The Annals of Applied Statistics
2024, Vol. 18, No. 2, 1319–1336
https://doi.org/10.1214/23-AOAS1836

RISK-AWARE RESTRICTED OUTCOME LEARNING FOR INDIVIDUALIZED TREATMENT REGIMES OF SCHIZOPHRENIA

BY SHUYING ZHU[1,a], WEINING SHEN[2,b], HAODA FU[3,d] AND ANNIE QU[2,c]

[1]*Meta, Seattle,* [a]*shuyingzhu@meta.com*
[2]*Department of Statistics, University of California, Irvine,* [b]*weinings@uci.edu,* [c]*aqu2@uci.edu*
[3]*Eli Lilly and Company,* [d]*fu_haoda@lilly.com*

Schizophrenia is a severe mental disorder that distorts patients' perception of reality, and its treatment with antipsychotics can lead to significant side effects. Despite the heterogeneity in patient responses to treatments, most existing studies on individualized treatment regimes only focus on optimizing treatment efficacy, disregarding potential negative effects. To fill this gap, we propose a restricted outcome weighted learning method that optimizes efficacy outcomes while adhering to individual-level negative effect constraints. Our method is developed for multistage treatment decision problems that include single-stage decision as a special case. We propose an efficient learning algorithm that utilizes the difference-of-convex algorithm and the Lagrange multiplier to solve nonconvex optimization with nonconvex risk constraints. We also establish theoretical properties, including Fisher consistency and strong duality results, for the proposed method. We apply our method to a clinical study to design effective schizophrenia treatment [Stroup et al. (*Schizophr. Bull.* **29** (2003) 15–31)] and find that our approach reduces side-effect risk by at least 22.5% and improves efficacy by at least 26.3% compared to competing methods. In addition, we discover that certain covariates, such as the PANSS score, clinician global impressions severity score, and BMI, have a significant impact on controlling side effects and determining optimal treatment recommendations. These results are valuable in identifying subgroups of patients who need special attention when prescribing more aggressive treatment plans.

1. Introduction.

1.1. *Background*. Schizophrenia is a severe mental disorder that causes people to interpret reality abnormally. Antipsychotics are the primary treatments for this condition, but some can cause significant side effects such as weight gain, diabetes, and metabolic syndrome (Mcgurk et al. (1997), Popli et al. (1997), Stroup and Gray (2018)). A recent Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study for schizophrenia found that 55% of patients assigned to atypical antipsychotics experienced weight gain and 20% of them gained at least 15 pounds. This side effect was particularly severe for patients assigned to Olanzapine, which caused an average weight gain of 10.1 pounds. Excess weight gain can lead to a range of health issues, including diabetes, heart disease, and cancer, making it crucial to control weight gain when selecting treatment choices for schizophrenia patients. However, *patient heterogeneity* poses a significant challenge to achieving this goal. Models that only consider the efficacy of antipsychotics are likely to recommend aggressive treatments that could result in severe side effects. Meanwhile, avoiding the most effective antipsychotics for all patients would mean that some patients, who are unlikely to develop severe side effects, would benefit less. Moreover, different patients could have different levels of tolerance to

Received July 2022; revised August 2023.
Key words and phrases. Dynamic treatment regimes, individual-level risk control, individualized treatment regimes, outcome weighted learning, restricted optimization, side effects.

side effects. Therefore, it is more desirable to control side effects at the subject level while maximizing the overall treatment benefits.

To address the challenges discussed above, our paper proposes a new individualized treatment regime (ITR) that recommends antipsychotic treatments by optimizing overall efficacy while controlling weight gain for individual patients. Besides its application to the CATIE study, our proposal is applicable to other medical studies where there is a need to control individual level side effects of treatments, such as complications, drug safety risks, and costs of treatments (Gewandter et al. (2021), Moreau and Wiebels (2021), Read and Williams (2019), Shaw (2018)). Certain medications can pose serious or even life-threatening risks to individuals. For instance, in 2004, the U.S. Food and Drug Administration (FDA) issued a Black Box warning on antidepressants, highlighting an increased risk of suicidal thinking, feelings, and behaviors in children and adolescents based on placebo-controlled trials of these medications. This decision sparked controversy, as there were concerns that it might discourage the appropriate use of antidepressants for depressed patients when clinically indicated. Numerous studies have been conducted to investigate and probe this issue (Fornaro et al. (2019), Friedman (2014), Spielmans, Spence-Sing and Parry (2020)). Another notable example is a study conducted by Watanabe et al. (2021), which revealed that long-term use of drugs commonly prescribed to treat high-blood pressure and heart failure could potentially contribute to kidney damage. While these medications are life-saving for many patients, the study emphasizes the importance of continuing to take them as prescribed. However, it also calls for additional research to better understand the extent of the drugs' long-term effects on the kidneys. These examples underscore the critical need for thorough examination and monitoring of medication side effects, especially those with potentially severe consequences. Balancing the risks and benefits of medications is crucial for ensuring patient safety and optimizing treatment outcomes.

1.2. *Literature review.* Individualized treatment regimes (ITRs) are critical in addressing patients' varying responses to treatments in health studies. The key idea of individualized treatment is to improve health outcomes by tailoring treatments to individual patients based on their information, including genes, health histories, environments, and lifestyles (Gillman and Hammond (2016), Hodson (2016), Kosorok and Laber (2019)). Existing literature primarily focuses on maximizing the expected efficacy of treatments across populations. These methods can be categorized into indirect and direct methods. Indirect methods estimate optimal ITRs by modeling the conditional mean outcomes or their differences among different treatment groups, such as Q-learning (Clifton and Laber (2020), Murphy (2005), Watkins (1989)), G-estimation (Lavori and Dawson (2004), Thall, Sung and Estey (2002)), and regret-based methods (A-learning, regret regression) (Murphy (2003), Robins (2004)). In contrast, direct methods model treatment regimes directly, including the inverse probability weighting estimator (IPW) (Robins, Hernan and Brumback (2000)), augmented inverse probability of treatment weighting (AIPW) (Zhang et al. (2012)), and outcome weighted learning (OWL) (Zhao et al. (2012), Zhao et al. (2015)).

Recent studies have taken side effects into consideration for ITR estimation. For example, Lee et al. (2015) defined a joint utility function, based on binary efficacy and binary toxicity outcomes, and proposed a Bayesian hierarchical latent variable model to find the best dosing for two-cycle treatment based on the utility function. Luckett et al. (2021) and Butler et al. (2018) considered a composite-outcome problem by taking a weighted combination of outcomes. Lakkaraju and Rudin (2017) constructed a decision list using a sequence of if-then-else rules to maximize outcomes and minimize overall costs for populations simultaneously. Wang, Fu and Zeng (2018) proposed to maximize population benefit under a population level risk constraint using regression-model-based learning and an outcome weighted learning algorithm. However, these methods are not directly applicable under our motivating problem

since they only control side effects at the *population level*, that is, requiring the average side effect to be less than a predefined threshold or balancing between the expected weighted average of efficacy and side effect; however, there is no risk control at the *individual level*. Several other studies have also pointed out that considering the population-level efficacy alone is not sufficient and proposed alternative solutions. For example, Fang, Wang and Wang (2023), Qi et al. (2019), and Qi, Pang and Liu (2023) proposed ITRs that maximize the average efficacy while controlling the average outcome from the less predominated subjects. Still, these approaches only consider a single outcome and the constraint targeting the average efficacy from a subpopulation instead of each individual. Liu and Kennedy (2021) suggested utilizing the median of the efficacy, instead of the mean in the optimization process, to add robustness to the decision rule.

Besides patient heterogeneity and the need for controlling individual-level side effects, multistage decision points impose additional challenges for ITR estimation since multiple experimental phases are involved. This is often called multistage ITR or dynamic treatment regimes (DTRs), which commonly arise in chronic disease studies with sequential decision points. Dynamic programming is a prevalent approach for solving multistage ITR problems. For instance, backward Q-learning and backward outcome weighted learning are quite successful in integrating dynamic programming with Q-learning and outcome weighted learning for multistage decision problems (Murphy (2005), Watkins (1989), Zhao et al. (2015)). Alternatively, Zhao et al. (2015) estimate the optimal decision rules in multiple stages simultaneously using a surrogate loss function.

1.3. *Contribution and key findings*. We present a novel approach for developing optimal multistage individualized treatment regimes (ITRs) that satisfy individual-level hard constraints on negative effects. Unlike existing methods that aim to optimize overall benefit for populations while only controlling the average negative effect, our approach considers the maximum negative effect threshold for each subject. To achieve this, we incorporate the individual-level constraint into the outcome weighted learning framework and propose an efficient algorithm that approximates the indicator functions in the objective function and constraints. We employ several computational techniques, such as the Difference of Convex (DC) functions algorithm, Lagrange multiplier, and quadratic programming to solve the optimization problem.

Our work is innovative as follows. First, to the best of our knowledge, we are among the first to consider and solve the ITR learning problem with *individual level* negative effect constraints. Our method is developed under a multistage treatment decision framework, where a single-stage problem is our special case. Second, we provide theoretical justification for our method by establishing Fisher consistency and strong duality properties under regularity conditions. These developments shed light on addressing individual-level constraints in OWL optimization problems.

We highlight two major findings when applying our method to the CATIE study for schizophrenia treatment. In terms of individual risk control, our method yields a remarkable reduction of at least 22.5% on the side-effect of weight gain, while still achieving at least 26.3% improvement in treatment efficacy compared to other competing methods. Additionally, we observe that specific covariates, such as the PANSS score, Clinician Global Impressions Severity Score, and BMI, have significant influences on controlling side effects and determining optimal treatment recommendations. These findings are valuable in identifying patient subgroups that require special attention when considering more aggressive treatment plans. Our findings offer promising avenues for enhancing schizophrenia treatment strategies and ultimately improving patient outcomes.

