargine as the initial treatment and reassigning patients who fail to reach HbA1c 7% with BBT (G-BBT) yields the lowest risk with an average increment of BMI equal to 1.20 and HbA1c reduction equal to 1.42%. When G-BBT is selected as the standard treatment, the results in Table 2 show that MRL still achieves an overall higher efficacy ratio than other methods. Moreover, MRL can yield treatment rules with an efficacy ratio close to or higher than the unconstrained optimal rules for all τ. This demonstrates that considering risk impact can also lead to a better treatment regimen design with more efficient benefit-risk balancing in DURABLE, and MRL tends to perform better than OWL, AOWL, and Q-learning.

The proportion of patients recommended to each treatment arm by MRL is displayed in Table 2. MRL recommends almost all patients to G-LMx2 rules when the risk constraint is $\tau = 1.50$ and almost all patients to LMx2-MMx3 when the risk constraint is above 1.60. When $\tau = 1.55$, the treatment rules estimated from MRL will recommend LMx2-MMx3 to 24.6% of patients and G-LMx2 to the remaining patients. We show the treatment recommendation distribution in Figure 2, which suggests that younger and more overweight patients with lower baseline HbA1c are more likely to be recommended with the less intensive insulin glargine therapy as the initial treatment. Younger patients with lower HbA1c levels are in better health conditions, and thus, less intensive insulin therapy is preferable following the general T2D management guidance. Moreover, clinical studies suggest that obesity is associated with insulin

Table 2. Analysis results of DURABLE study.

τ	Method	BMI increment	HbA1c reduction	Efficacy ratio	Percentage of LMx2 during Phase I	Percentage of LMx2/MMx3 during Phase II
1.50	MRL	1.508(0.190)	1.589(0.089)	0.516(0.118)	0.1	100.0
	OWL	1.497(0.083)	1.556(0.056)	0.497(0.167)	41.4	80.6
	AOWL	1.530(0.105)	1.578(0.056)	0.484(0.152)	51.0	84.1
	Q-learning	1.557(0.098)	1.583(0.057)	0.428(0.113)	64.4	64.2
1.55	MRL	1.541(0.187)	1.609(0.081)	0.547(0.150)	24.6	100.0
	OWL	1.492(0.083)	1.563(0.053)	0.477(0.153)	41.4	82.9
	AOWL	1.526(0.093)	1.564(0.064)	0.447(0.148)	49.7	83.2
	Q-learning	1.565(0.096)	1.584(0.046)	0.428(0.104)	69.3	68.1
1.60	MRL	1.615(0.139)	1.609(0.079)	0.527(0.115)	99.8	100.0
	OWL	1.499(0.079)	1.570(0.057)	0.478(0.167)	45.7	87.1
	AOWL	1.540(0.124)	1.579(0.081)	0.469(0.138)	59.6	88.8
	Q-learning	1.580(0.102)	1.585(0.055)	0.436(0.103)	72.9	68.8
1.65	MRL	1.622(0.140)	1.614(0.086)	0.502(0.125)	99.8	100.0
	OWL	1.509(0.084)	1.556(0.052)	0.455(0.152)	51.9	86.7
	AOWL	1.544(0.115)	1.581(0.072)	0.461(0.120)	62.6	87.3
	Q-learning	1.594(0.092)	1.592(0.056)	0.423(0.104)	75.4	70.9
∞	Unconstrained	1.750(0.052)	1.699(0.058)	0.501(0.072)	100.0	100.0

NOTE: Results are reported in median(dev) format as the simulation study. BMI, HbA1c, and efficacy ratio are estimated on repeatedly sampled testing data. Efficacy ratios are calculated using G-BBT as reference rules. The percentage of LMx2 during phase I is the proportion of patients recommended with LMx2 treatment as initial treatment. The percentage of LMx2/MMx3 during phase II is the proportion of patients recommended with LMx2/MMx3 as second phase intensification treatment when failed to reach HbA1c $\leq$ 7.0%. Treatment recommendation is estimated for all patients using maximum voting based on 100 repeated analyses.