2. Methodology.

2.1. *Background and notations.* Consider a T-stage decision problem where the number of decision points T is finite and $T \geq 1$. For $t = 1, 2, \ldots, T$, let A_t be the treatment assignment at stage t, where $A_t \in \{-1, 1\}$, and X_t are the covariates observed between stage $t-1$ and t. Without loss of generality, we consider two potential outcomes, where Y_t is the efficacy, benefit, or reward received after receiving treatment A_t, and Z_t is the side effect, risk, adverse event, or cost incurred after receiving treatment A_t. In the CATIE study, there are $T = 2$ stages and two treatment groups. The efficacy is defined as the difference in total scores of positive and negative symptoms between the start and end point at each stage, and the side effect here is weight gain.

We use subscript i to denote patient i and let $\tau_{i,t}$ be the threshold for the side effect at stage t, where $\tau_{i,t}$ can be prespecified. Here (Y, Z, A, X) follow a nondegenerate distribution P. We use H_t to denote the patients' history information up to point t, where $H_1 = X_1$, and $H_t = (X_1, A_1, Y_1, Z_1, \ldots, A_{t-1}, Y_{t-1}, Z_{t-1}, X_t)$ for $t = 2, 3, \ldots, T$. For the single-stage problem with $T = 1$, the subscript t can be dropped for convenience.

For the multistage decision problem, it is common to combine the efficacy and side-effect outcomes over different stages by considering, for example, the endpoint efficacy Y_T, the total efficacy $\sum_{t=1}^{T} Y_t$, and a weighted average of the efficacy $\sum_{t=1}^{T} W_t Y_t$, where W_t is a weighting function for stage t. Here we focus on the total efficacy $\sum_{t=1}^{T} Y_t$ and the total side effect $\sum_{t=1}^{T} Z_t$, although our proposed method can be extended to address other forms of outcomes. We also define $\tau_i = \sum_{t=1}^{T} \tau_{i,t}$ as the summation of the side-effect threshold for subject i.

A dynamic treatment regime (DTR) for a multistage decision problem is a sequence of decision rules, $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_T)$, where $\mathcal{D}_t$ is a map from subject's history information H_t to treatment A_t. The value function under a decision rule $\mathcal{D}$ is defined as an expected efficacy,

$$V^{\mathcal{D}} = E^{\mathcal{D}}\left[\sum_{t=1}^{T} Y_t\right] = \int \sum_{t=1}^{T} Y_t \, dP^{\mathcal{D}}, \tag{1}$$

where $E^{\mathcal{D}}(\cdot)$ is the conditional expectation under probability measure $P^{\mathcal{D}}$ for (H, A, Y, Z), given $A = \mathcal{D}(H)$.

Assume that $P(A = a|H) > 0$ for $a \in \{-1, 1\}$. It is obvious that $P^{\mathcal{D}}$ is absolutely continuous with respect to P. Since $dP^{\mathcal{D}}/dP = I(a = \mathcal{D}(h))/P(A = a|h)$, the expected benefit under a treatment rule $\mathcal{D}$ is

$$V^{\mathcal{D}} = \int \sum_{t=1}^{T} Y_t \, dP^{\mathcal{D}} = \int \sum_{t=1}^{T} Y_t \frac{dP^{\mathcal{D}}}{dP} \, dP = E\left[\frac{\sum_{t=1}^{T} Y_t \prod_{t=1}^{T} I(A_t = d_t(H_t))}{\prod_{t=1}^{T} P(A_t|H_t)}\right]. \tag{2}$$

2.2. *Single-stage restricted outcome weighted learning* (*ROWL*). We first consider a single-stage decision problem, and extend a single-stage decision idea to a multistage problem in the next section. Our goal is to develop an individualized treatment regime (ITR), which maximizes the expected benefit of the population under the constraint that the expected risk for each individual is no greater than a given threshold. Specifically, by the value function presented in (2) and that

$$E\left[\frac{I(A = \mathcal{D}(H))}{P(A|H)} Y\right] = E\left[\frac{Y}{P(A|H)}\right] - E\left[\frac{I(A \neq \mathcal{D}(H))}{P(A|H)} Y\right],$$

we target the optimal treatment rule $\mathcal{D}^*$ that solves the following optimization:

$$\max_{\mathcal{D}} E^{\mathcal{D}}(Y) = \max_{\mathcal{D}} E\left[\frac{I(A=\mathcal{D}(H))}{P(A|H)}Y\right] = \min_{\mathcal{D}} E\left[\frac{I(A\neq\mathcal{D}(H))}{P(A|H)}Y\right] \quad \text{s.t. } r^{\mathcal{D}}(H_i) \le \tau_i \text{ for any } i, \tag{3}$$

where $E^{\mathcal{D}}(\cdot)$ is the conditional expectation under probability measure $P^{\mathcal{D}}$ for (H, A, Y, Z), given $A = \mathcal{D}(H)$, $r^{\mathcal{D}}(H_i)$ is the expected risk of individual i under decision rule $\mathcal{D}$, and τ_i is the risk tolerance threshold for i. Here the threshold τ_i can differ for different subjects.

A key component that distinguishes our method from the existing literature is that the risk under our framework can be controlled at different levels. In particular, the existing methods (Wang, Fu and Zeng (2018), Zhao et al. (2012), Zhou et al. (2017)) only control the overall risk at the population level, while our method targets controlling individual risk by allowing $r^{\mathcal{D}}(H_i) \le \tau_i$. Individual risk control is indeed more challenging; however, it is practically more sensible, especially under scenarios when the side effect could be life-threatening or when there is a hard budget constraint from treatment cost. To deal with the challenges induced by the large number of subject-level risk constraints, we introduce a new restricted outcome weighted learning (ROWL) method.

In the following we first reformulate the objective function to address issues of the objective function in the original outcome weighted learning method raised by Zhou et al. (2017) and Wang, Fu and Zeng (2018): (1) the optimal decision rule for the objective function is not invariant for a simple shift, namely, adding or subtracting a positive constant on the efficacy, and (2) the efficacy in the objective function must be positive.

To solve (1), we use residuals $Y - \mathcal{M}(H)$ instead of relying on the original efficacy Y as outcomes, where $\mathcal{M}(H)$ is a regression function of Y on H. This is because, if the residual is positive, that is, using the current treatment yields efficacy better than expected, then the decision rule should align with the current treatment assignment and vice versa. To solve (2), we define $Y^* = |Y - \mathcal{M}(H)|$, and $A^* = \text{sign}\{Y - \mathcal{M}(H)\}A$, and replace (Y, A) by (Y^*, A^*) in the original optimization problem. The proposed strategy is valid since

$$\begin{aligned}
&\operatorname*{argmax}_{\mathcal{D}} E\left[\frac{I(A=\mathcal{D}(H))}{P(A|H)}Y\right] \\
&= \operatorname*{argmax}_{\mathcal{D}} E\left[\frac{I(A=\mathcal{D}(H))}{P(A|H)}Y^+ - \frac{I(A=\mathcal{D}(H))}{P(A|H)}Y^-\right] \\
&= \operatorname*{argmax}_{\mathcal{D}} E\left[\frac{I(A=\mathcal{D}(H))}{P(A|H)}Y^+ - \left(\frac{Y^-}{P(A|H)} - \frac{I(A\neq\mathcal{D}(H))}{P(A|H)}Y^-\right)\right] \\
&= \operatorname*{argmax}_{\mathcal{D}} E\left[\frac{I(A=\mathcal{D}(H))}{P(A|H)}Y^+ + \frac{I(A\neq\mathcal{D}(H))}{P(A|H)}Y^-\right] \\
&= \operatorname*{argmax}_{\mathcal{D}} E\left[\frac{I(A\times\text{sign}(Y)=\mathcal{D}(H))}{P(A|H)}|Y|\right],
\end{aligned}$$

where $Y^+ = \max(Y, 0)$ is the positive part of Y and $Y^- = \max(-Y, 0)$ is the negative part.

Thus, the optimization in (3) can be written as

$$\min_{\mathcal{D}} E\left[\frac{I(A^*\neq\mathcal{D}(H))}{P(A|H)}Y^*\right] \quad \text{s.t. } r^{\mathcal{D}}(H_i) \le \tau_i \text{ for any } i. \tag{4}$$

For binary treatment problems, the treatment regime $\mathcal{D}(H)$ is usually defined as the sign of a decision function, that is, $\mathcal{D}(H) = \text{sign}(f(H))$, where $f(\cdot)$ can be either linear or nonlinear. To ensure the expected risk $r^{\mathcal{D}}(H_i)$ satisfying these constraints, we let

$r^D(H_i) = m_i + s_i I(D(H_i) = 1)$, where m_i and s_i are functions of H_i, corresponding to the main effect and the treatment effect on the risk, respectively. Moreover, we replace the expectation by its empirical version in the objective function, which leads to an equivalent optimization problem in (4) as follows:

$$\begin{aligned} &\min_f n^{-1} \sum_{i=1}^n \frac{Y_i^*}{P(A_i|H_i)} I\big(A_i^* f(H_i) < 0\big) \\ &\text{s.t. } m_i + s_i I_{f(H_i)>0} \le \tau_i, i = 1, 2, \ldots, n. \end{aligned} \tag{5}$$

There are indicator functions in both the objective function and the constraints in (5). However, the nonconvexity in (5) leads to an NP-hard problem. In order to handle the nonconvexity, a natural way is to approximate the indicator functions (0–1 loss) by surrogate loss functions. Instead, for the indicator function in the objective function, we consider a hinge loss $(1 - A_i^* f(H_i))^+$ (Cortes and Vapnik (1995), Zhao et al. (2012)). For handling the indicator functions in constraints, we approximate them with a shifted ψ-loss (Huang, Shi and Suykens (2014), Wang, Fu and Zeng (2018)), where a shifted ψ-loss is defined as

$$\psi(\delta, x) = q_1(\delta, x) - q_2(\delta, x) = \delta^{-1}(x + \delta)^+ - \delta^{-1}(x)^+,$$

where δ is a small positive number (see Figure S1 in Supplementary Material Section 1 for illustration (Zhu et al. (2024))). There are two main purposes for employing the shifted ψ-loss. The first one is to replace the treatment term $s_i I_{f(H_i)>0}$ with an upper bound to guarantee that the expected risk will be under a threshold. The second one is for computational convenience. Since shifted ψ-loss is the difference between two convex functions, we can utilize the DC algorithm to remove the nonconvexity, as shown in Supplementary Material Section 2.

Let $f(H_i|\beta) = \beta_0 + \sum_{j=1}^n \beta_j K(H_i, H_j)$, where $K(\cdot, \cdot)$ is a kernel function. Here the function $f(H_i|\beta)$ is quite flexible, which can either be a linear function of H_i, by setting the kernel function K to be a linear kernel, or can be nonlinear, by using nonlinear kernel functions such as Gaussian kernels. To avoid overfitting, we add penalization on the parameters to the objective function. After plugging in the surrogate functions and adding the penalization, the optimization problem (5) is converted to

$$\begin{aligned} &\min_{\beta_0, \beta_{(0)}, \xi_i} C \sum_{i=1}^n \frac{Y_i^*}{P(A_i|H_i)} \xi_i + \frac{1}{2} \beta_{(0)}^T K \beta_{(0)} \\ &\text{s.t. } A_i^* f(H_i) \ge 1 - \xi_i, \xi_i \ge 0, \\ &m_i + s_i \delta^{-1} \big[\big(f(H_i) + \delta\big)^+ - \big(f(H_i)\big)^+ \big] \le \tau_i, i = 1, 2, \ldots, n, \end{aligned} \tag{6}$$

where C is a hyperparameter playing a trade-off between the "fitting error" and the regularization. The penalty term $\frac{1}{2}\beta_{(0)}^T K \beta_{(0)}$ is the norm of the reproducing kernel Hilbert space, with K as a kernel matrix, and $\beta_{(0)} = \{\beta_1, \ldots, \beta_n\}^T$.

The optimization problem in (6) can be solved by applying the DC algorithm (Tao and An (1997)) to iteratively update the estimation. Specifically, at each iteration we employ the Lagrange multiplier to convert the primary problem to a dual problem and then utilize quadratic programming to solve the dual problem. We will discuss more details regarding the computational algorithm in Section 6 of the Supplementary Material.

2.3. *Multistage restricted outcome weighted learning.* In this subsection we extend the single-stage ROWL to multistage treatment decision problems, where the total efficacy $\sum_{t=1}^T Y_t$ is the primary outcome of interest and cumulative toxicity $\sum_{t=1}^T Z_t$ is the constraint

with threshold τ_i for subject i. Dynamic programming is a natural approach for solving multistage optimization problems. However, dynamic programming cannot handle the optimization problem with a total constraint for all stages. To solve this issue, we consider estimating the optimal decision rules for all stages simultaneously (Zhao et al. (2015)).

In particular, under the multistage decision setting, the optimization problem is

$$\begin{aligned}(7)\qquad &\max_{\mathcal{D}} E^{\mathcal{D}}\left(\sum_{t=1}^{T} Y_t\right) = \max_{\mathcal{D}} E\left[\frac{\sum_{t=1}^{T} Y_t \prod_{t=1}^{T} I(A_t = d_t(H_t))}{\prod_{t=1}^{T} P(A_t|H_t)}\right],\\ &\text{s.t.} \sum_{t=1}^{T} r_t^{\mathcal{D}}(H_{i,t}) \le \tau_i \text{ for any } i.\end{aligned}$$

The above optimization problem is also an NP-hard problem due to the nonconvexity of the indicator functions from the objective function and the constraints. Similar to the single-stage optimization, we let $r_t^{\mathcal{D}}(H_{i,t}) = m_{i,t} + s_{i,t} I_{f(H_{i,t})>0}$ and apply a shifted ψ-loss function to substitute the indicator functions in the constraints. Consequently, the objective function involves the product of indicator functions $\prod_{t=1}^{T} I(A_t = d_t(H_t)) = \min\{I(A_1 = d_1(H_1)), \ldots, I(A_T = d_T(H_T))\}$. In a single-stage treatment, the corresponding hinge loss is optimized by forcing $Af(H) - 1 \ge \xi$, and $\xi \ge 0$. Similarly, in a multistage process, we let $A_t f(H_t) - 1 \ge \xi$, for $t = 1, 2, \ldots, T$, and $\xi \ge 0$ for optimizing the multi-dimension hinge loss. This is equivalent to approximating the product of indicator functions $\prod_{t=1}^{T} I(A_t = d_t(H_t))$ by $\min(A_1 f(H_1), \ldots, A_T f(H_T), 1)$.

Similar to (6) for the single-stage problem, the empirical form of the optimization problem for the multistage decision problem after plugging in the corresponding surrogate functions becomes

$$\begin{aligned}(8)\qquad &\min_{\beta_{t0}, \beta_{t(0)}, \xi_i} C \sum_{i=1}^{n} \frac{\sum_{t=1}^{T} Y_i^*}{\prod_{t=1}^{T} P_t(A_{i,t}|H_{i,t})} \xi_i + \frac{1}{2} \sum_{t=1}^{T} \beta_{t(0)}^T K \beta_{t(0)},\\ &\text{s.t. } A_{i,t}^* f(H_{i,t}) \ge 1 - \xi_i, \xi_i \ge 0, \text{ for } t = 1, \ldots, T,\\ &\sum_{t=1}^{T} \left[m_{i,t} + s_{i,t}\delta^{-1}\left\{\left(f(H_{i,t}) + \delta\right)^+ - \left(f(H_{i,t})\right)^+\right\}\right] \le \tau_i, i = 1, 2, \ldots, n,\end{aligned}$$

where C is a hyperparameter, K is a kernel matrix of the history information, and $\beta_{t(0)} = \{\beta_{t,1}, \ldots, \beta_{t,n}\}^T$. This optimization problem can be solved by using similar techniques as in the single-stage ROWL.

The multistage ROWL requires a key assumption in that the individual risk process $\{Z_{it}\}_{t=1}^{T}$ can be predicted by $\{m_{it} + s_{it}\mathbb{1}(A_{it} > 0)\}_{t=1}^{T}$ so that $\{m_{it} + s_{it}\mathbb{1}(a_{it}^* > 0)\}_{t=1}^{T}$ corresponds to the counterfactual risks under the treatment $\{a_{it}^*\}_{t=1}^{T}$. This assumption does not necessarily hold when time-varying effect moderation on the risk process exists (Almirall, Ten Have and Murphy (2010)). However, the assumption still holds if the treatment effect on risk is "instantaneous." That is, A_t only affects Z_t, and not $Z_{t+1}, \ldots, Z_T$ via any other time-varying variables. For example, in our simulation setting 2, $X_{2,1}$ is considered an efficacy moderator, and there is a relationship of $A_1 \to X_{2,1} \to Y_2$. Such a moderation effect does not apply to the risk variable Z_2.

3. Theoretical results.

3.1. *Fisher consistency.* We first establish the Fisher consistency property of our method, and show that the obtained optimizer of the proposed method (ROWL) is indeed the optimal

decision rule. Specifically, we show that the decision rule learned by maximizing the objective function in (7) with the 0–1 loss substituted by the surrogate loss function is equivalent to the optimal decision rule in theory. Here we only consider relaxation of the indicators in efficacy, while treating the risk as known; that is, the DC relaxation of indicators in the risk constraints does not need to be considered.

The following assumptions are made to establish our theorems:

(A1) Stable unit treatment value assumption (SUTVA): A subject's potential outcome is not affected by the treatment assignment of other subjects.

(A2) No unmeasured confounders: All sources of confounding are measured.

(A3) Positivity assumption: For a feasible regime d, the probability of treatment regime d $\prod_{t=1}^{T} P_t(d(H_t) = A_t | H_t = h_t) > 0$ for any A_t and $h_t \in \mathcal{H}_t$, where $\mathcal{H}_t$ is the support of h_t.

Assumptions (A1)–(A3) are standard and commonly used in the causal inference literature (Robins (1997), Rubin (1980)). Note that the objective function in our proposed method is developed under the outcome weighted learning framework. We denote

$$V(f_1, \ldots, f_T) = E\left[\frac{\sum_{t=1}^{T} Y_t \min(A_1 f_1(H_1), \ldots, A_T f_T(H_T), 1)}{\prod_{t=1}^{T} P_t(A_t | H_t)}\right],$$

where $f_t \in \mathcal{F}_t$ is a function defined as $f_t(H_{i,t}) = \beta_{t,0} + \sum_{j=1}^{n} \beta_{t,j} K(H_{t,i}, H_{t,j})$, $K(\cdot, \cdot)$ is a kernel function defined in the Section 2.2, and $\mathcal{F}_t$ is the support of f_t. Hence, $V(f_1, \ldots, f_T)$ is the expected total benefit for all T stages if the sequence of decisions $\{\text{sign}(f_1), \ldots, \text{sign}(f_T)\}$ is implemented at each stage. We also define d_t^* as the optimal decision rule at stage t and let $\tilde{f}_t$ be the function learned by the proposed model at stage t, where $t \in \{1, 2, \ldots, T\}$. Then we have the following theorem.