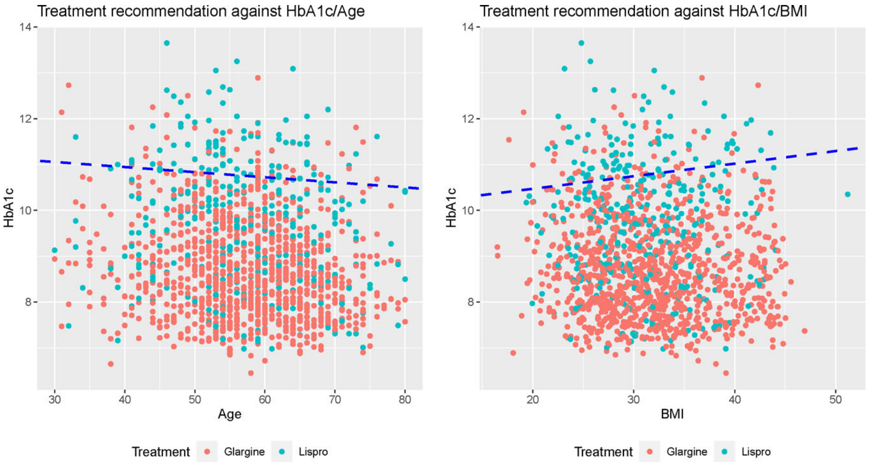

Figure 2. Scatterplot of baseline HbA1c against age or baseline BMI. The color indicates the treatment recommendation estimated from MRL given $\tau = 1.55$. The linear decision boundary is calculated using logistic regression.

Table 3. Mean HbA1c reduction/BMI increment at week 48 under four one-size-fits-all treatment rules.

Treatment rules	Mean BMI increment	Mean HbA1c reduction	Efficacy ratio
LMx2-MMx3	1.738	1.699	0.519
LMx2-BBT	1.683	1.640	0.456
G-LMx2	1.437	1.563	0.610
G-BBT	1.205	1.422	Ref

NOTE: Efficacy ratios are calculated using G-BBT as reference rules.

resistance (Saha and Schwarz 2017), and additional exogenous insulin will cause increased weight gain among T2D patients with insulin resistance (McFarlane 2009). Therefore, patients with higher BMI are more likely to be resistant to insulin and should be treated with less intensive insulin therapy to reduce the risk of weight gain unless the patient's HbA1c level is high. Figure 2 indicates that the treatment rules learned from MRL are consistent with clinical evidence and practices. These results suggest that our proposed method is capable of learning treatment rules that are clinically meaningful in practice while meeting the risk constraint in a real-world application.

6. Discussion

In this work, we proposed a general estimation procedure to solve the CBR problem where the goal is to find optimal treatment rules that maximize the cumulative reward, but the induced risk is no more than a pre-specified threshold. Our approach converts constrained optimization into solving a series of unconstrained optimization problems. Consequently, the proposed procedure can be easily implemented using many existing standard DTR methods or the proposed simultaneous algorithm. Simulation studies and the real data example indicated that using MRL, O-learning, or Q-learning along with the proposed procedure would yield well-performed DTRs with the risk being controlled under or close to the risk constraint.

The proposed MRL can be used to solve unconstrained DTR problems. The key advantage of MRL is that it estimates the DTRs jointly without distinguishing early stages from later stages. This special property would allow one to impose a joint structure on stagewise decision rules to conduct simultaneous variable selection across all stages, which is not feasible using backward Q-learning or O-learning. From the computational perspective, the DC algorithm may be inefficient with a large sample size or a large number of stages. Possible improvement can be using coordinate descent along with stochastic gradient descent to reduce the computation cost of each DC iteration or considering a smoother approximation to the objective functions so that quasi-Newton's methods are applicable.