THEOREM 1. *If* $(\tilde{f}_1, \ldots, \tilde{f}_T) \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_T$, *maximize* $V(f_1, \ldots, f_T)$ *over* $\mathcal{F}_1 \times \cdots \times \mathcal{F}_T$ *under the constraint that the expected side effect does not exceed the corresponding threshold, then for* $h_t \in \mathcal{H}_t$, $d_t^*(h_t) = sign(\tilde{f}_t(h_t))$, $t = 1, \ldots, T$.

Theorem 1 states that the proposed optimizer is sign consistent with the optimal decision rule. In Zhao et al. (2015), they propose a simultaneous outcome weighted learning (SOWL) method, which optimizes the efficacy and establishes its Fisher consistency property. However, their method does not impose any constraints on the side effect. Our Theorem 1 can be viewed as a generalization of their method satisfying the side effect constraint. A detailed proof is given in Supplementary Material Section 3.

3.2. *Duality and optimality.* In our algorithm we propose to solve the primal optimization problem by transforming to a dual problem. In this section we provide a theoretical justification for our algorithm and establish its strong duality and optimality property. For simplicity, we focus on a single stage problem, although our theorem and the corresponding proof can also be directly extended for the multistage problem.

As discussed in Section 2.2, we approximate the indicator functions with surrogate functions in both the objective function and the constraints and then apply the DC algorithm to iteratively update the estimate. As shown in Supplementary Material Section 2, the primal

problem at the $(l+1)$th step is

$$
\begin{aligned}
&\min_{\beta} \frac{1}{2}\beta_{(0)}^T K \beta_{(0)} + C\sum_{i=1}^{n}\Big(\frac{Y_i^*}{P(A_i|H_i)}\xi_i + \frac{\varphi_i}{n} + \frac{\zeta_i}{n}\Big) \\
&\quad \text{s.t. } A_i^* f(H_i) \geq 1-\xi_i, \xi_i \geq 0, \\
&\qquad \varphi_i \geq f(H_i|\beta) + \delta, \varphi_i \geq 0, \\
&\qquad \zeta_i \geq f(H_i|\beta), \zeta_i \geq 0, \\
&\qquad \delta m_i + s_i\big(\varphi_i - \hat{\upsilon}_i\big(\beta, \beta^{(l)}\big)\big) I_{s_i>0} - s_i\big(\zeta_i - \hat{\upsilon}_i\big(\beta, \beta^{(l)}\big)\big) I_{s_i\leq 0} \leq \delta\tau_i,
\end{aligned} \tag{9}
$$

for $i = 1, 2, \ldots, n$, where $\beta_{(0)} = \{\beta_1, \ldots, \beta_n\}^T$, $f(H_i|\beta) = \beta_0 + \sum_{j=1}^{n}\beta_j K(H_i, H_j)$, K is a kernel function, $Y_i^* = |Y_i|$, $A_i^* = \text{sign}(Y_i) \times A_i$, $\hat{\upsilon}(\beta, \beta^{(l)}) = f(H_i|\beta) I_i^{(l)}$, $I_i^{(l)} = I_{\{f(H_i|\beta^{(l)})>0\}}$, $\hat{\upsilon}_i(\beta, \beta^{(l)}) = (f(H_i|\beta)+\delta) I_i^{\prime(l)}$, $I_i^{\prime(l)} = I_{f(H_i|\beta^{(l)})+\delta>0}$, and C is a hyperparameter which balances the misclassification error and regularization penalty. The first three inequalities in the constraints in (9) are alternative ways to present the hinge losses $(1 - A_i^* f(H_i))^+$, $(f(H_i|\beta)+\delta)^+$, and $f(H_i|\beta)^+$ such that their positive part functions can be removed from the optimization problem and thus to be solved more easily. We use $C_j \geq 0$ to represent each constraint.

However, the prime optimization problem (9) is still hard to solve directly since it contains many constraints. Therefore, we propose to transform the prime problem to a dual problem through a Lagrange multiplier and then solve the dual problem to obtain an equivalent solution for the prime problem.

THEOREM 2. *Assume K in (9) is a linear or Gaussian kernel function, and let $\hat{\beta}$ be a solution to the corresponding dual problem of (9). If the primal optimization problem in (9) has at least one feasible solution, then $\hat{\beta}$ must be an optimal solution to the primal optimization problem.*

Theorem 2 implies the equivalence between solving the primal problem and its dual problem and further ensures that the estimated $\hat{\beta}$ is optimal for the primal optimization problem under regularity conditions. To prove the optimality of the solution obtained from the dual problem, it suffices to prove the strong duality and the differentiability of C_j with respect to $(\beta, \xi, \varphi, \zeta)$, where $C_j \geq 0$ represents each constraint, as well as the fulfillment of the Karush–Kuhn–Tucker (KKT) conditions, where strong duality means the optimal objective of the primal problem is equivalent to the optimal objective of the dual problem (Boyd and Vandenberghe (2004)). A detailed proof is provided in Supplementary Material Section 4. Theorem 2 still holds if K takes other kernel functions, in addition to linear or Gaussian functions, as long as the strong duality holds, that is, the second derivative of the penalty term with respect to any β_j is greater than 0.

4. Simulation. In this section we investigate the numerical performance of the proposed method in multistage decision problems. We consider four competitive methods for comparison, including the backward outcome weighted learning method (OWL) (Zhao et al. (2015)), the backward Q-learning method (QL) (Murphy (2005)), the benefit and risk personalized treatment rules (BRITR) (Wang, Fu and Zeng (2018)), and the oracle method. Among them OWL and QL are the most popular methods for learning individualized treatment rules without considering side-effect constraints. The BRITR method is a competitive method for maximizing the benefit under the population-level side-effect constraint. In our numerical studies, since we set up a threshold for the summation of the expected risk over all stages for each

TABLE 1

Simulation results under setting 1: *Average efficacy* (*Y-avg*), *median efficacy* (*Y-med*), *average side effect* (*Z-avg*), 90*th percentile of side effect* (*Z-90*), *maximum side effect* (*Z-max*), *percentage of subjects over the threshold* (*Z-pct*), *and their associated standard deviation* (*sd*) *for the proposed method* (*ROWL*) *and four competing methods under different thresholds* τ. *Bold indicates the best performer among all methods, except for the oracle*

τ	Method	Y-avg (sd)	Y-med (sd)	Z-avg (sd)	Z-90 (sd)	Z-max (sd)	Z-pct (sd)
3.5	Oracle	3.31 (0.02)	3.22 (0.03)	1.77 (0.06)	3.22 (0.02)	3.88 (0.05)	2.05 (0.34)
	OWL	**3.14 (0.14)**	**3.18 (0.14)**	2.46 (0.18)	4.38 (0.27)	6.56 (0.41)	29.06 (5.22)
	QL	1.53 (0.53)	1.44 (0.59)	2.94 (0.36)	5.03 (0.77)	7.79 (0.54)	34.07 (8.54)
	BRITR	2.74 (0.32)	2.78 (0.33)	2.00 (0.69)	3.43 (0.76)	5.46 (0.98)	14.99 (13.99)
	ROWL	2.99 (0.19)	3.05 (0.17)	**1.51 (0.11)**	**2.95 (0.15)**	**4.34 (0.25)**	**2.10 (1.31)**
4.0	Oracle	3.41 (0.02)	3.40 (0.03)	1.98 (0.07)	3.65 (0.03)	4.36 (0.05)	1.50 (0.31)
	OWL	**3.14 (0.14)**	3.18 (0.14)	2.46 (0.18)	4.38 (0.27)	6.56 (0.41)	15.81 (4.26)
	QL	1.53 (0.53)	1.44 (0.59)	2.94 (0.36)	5.03 (0.77)	7.79 (0.54)	19.88 (8.65)
	BRITR	2.78 (0.34)	2.82 (0.35)	2.18 (0.83)	3.60 (0.89)	5.60 (1.10)	8.59 (9.66)
	ROWL	3.12 (0.14)	**3.20 (0.11)**	**1.71 (0.14)**	**3.37 (0.18)**	**4.78 (0.24)**	**1.76 (1.13)**
4.5	Oracle	3.46 (0.02)	3.49 (0.03)	2.14 (0.08)	4.01 (0.05)	4.84 (0.06)	1.06 (0.21)
	OWL	3.14 (0.14)	3.18 (0.14)	2.46 (0.18)	4.38 (0.27)	6.56 (0.41)	8.70 (3.24)
	QL	1.53 (0.53)	1.44 (0.59)	2.94 (0.36)	5.03 (0.77)	7.79 (0.54)	14.77 (7.61)
	BRITR	2.77 (0.35)	2.80 (0.35)	2.28 (0.91)	3.69 (0.99)	5.65 (1.14)	5.63 (7.78)
	ROWL	**3.27 (0.09)**	**3.34 (0.08)**	**1.93 (0.16)**	**3.75 (0.18)**	**5.31 (0.29)**	**1.41 (0.95)**

subject, we split the threshold equally at each stage and use it as the population-level constraint when implementing BRITR. For the oracle method, we assume the model structures for the efficacy and risk are known, but the model parameters still need to be estimated from the simulated data. To implement OWL and QL models, we employ the "bowl" and "qLearn" functions in R package "DynTxRegime." The BRITR, the oracle and the proposed method are implemented in R.

Since the CATIE study has two stages with 634 patients and eight covariates, we adopt a similar setting for our simulations. We consider a two-stage setting for illustration. In the following, if not specified, the sample sizes for the training and test sets are 100 and 10,000, respectively. We let the tuning parameter C take values in $\{0.05, 0.1, 0.5, 1, 10, 100\}$, and δ in $\{0.01, 0.05, 0.1, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}$. All the simulation results are based on 100 Monte-Carlo replications. We consider four data generation settings.

For setting 1 and in the first stage, we generate covariates $X_{1,1}, \ldots, X_{1,10}$, followed by an identical and independent Unif$(0, 1)$ distribution, and randomly generate the treatment assignment A_1 from $\{-1, 1\}$ with equal probability. The outcome $Y_1 \sim N(\mu_1, 0.5^2)$, where $\mu_1 = 0.3 - X_{1,1} + X_{1,2} + X_{1,3} + 5(1 - X_{1,1} - X_{1,2})A_1$, and the negative effect $Z_1 \sim N(u_{z1}, 0.1^2)$, where $u_{z1} = 1 + X_{1,1} + (X_{1,1} + X_{1,2})A_1$. For the second stage, we randomly generate the treatment assignment A_2 from $\{-1, 1\}$ with equal probability. The outcome $Y_2 \sim N(\mu_2, 0.5^2)$, where $\mu_2 = 0.2 - X_{1,1} + X_{1,2} + X_{1,3} + (1 - X_{1,1} - X_{1,2})A_2$, and the negative effect $Z_2 \sim N(u_{z2}, 0.1^2)$, where $u_{z2} = 1 + X_{1,1} + (X_{1,1} + X_{1,2})A_2$.

Table 1 and Figure 1 summarize the results for our method and four competing methods. The tuning parameters for our method are chosen as $C = 10, \delta = 3.5$, and we use the linear kernel. The proposed method (ROWL) achieves a very satisfactory performance, as its efficacy and side-effect risks are very close to those obtained by the oracle method. Specifically, the estimated average efficacy obtained by the proposed method is $90.3\% \sim 94.5\%$ of the oracle, and the median efficacy is $94.7\% \sim 95.7\%$ of the oracle. The percentage of subjects exceeding the threshold is also at the same level for the proposed method and the oracle method, and the average side effect as well as the 90th percentile of the side effect of the

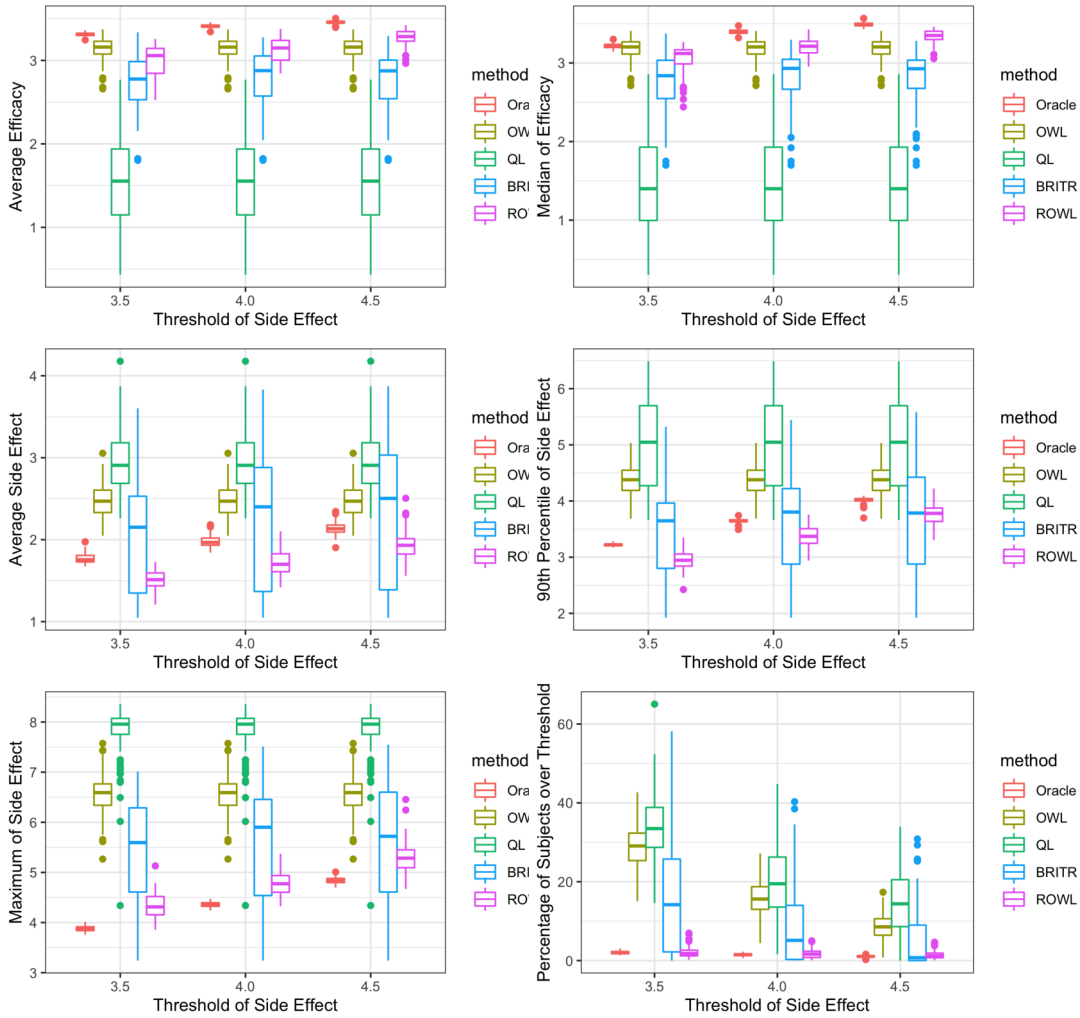

FIG. 1. *Simulation results under setting* 1: *Boxplots of average efficacy* (*upper left*), *median efficacy* (*upper right*), *average side effect* (*middle left*), 90*th percentile of side effect* (*middle right*), *maximum side effect* (*bottom left*), *and percentage of subjects over threshold* (*bottom right*) *for the proposed method* (*ROWL*) *and four competing methods*.

proposed method are smaller compared to those of the oracles. Compared to the other three competing methods, the proposed method achieves a much lower risk and smaller standard deviations at various thresholds τ. This advantage is particularly noticeable in terms of the percentage of subjects exceeding the thresholds, which is reduced by at least 74% for our method compared to competing method. This is expected, since neither the OWL nor the QL methods take risk into consideration; and the BRITR method only considers the population-level risk constraint, which inevitably allows a higher portion of subjects' risk to enable a higher overall efficacy.

We discuss the differences between our method and BRITR as follows. Both methods have the same objective function, that is, to maximize the expected efficacy. However, there are two major differences. The first one is that BRITR only applies to single stage ITRs, which is not applicable to our motivating data study since it consists of more than one stage. Second, the constraints are different. BRITR only requires that the expected population risk not exceed a threshold, while our algorithm requires the expected individual risk to be below a required threshold. Therefore, it is not surprising to observe that, in numerical studies, BRITR tends to

TABLE 2

Simulation results under setting 1 *for different training size* N: *Average efficacy* (*Y-avg*), *median efficacy* (*Y-med*), *average side effect* (*Z-avg*), 90*th percentile of side effect* (*Z-90*), *maximum side effect* (*Z-max*), *percentage of subjects over the threshold* (*Z-pct*), *and their associated standard deviation* (*sd*) *for different threshold* τ. *Bold indicates the best performer among all methods, except for the oracle*

N	Method	Y-avg (sd)	Y-med (sd)	Z-avg (sd)	Z-90 (sd)	Z-max (sd)	Z-pct (sd)
50	Oracle	3.40 (0.02)	3.39 (0.03)	2.00 (0.10)	3.65 (0.04)	4.37 (0.06)	1.55 (0.43)
	OWL	2.86 (0.23)	2.90 (0.24)	2.56 (0.28)	4.39 (0.43)	6.93 (0.54)	15.27 (5.91)
	QL	1.59 (0.57)	1.53 (0.65)	2.88 (0.37)	4.94 (0.80)	7.79 (0.45)	18.64 (8.51)
	BRITR	2.34 (0.40)	2.36 (0.44)	2.27 (0.87)	3.84 (1.10)	6.11 (1.34)	10.04 (12.11)
	ROWL	**3.00 (0.13)**	**3.09 (0.13)**	**1.89 (0.23)**	**3.45 (0.22)**	**4.94 (0.37)**	**2.48 (1.68)**
100	Oracle	3.41 (0.02)	3.40 (0.03)	1.98 (0.07)	3.65 (0.03)	4.36 (0.05)	1.50 (0.31)
	OWL	**3.14 (0.14)**	3.18 (0.14)	2.46 (0.18)	4.38 (0.27)	6.56 (0.41)	15.81 (4.26)
	QL	1.53 (0.53)	1.44 (0.59)	2.94 (0.36)	5.03 (0.77)	7.79 (0.54)	19.88 (8.65)
	BRITR	2.78 (0.34)	2.82 (0.35)	2.18 (0.83)	3.60 (0.89)	5.60 (1.10)	8.59 (9.66)
	ROWL	3.12 (0.14)	**3.20 (0.11)**	**1.71 (0.14)**	**3.37 (0.18)**	**4.78 (0.24)**	**1.76 (1.13)**
200	Oracle	3.41 (0.02)	3.40 (0.03)	1.97 (0.05)	3.65 (0.03)	4.37 (0.06)	1.51 (0.25)
	OWL	**3.30 (0.08)**	**3.32 (0.08)**	2.41 (0.14)	4.41 (0.20)	6.34 (0.30)	16.72 (3.20)
	QL	1.71 (0.55)	1.60 (0.63)	2.81 (0.36)	4.61 (0.64)	7.46 (0.54)	16.65 (7.81)
	BRITR	3.04 (0.20)	3.05 (0.18)	2.15 (0.70)	3.61 (0.72)	5.37 (0.95)	7.39 (7.75)
	ROWL	3.16 (0.15)	3.19 (0.10)	**1.64 (0.12)**	**3.24 (0.14)**	**4.51 (0.19)**	**0.70 (0.64)**

sacrifice a significant portion of subjects in terms of having their risk above the threshold to achieve a higher overall efficacy. This is also reflected in our CATIE data study, where over 19% of patients exceed their risk threshold based on the decision rule from BRITR, which is the highest among all four competing methods.