In some applications, more than one adverse event needs to be controlled in the long run, and CBR can be generalized to handle multiple risk constraints. In addition, CBR can be extended to solve combined short-term stagewise risk control along with cumulative risk control in DTRs. However, further development is needed to address the computational challenges of the case with multiple constraints. The proposed method can also be extended to consider multicategory or continuous treatments (Laber et al. 2018). Although we focused on clinical trials, as discussed in Section 3, our method is applicable to analyze observational studies with the treatment assignment probabilities, that is, propensity scores, being estimated from the data. Finally, when the positivity assumption is a concern, especially for observational studies, some existing techniques such as pessimistic learning (Fu et al. 2022; Zhou et al. 2023) can be incorporated into our framework to learn suboptimal DTRs by working on pessimistic Q-functions.

Acknowledgment

The original research by Dr. Donglin Zeng was conducted in the Department of Biostatistics at the University of North Carolina, Chapel Hill, NC, before the author moved to the current institution.

Supplementary Materials

The supplementary materials contain the algorithms for AOWL and MRL, the proofs to Lemmas 1 and 2 and Theorems 1 and 2, and some additional simulation results.

Disclosure Statement

The authors report there are no competing interests to declare.

References

American Diabetes Association. (2022), "Glycemic Targets: *Standards of Medical Care in Diabetes–2022*," *Diabetes Care*, 45, S83–S96. [1]

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018), "Bandits with Knapsacks," *Journal of the ACM*, 65, 1–55. [1]

Barrett, J. K., Henderson, R., and Rosthøj, S. (2014), "Doubly Robust Estimation of Optimal Dynamic Treatment Regimes," *Statistics in Biosciences*, 6, 244–260. [1]

Butler, E. L., Laber, E. B., Davis, S. M., and Kosorok, M. R. (2018), "Incorporating Patient Preferences into Estimation of Optimal Individualized Treatment Rules," *Biometrics*, 74, 18–26. [1]

Cayci, S., Eryilmaz, A., and Srikant, R. (2020), Budget-Constrained Bandits over General Cost and Reward Distributions, in *PMLR*, pp. 4388–4398. [1]

Chakraborty, B., and Moodie, E. E. M. (2013), *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*, Statistics for Biology and Health, New York: Springer. [1,2]

Cortes, C., and Vapnik, V. (1995), "Support-Vector Networks," *Machine Learning*, 20, 273–297. [4]

Ding, D., Wei, X., Yang, Z., Wang, Z., and Jovanovic, M. (2021), "Provably Efficient Safe Exploration via Primal-Dual Policy Optimization," in PMLR, pp. 3304–3312. [1]

Fahrbach, J., Jacober, S., Jiang, H., and Martin, S. (2008), "The DURABLE Trial Study Design: Comparing the Safety, Efficacy, and Durability of Insulin Glargine to Insulin Lispro Mix 75/25 Added to Oral Antihyperglycemic Agents in Patients with Type 2 Diabetes," *Journal of Diabetes Science and Technology*, 2, 831–838. [8]

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. (2022), "Offline Reinforcement Learning with Instrumental Variables in Confounded Markov Decision Processes," arXiv:2209.08666 [cs, stat]. [11]

Guo, J. J., Pandey, S., Doyle, J., Bian, B., Lis, Y., and Raisch, D. W. (2010), "A Review of Quantitative Risk–Benefit Methodologies for Assessing Drug Safety and Efficacy—Report of the ISPOR Risk–Benefit Management Working Group," *Value in Health*, 13, 657–666. [7]

Huang, X., Shi, L., and Suykens, J. A. (2014), "Ramp Loss Linear Programming Support Vector Machine," *The Journal of Machine Learning Research*, 15, 2185–2211. [4]

Kumar, A., and Wakelee, H. (2006), "Second- and Third-Line Treatments in Non-Small Cell Lung Cancer," *Current Treatment Options in Oncology*, 7, 37–49. [1]