As τ decreases, the percentage of subjects exceeding the threshold for other nonoracle methods increases, while our method retains a low percentage consistently. The above numerical findings confirm the advantage of our method in terms of controlling individual risk while maintaining a high level of efficacy for multistage treatment decision problems.

We also investigate the impact of the training sample size under this setting and summarize the results in Table 2 and Figure 2. We use the same tuning parameters ($C = 10$, $\delta = 3.5$) and linear kernel K for implementation. From Table 2 we observe that, as the training size N increases, the performance becomes better, that is, the efficacy increases and the side effect decreases. Compared to the oracle, our proposed method achieves a similar level of efficacy and risk, for example, the average efficacy for the proposed method is around 90% of the oracle's. The risk for the proposed method is lower than or comparable to the oracle's under different training sample sizes. Compared to other nonoracle methods, the proposed method has much lower side-effect risks while maintaining competitive efficacy levels.

The computational complexity of our method is $O(n^3L)$, where n is the sample size, and L is the bit-length of matrix W in equation (1) provided in the Supplementary Material File. Here bit-length is defined as $L = \sum_i \sum_j \lceil \log_2(|w_{i,j}| + 1) + 1 \rceil$, where $w_{i,j}$ is the entry of matrix W. The above rate is derived based on the fact that our method mainly uses the interior-point method in each iteration, and the complexity of the interior point is $O(n^3)$ according to Potra and Wright (2000). We further evaluate the computational time of our method and other competing approaches under simulation setting 1 when the training sample size is 100 and τ is 4.0. The comparison results are summarized in Table 3. Specifically, the OWL and QL methods take the least amount of time, followed by BRITR, and the oracle and our proposed method. This is expected, since the OWL and QL methods only consider the average efficacy among the entire population, while BRITR also considers the average risk constraint.

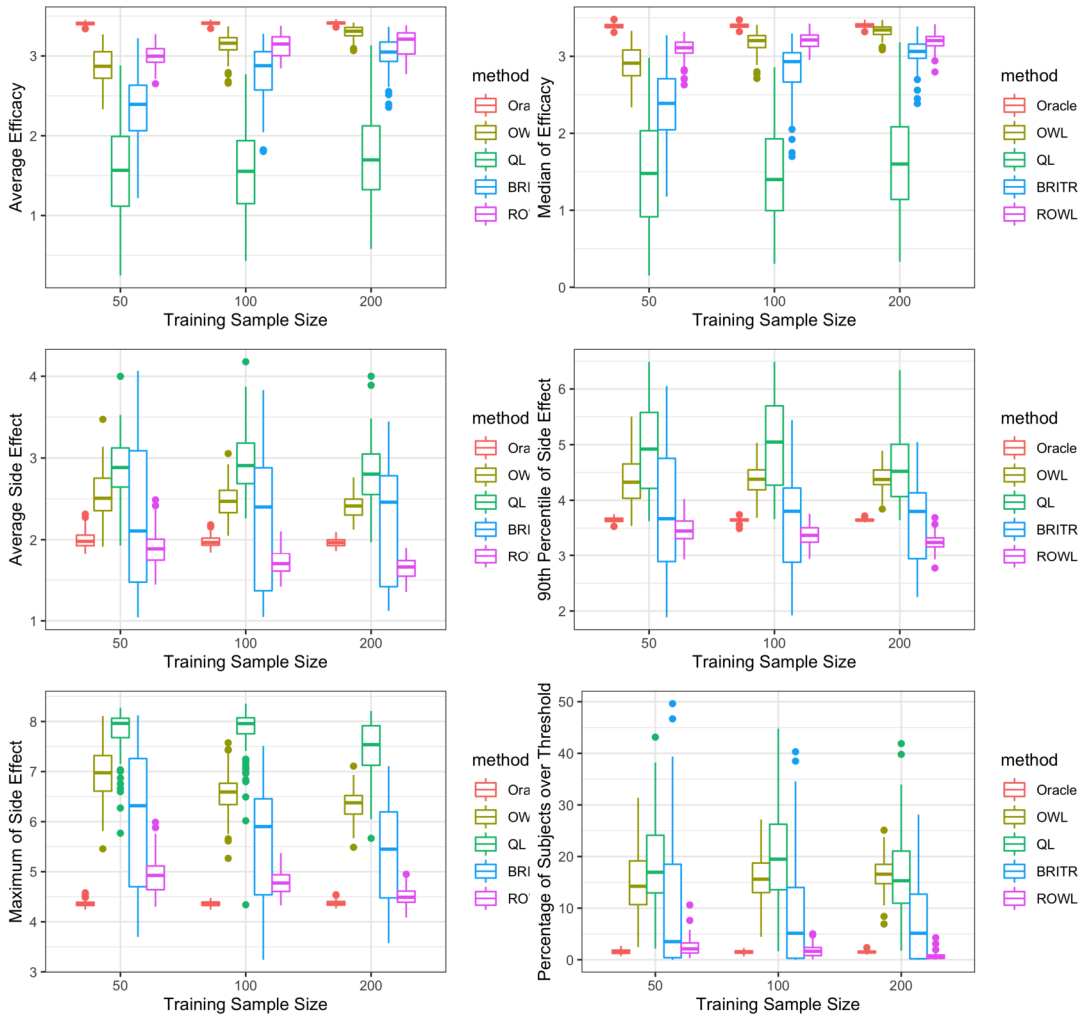

FIG. 2. *Simulation results under setting* 1 *for different training size* N: *Boxplots of average efficacy* (*upper left*), *median efficacy* (*upper right*), *average side effect* (*middle left*), 90*th percentile of side effect* (*middle right*), *maximum side effect* (*bottom left*), *and percentage of subjects over threshold* (*bottom right*).

In contrast, both the oracle and our method impose individual-level risk constraints, which significantly increases the computational cost.

We further evaluate the robustness of the proposed method under treatment mislabeling, where treatment assignment is labeled incorrectly, and summarize the results in Table 4. Both efficacy and side effect do not change much, as the mislabeling rate increases up to 10%. The side-effect risks are more sensitive to treatment mislabeling than efficacy since they are directly related to the accuracy of the individual side-effect prediction. Nevertheless, the risk for the proposed method with 10% mislabeled treatment is still much lower than those

TABLE 3

The average running time (*in seconds*) *of the methods in simulation setting* 1 *when training sample size is* 100 *and* $\tau = 4.0$

Model	Oracle	OWL	QL	BRITR	ROWL
Time (s)	76.58	0.83	0.69	2.78	99.46

TABLE 4

Simulation results for ROWL under setting 1 *with treatment label mislabeling rates of* {0, 2, 5, 10}%: *Average efficacy* (*Y-avg*), *median efficacy* (*Y-med*), *average side effect* (*Z-avg*), 90*th percentile of side effect* (*Z-90*), *maximum side effect* (*Z-max*), *percentage of subjects over the threshold* (*Z-pct*), *and their associated standard deviation* (*sd*) *for different threshold* τ

Rate	Y-avg (sd)	Y-med (sd)	Z-avg (sd)	Z-90 (sd)	Z-max (sd)	Z-pct (sd)
0%	3.12 (0.14)	3.20 (0.11)	1.71 (0.14)	3.37 (0.18)	4.78 (0.24)	1.76 (1.13)
2%	3.14 (0.13)	3.21 (0.11)	1.75 (0.20)	3.40 (0.20)	4.83 (0.24)	2.07 (1.37)
5%	3.13 (0.14)	3.21 (0.13)	1.78 (0.19)	3.44 (0.21)	4.89 (0.25)	2.50 (1.78)
10%	3.11 (0.15)	3.19 (0.13)	1.84 (0.23)	3.51 (0.24)	5.05 (0.32)	3.52 (2.26)

obtained from the other nonoracle methods without mislabeling. We have also conducted three more simulation studies under more complex settings and observed a similar pattern as in Simulation 1. More details can be found in Section 5 in the Supplementary Material.

5. Real data analysis. We apply the proposed method to analyze the data collected from the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study for schizophrenia. The data is obtained from the NIMH Repository and Genomics Resource, available upon request at https://www.nimhgenetics.org/download-tool/SZ under study 17. A detailed study description can be found in Stroup et al. (2003). The CATIE study is an 18-month sequential randomized trial which has a maximum of four phases. Subjects started at Phase 1 upon entering the study, and they were randomly assigned to five treatment groups: (Quetiapine (QUET), Olanzapine (OLZ), Risperidone (RIS), Ziprasidone (ZIP), and Perphenazine (PER). Among these treatments PER is a conventional antipsychotic, and the others are new atypical antipsychotics. During Phase 1 subjects could choose to discontinue their current phase (due to tolerance failure or lack of treatment efficacy) and enter the next phase to receive a new treatment. In Phase 2 the patients who discontinued their treatment in Phase 1, due to tolerance failure, were randomly assigned to a double-blind treatment with a 50% probability of receiving one of the newer atypical antipsychotics (OLZ, RIS, QUET), which they had not previously received, and 50% probability of receiving Ziprasidone (ZIP). Similarly, the patients who discontinued treatment in Phase 1, due to efficacy failure, were randomly assigned to a double-blind treatment with a 50% probability of receiving one of OLZ, RIS, and QUET, which they had not previously taken, and a 50% probability of receiving open-label clozapine (CLO). Since the number of subjects in Phase 3 drops dramatically, we only consider Phase 1 and 2 in our study and treat them as two main stages.

In our analysis we consider the difference in total scores of positive and negative symptoms (PANSSTOT) between the start and end point at each stage as the efficacy outcomes. Let the change of weight at each stage be the side effect. We use the same threshold of 15 pounds for each subject. The covariates include baseline BMI, age, gender, baseline weight, B1_PANSS (baseline PANSS), CGI_SEV (Clinician Global Impressions: Severity), EXACER (hospitalization), and TD (tardive dyskinesia). The total number of subjects is 634.

Based on the previous results in Stroup et al. (2003), we focus on two groups of treatments. One includes OLZ, RIS, QUET, and CLO, which are atypical antipsychotics and are expected to have a higher efficacy and a higher risk of gaining weight, and the other includes PER and ZIP (treatments that are expected to have a lower efficacy and a lower weight gain). For subjects who did not change their treatments, that is, they never entered Phase 2, we split their Phase 1 observations at the middle time point to create two-stage data for further analysis.

For implementation we randomly select 100 subjects as a training set and the rest of the data as a test set. This process is repeated 100 times to obtain the average efficacy and side-effect estimates. The tuning parameters are chosen as $C = 1, \delta = 0.2$, and the linear kernel

TABLE 5

CATIE data analysis results: *Estimated average efficacy* (*Y-avg*), *median efficacy* (*Y-med*), *average side effect* (*Z-avg*), 90*th percentile of side effect* (*Z-90*), *maximum side effect* (*Z-max*), *percentage of subjects over the threshold* (*Z-pct*), *and their associated standard deviation* (*sd*)

Method	Y-avg (sd)	Y-med (sd)	Z-avg (sd)	Z-90 (sd)	Z-max (sd)	Z-pct (sd)
OWL	8.75 (2.44)	7.53 (1.95)	1.48 (2.86)	23.91 (4.76)	86.08 (14.74)	17.71 (3.87)
QL	7.95 (2.30)	7.69 (1.90)	1.48 (2.45)	24.60 (4.37)	85.87 (13.53)	18.62 (4.16)
BRITR	9.92 (3.54)	7.60 (1.87)	2.79 (4.10)	24.76 (4.87)	83.11 (14.77)	19.76 (5.57)
ROWL	**12.53 (4.98)**	**7.91 (1.91)**	**−2.18 (4.05)**	**20.03 (4.93)**	**68.73 (16.53)**	**13.73 (4.98)**

is used. For treatment assignment probability at phase 1, $P(A_1|H_1)$, we use the assignment probability directly since treatments were randomly assigned to the subjects at stage 1. At phase 2 we use logistic regression to predict the probability $P(A_2|H_2)$ with history information. Since we have split the treatment trajectory of patients who only participated in Phase 1 at these two stages. This makes the treatment in the second stage dependent to the first one. In the analysis we standardize the covariates and add their interactions for predicting the side effect (weight gain).

We summarize the results of our method and three competing approaches (OWL, QL, and BRITR) in Table 5 and display their corresponding boxplots in Figure 3. We observe that the proposed method achieves the lowest risk, that is, the average weight gain is, in fact, −2.18 pounds while all other methods have at least 1.48 pounds of weight gain on average. The maximum weight gain for our method is 68.73 pounds compared to at least 83.11 pounds

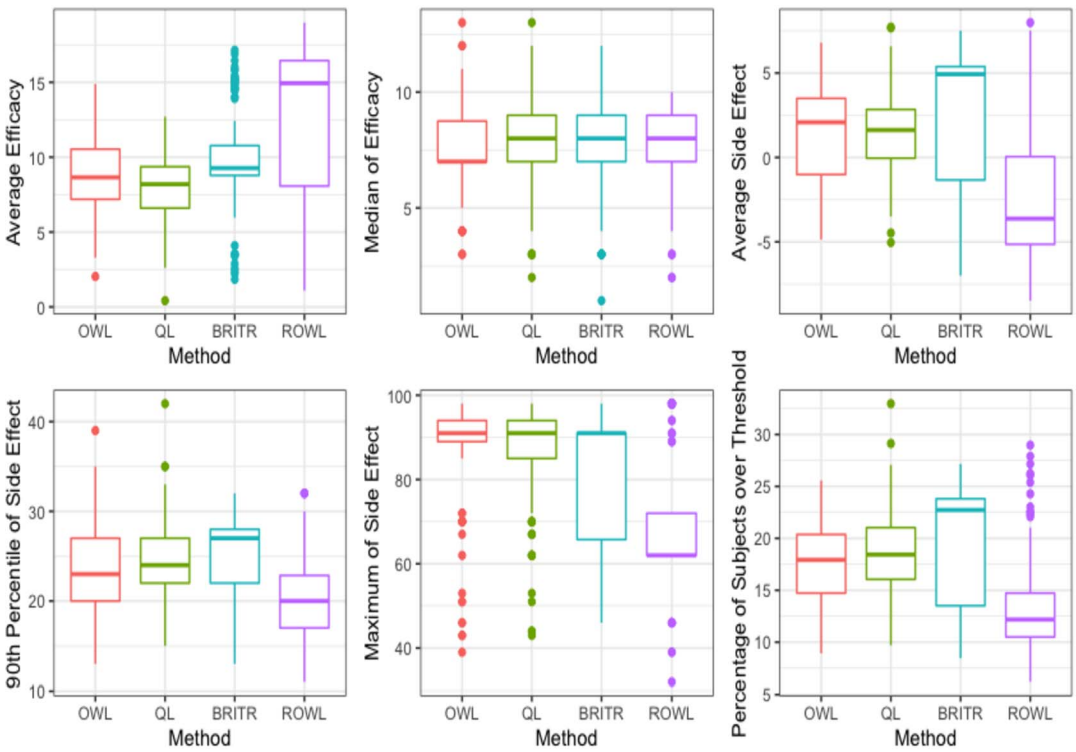

FIG. 3. *CATIE data analysis results*: *Boxplots of average efficacy* (*upper left*), *median efficacy* (*upper middle*), *average side effect* (*upper right*), 90*th percentile of side effect* (*bottom left*), *maximum of side effect* (*bottom middle*), *and percentage of subjects over threshold* (*bottom right*).

from the competing methods. Notably, our method also achieves the highest efficacy. Specifically, the change in PANSSTOT score is 12.53 for our method compared to at most 9.92 from the competing methods. In terms of the percentages of subjects over the threshold, our method reduces the side-effect risk by 22.5%, 26.3%, and 30.5% compared to OWL, QL and BRITR, respectively. In terms of average efficacy, our method significantly increases efficacy by 43.2%, 57.6%, and 26.3% compared to OWL, QL and BRITR. The improvement in side effect is not surprising since we explicitly take individual risk control into consideration in our model. The improvement in the efficacy is likely due to the fact that our method simultaneously learns the model parameters at different stages. In addition, the proposed method has addressed the two issues, that is, the shift invariance and the requirement that the efficacy in the objective function must be positive, which arise in the original outcome weighted learning framework discussed in Section 2.2. These findings illustrate the effectiveness and significant advantages of the proposed individualized risk-control outcome weighted learning method over existing approaches. We have also implemented other values for the side-effect thresholds, and similar results are observed.

In our analysis we have also identified several covariates which have a higher impact than others. When making treatment decisions, it is crucial to consider covariates, such as baseline PANSS score, baseline clinician global impressions severity score (CGI), and BMI, as they are among the most important variables in our model. When predicting side effects, we find that the baseline PANSS score, the weight in phase 2, and the interaction effect between weight and PANSS in both phases are the most important covariates. These results are useful in determining patient subgroups, which may require additional attention when recommending more aggressive treatment plans.

6. Discussion. In this paper we propose a risk-constrained outcome weighted learning method to optimize efficacy while controlling negative effects individually. The proposed method is developed for multistage decision problems, based on the simultaneous outcome weighted learning framework, with the single-stage setting as a special case. Numerical results have confirmed the excellent utility of our method in terms of achieving risk control and treatment efficacy simultaneously.

In the CATIE data analysis, we opt for a linear risk predictive model, as it strikes a good balance between ease of interpretation and predictive accuracy. Our simulation study, specifically Setting 3 as presented in Section 5 of the Supplementary Material, demonstrates the desired robustness of our method against linear model misspecification. Selecting an appropriate risk model and evaluating assumptions, for example, the absence of time-varying moderation on risk as discussed in Section 2, are critical aspects in model building. In practical scenarios if there is prior knowledge available regarding which covariates (or their interactions) should be included or what transformations are needed, then we can utilize this information in constructing a risk predictive model. Otherwise, one can consider fitting several different models (e.g., linear vs. nonlinear) and selecting the most appropriate one based on their performance.

Several research directions remain open for future investigation. First, we focus on predicting and controlling the mean of side effects (over different stages) in this paper. Other statistics, such as percentiles and maxima, can also be used. Second, our proposed method is geared toward binary treatment problems, but it can be extended to solving multitreatment decision problems. For example, one can transform the multitreatment problems to multiple binary treatment problems and then combine pairwise comparisons between treatments using the proposed method. Another approach is to implement comparison between one treatment vs. the others. By sequentially applying our method for binary comparison, an optimal treatment can be obtained. Another future direction is to consider multiple side effects. This can be achieved by adding more constraints to the optimization or using a weighted average of the side effects instead, then applying the proposed procedure to solve the problem.