Laber, E. B., Wu, F., Munera, C., Lipkovich, I., Colucci, S., and Ripa, S. (2018), "Identifying Optimal Dosage Regimes under Safety Constraints: An Application to Long Term Opioid Treatment of Chronic Pain," *Statistics in Medicine*, 37, 1407–1418. [11]

Laber, E. B., and Zhao, Y. (2015), "Tree-Based Methods for Individualized Treatment Regimes," *Biometrika*, 102, 501–514. [1]

Lee, J., Thall, P. F., Ji, Y., and Müller, P. (2015), "Bayesian Dose-Finding in Two Treatment Cycles Based on the Joint Utility of Efficacy and Toxicity," *Journal of the American Statistical Association*, 110, 711–722. [1]

Liu, Y., Wang, Y., Kosorok, M. R., Zhao, Y., and Zeng, D. (2018), "Augmented Outcome-Weighted Learning for Estimating Optimal Dynamic Treatment Regimens," *Statistics in Medicine*, 37, 3776–3788. [1,4,7]

Mahdavi, M., Jin, R., and Yang, T. (2012), "Trading Regret for Efficiency: Online Convex Optimization with Long Term Constraints," *The Journal of Machine Learning Research*, 13, 2503–2528. [1]

McFarlane, S. I. (2009), "Insulin Therapy and Type 2 Diabetes: Management of Weight Gain," *The Journal of Clinical Hypertension*, 11, 601–607. [10]

Murphy, S. A. (2003), "Optimal Dynamic Treatment Regimes," *Journal of the Royal Statistical Society*, Series B, 65, 331–355. [1]

Murphy, S. A. (2005), "An Experimental Design for the Development of Adaptive Treatment Strategies," *Statistics in Medicine*, 24, 1455–1481. [2]

Park, C. S., Choi, Y.-J., Rhee, T.-M., Lee, H. J., Lee, H.-S., Park, J.-B., Kim, Y.-J., Han, K.-D., and Kim, H.-K. (2022), "U-Shaped Associations Between Body Weight Changes and Major Cardiovascular Events in Type 2 Diabetes Mellitus: A Longitudinal Follow-up Study of a Nationwide Cohort of Over 1.5 Million," *Diabetes Care*, 45, 1239–1246. [1]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180–1210. [1,2,3]

Qiu, X., and Wang, Y. (2019), 'Composite Interaction Tree for Simultaneous Learning of Optimal Individualized Treatment Rules and Subgroups," *Statistics in Medicine*, 38, 2632–2651. [1]

Saha, S., and Schwarz, P. E. H. (2017), "Impact of Glycated Hemoglobin (HbA1c) on Identifying Insulin Resistance among Apparently Healthy Individuals," *Journal of Public Health*, 25, 505–512. [10]

Socinski, M. A., and Stinchcombe, T. E. (2007), "Duration of First-Line Chemotherapy in Advanced Non–Small-Cell Lung Cancer: Less Is More in the Era of Effective Subsequent Therapies," *Journal of Clinical Oncology*, 25, 5155–5157. [1]

Steinwart, I., and Scovel, C. (2007), "Fast Rates for Support Vector Machines Using Gaussian Kernels," *The Annals of Statistics*, 35, 575–607. [5]

Tao, P. D., and An, L. T. H. (1997), "Convex Analysis Approach to DC Programming: Theory, Algorithms and Applications," *Acta Mathematica Vietnamica*, 22, 289–355. [5]

Wu, Y., Zhang, H. H., and Liu, Y. (2010), "Robust Model-Free Multiclass Probability Estimation," *Journal of the American Statistical Association*, 105, 424–436. [7]

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012), "A Robust Method for Estimating Optimal Treatment Regimes," *Biometrics*, 68, 1010–1018. [1]

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015), "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes," *Journal of the American Statistical Association*, 110, 583–598. [1,3,5,7]

Zhou, Y., Qi, Z., Shi, C., and Li, L. (2023), "Optimizing Pessimism in Dynamic Treatment Regimes: A Bayesian Learning Approach," arXiv: 2210.14420 [cs, stat]. [11]