Funding. Qu's research is supported by the National Science Foundation grant DMS-1952406 and 2210640. Shen's research is supported by Simons Foundation Award 512620.

SUPPLEMENTARY MATERIAL

Examples and algorithms (DOI: 10.1214/23-AOAS1836SUPPA; .pdf). Additional numerical results and the algorithm description are provided.

Sample code (DOI: 10.1214/23-AOAS1836SUPPB; .pdf). Computational codes are provided.

REFERENCES

ALMIRALL, D., TEN HAVE, T. and MURPHY, S. A. (2010). Structural nested mean models for assessing time-varying effect moderation. *Biometrics* **66** 131–139. MR2756699 https://doi.org/10.1111/j.1541-0420.2009.01238.x

BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 https://doi.org/10.1017/CBO9780511804441

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. MR3777922 https://doi.org/10.1111/biom.12743

CLIFTON, J. and LABER, E. (2020). Q-learning: Theory and applications. *Annu. Rev. Stat. Appl.* **7** 279–301. MR4104194 https://doi.org/10.1146/annurev-statistics-031219-041220

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

FANG, E. X., WANG, Z. and WANG, L. (2023). Fairness-oriented learning for optimal individualized treatment rules. *J. Amer. Statist. Assoc.* **118** 1733–1746. MR4646602 https://doi.org/10.1080/01621459.2021.2008402

FORNARO, M., ANASTASIA, A., VALCHERA, A., CARANO, A., ORSOLINI, L., VELLANTE, F., RAPINI, G., OLIVIERI, L., DI NATALE, S. et al. (2019). The FDA "black box" warning on antidepressant suicide risk in young adults: More harm than benefits? *Frontiers in Psychiatry* **10** 294.

FRIEDMAN, R. A. (2014). Antidepressants' black-box warning—10 years later. *N. Engl. J. Med.* **371** 1666–1668. https://doi.org/10.1056/NEJMp1408480

GEWANDTER, J. S., MCDERMOTT, M. P., EVANS, S., KATZ, N. P., MARKMAN, J. D., SIMON, L. S., TURK, D. C. and DWORKIN, R. H. (2021). Composite outcomes for pain clinical trials: Considerations for design and interpretation. *Pain* **162** 1899–1905.

GILLMAN, M. W. and HAMMOND, R. A. (2016). Precision treatment and precision prevention: Integrating "below and above the skin". *JAMA Pediatr* **170** 9–10. https://doi.org/10.1001/jamapediatrics.2015.2786

HODSON, R. (2016). Precision medicine. *Nature* **537** S49. https://doi.org/10.1038/537S49a

HUANG, X., SHI, L. and SUYKENS, J. A. K. (2014). Ramp loss linear programming support vector machine. *J. Mach. Learn. Res.* **15** 2185–2211. MR3231595

KOSOROK, M. R. and LABER, E. B. (2019). Precision medicine. *Annu. Rev. Stat. Appl.* **6** 263–286. MR3939521 https://doi.org/10.1146/annurev-statistics-030718-105251

LAKKARAJU, H. and RUDIN, C. (2017). Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics* 166–175. PMLR.

LAVORI, P. W. and DAWSON, R. (2004). Dynamic treatment regimes: Practical design considerations. *Clin. Trials* **1** 9–20. https://doi.org/10.1191/1740774s04cn002oa

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. *J. Amer. Statist. Assoc.* **110** 711–722. MR3367259 https://doi.org/10.1080/01621459.2014.926815

LIU, L. and KENNEDY, E. H. (2021). Median optimal treatment regimes. ArXiv preprint. Available at arXiv:2103.01802.

LUCKETT, D. J., LABER, E. B., KIM, S. and KOSOROK, M. R. (2021). Estimation and optimization of composite outcomes. *J. Mach. Learn. Res.* **22** Paper No. 167. MR4318523

MCGURK, S. R., GREEN, M. F., WIRSHING, W. C., AMES, D., MARSHALL, B., MARDER, S. R. and MINTZ, J. (1997). The effects of risperidone vs haloperidol on cognitive functioning in treatment-resistant schizophrenia: The trail making test. *CNS Spectr.* **2** 60–64.

MOREAU, D. and WIEBELS, K. (2021). Assessing change in intervention research: The benefits of composite outcomes. *Adv. Methods Pract. Psychol. Sci.* **4** 1–14.

MURPHY, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **65** 331–366. MR1983752 https://doi.org/10.1111/1467-9868.00389

MURPHY, S. A. (2005). A generalization error for Q-learning. *J. Mach. Learn. Res.* **6** 1073–1097. MR2249849

POPLI, A. P., KONICKI, P. E., JURJUS, G. J., FULLER, M. A. and JASKIW, G. E. (1997). Clozapine and associated diabetes mellitus. *J. Clin. Psychiatry* **58** 108–111. https://doi.org/10.4088/jcp.v58n0304

POTRA, F. A. and WRIGHT, S. J. (2000). Interior-point methods. *J. Comput. Appl. Math.* **124** 281–302.

QI, Z., CUI, Y., LIU, Y. and PANG, J.-S. (2019). Estimation of individualized decision rules based on an optimized covariate-dependent equivalent of random outcomes. *SIAM J. Optim.* **29** 2337–2362. MR4008648 https://doi.org/10.1137/18M1190975

QI, Z., PANG, J.-S. and LIU, Y. (2023). On robustness of individualized decision rules. *J. Amer. Statist. Assoc.* **118** 2143–2157. MR4646632 https://doi.org/10.1080/01621459.2022.2038180

READ, J. and WILLIAMS, J. (2019). Positive and negative effects of antipsychotic medication: An international online survey of 832 recipients. *Curr. Drug. Saf.* **14** 173–181. https://doi.org/10.2174/1574886314666190301152734

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality* (*Los Angeles*, *CA*, 1994). *Lect. Notes Stat.* **120** 69–117. Springer, New York. MR1601279 https://doi.org/10.1007/978-1-4612-1842-5_4

ROBINS, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*. *Lect. Notes Stat.* **179** 189–326. Springer, New York. MR2129402 https://doi.org/10.1007/978-1-4419-9076-1_11

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* 550–560.

RUBIN, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *J. Amer. Statist. Assoc.* **75** 591–593.

SHAW, P. A. (2018). Use of composite outcomes to assess risk-benefit in clinical trials. *Clin. Trials* **15** 352–358. https://doi.org/10.1177/1740774518784010

SPIELMANS, G. I., SPENCE-SING, T. and PARRY, P. (2020). Duty to warn: Antidepressant black box suicidality warning is empirically justified. *Front Psychiatry* **11** 18. https://doi.org/10.3389/fpsyt.2020.00018

STROUP, T. S. and GRAY, N. (2018). Management of common adverse effects of antipsychotic medications. *World Psychiatry* **17** 341–356. https://doi.org/10.1002/wps.20567

STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., BYERLY, M. J., GLICK, I. D., CANIVE, J. M., MCGEE, M. F., SIMPSON, G. M., STEVENS, M. C. et al. (2003). The national institute of mental health clinical antipsychotic trials of intervention effectiveness (CATIE) project: Schizophrenia trial design and protocol development. *Schizophr. Bull.* **29** 15–31. https://doi.org/10.1093/oxfordjournals.schbul.a006986

TAO, P. D. and AN, L. T. H. (1997). Convex analysis approach to d.c. programming: Theory, algorithms and applications. *Acta Math. Vietnam.* **22** 289–355. MR1479751

THALL, P. F., SUNG, H.-G. and ESTEY, E. H. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Amer. Statist. Assoc.* **97** 29–39. MR1947271 https://doi.org/10.1198/016214502753479202

WANG, Y., FU, H. and ZENG, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: With an application to treating type 2 diabetes patients with insulin therapies. *J. Amer. Statist. Assoc.* **113** 1–13. MR3803435 https://doi.org/10.1080/01621459.2017.1303386

WATANABE, H., MARTINI, A. G., BROWN, E. A., LIANG, X., MEDRANO, S., GOTO, S., NARITA, I., AREND, L. J., SEQUEIRA-LOPEZ, M. L. S. et al. (2021). Inhibition of the renin-angiotensin system causes concentric hypertrophy of renal arterioles in mice and humans. *JCI Insight* **6** e154337.

WATKINS, C. J. C. H. (1989). Learning from delayed rewards (Ph.D. thesis).

ZHANG, B., TSIATIS, A. A., LABER, E. B. and DAVIDIAN, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68** 1010–1018. MR3040007 https://doi.org/10.1111/j.1541-0420.2012.01763.x

ZHAO, Y., ZENG, D., RUSH, A. J. and KOSOROK, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.* **107** 1106–1118. MR3010898 https://doi.org/10.1080/01621459.2012.695674

ZHAO, Y.-Q., ZENG, D., LABER, E. B. and KOSOROK, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.* **110** 583–598. MR3367249 https://doi.org/10.1080/01621459.2014.937488

ZHOU, X., MAYER-HAMBLETT, N., KHAN, U. and KOSOROK, M. R. (2017). Residual weighted learning for estimating individualized treatment rules. *J. Amer. Statist. Assoc.* **112** 169–187. MR3646564 https://doi.org/10.1080/01621459.2015.1093947

ZHU, S., SHEN, W., FU, H. and QU, A. (2024). Supplement to "Risk-aware restricted outcome learning for individualized treatment regimes of schizophrenia." https://doi.org/10.1214/23-AOAS1836SUPPA, https://doi.org/10.1214/23-AOAS1836SUPPB