Published in final edited form as:
J Am Stat Assoc. 2016 ; 111(516): 1509–1521. doi:10.1080/01621459.2016.1148611.

Personalized Dose Finding Using Outcome Weighted Learning

Guanhua Chen,
Assistant Professor, Department of Biostatistics, Vanderbilt University, Nashville, TN 37203

Donglin Zeng, and
Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599

Michael R. Kosorok
W. R. Kenan, Jr. Distinguished Professor and Chair, Department of Biostatistics, and Professor, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599

Abstract

In dose-finding clinical trials, it is becoming increasingly important to account for individual level heterogeneity while searching for optimal doses to ensure an optimal individualized dose rule (IDR) maximizes the expected beneficial clinical outcome for each individual. In this paper, we advocate a randomized trial design where candidate dose levels assigned to study subjects are randomly chosen from a continuous distribution within a safe range. To estimate the optimal IDR using such data, we propose an outcome weighted learning method based on a nonconvex loss function, which can be solved efficiently using a difference of convex functions algorithm. The consistency and convergence rate for the estimated IDR are derived, and its small-sample performance is evaluated via simulation studies. We demonstrate that the proposed method outperforms competing approaches. Finally, we illustrate this method using data from a cohort study for Warfarin (an anti-thrombotic drug) dosing.

Keywords

DC Algorithm; Dose Finding; Individualized Dose Rule; Weighted Support Vector Regression; Risk Bound

1 Introduction

Dose finding plays an important role in clinical trials, which aim to assess drug toxicity, identify maximum tolerated doses for safety, and determine drug efficacy, while at the same time recognizing the potential that overdosing can increase the risk of side effects and underdosing can diminish the effects of the therapeutic drug. A typical design for dose finding trials is a double-blinded Phase II trial that is conducted to identify the no-effect, the mean effective, and the maximal effective doses (Chevret, 2006). In such trials, patients are randomized to a few fixed safe dose levels for a candidate drug. A single dose level is usually determined by comparing the average outcome of each dose level and then extrapolating for future recommendations. However, this one-size-fits-all approach is not optimal when treatment responses to the drug are heterogeneous among patients, since what works for some patients may not necessarily work for others. Hence, personalized treatment, which tailors drug dose levels according to individual health conditions and disease prognosis, is desirable. Furthermore, the flexibility of using continuous individualized dose levels can help improve clinical practice and increase a patient's compliance. For example, it is widely known that Warfarin, a common drug for the prevention of thrombosis and thromboembolism, is administered in varying doses, ranging from 10mg to 100mg per week depending on the patient's individual clinical and genetic factors (The International Warfarin Pharmacogenetics Consortium, 2009).

As the future of personalized medicine continues to gain prominence as a vital mechanism for the effective treatment of multiple diseases and conditions, there is a growing amount of literature dedicated to the development of study designs and methods for personalized treatment, yet these approaches remain restricted to a finite number of options. A randomized trial design, where the finite treatment options are randomly assigned to patients, is the most valid design to estimate the optimal individualized treatment rule (ITR), which is the treatment decision rule determined by a patient's pre-treatment condition that aims to maximize the expected beneficial clinical outcome, also known as the reward. To infer the ITR using data from such a trial, earlier work has proposed certain parametric or semiparametric models to group patients into subgroups according to their risk levels (Eagle et al., 2004; Marlowe et al., 2007; Cai et al., 2010). However, the parametric/semiparametric model assumption is likely to be invalid due to the complexity of disease mechanisms and individual heterogeneity. More recently, the literature has proposed the use of powerful statistical learning methods, which perform well in analyzing data with high complexity and can be categorized into two groups, namely indirect methods and direct methods. The most popular indirect method is called Q-learning, which is essentially a two-step regression-based approach (Qian and Murphy, 2011; Chakraborty and Moodie, 2013; Moodie et al., 2014; Zhao et al., 2009). Its first step consists of fitting a regression model for the conditional expectation of the reward given the treatments, covariates, and the treatment-covariate interactions. The regression model can be either a parametric regression with complexity penalization, such as LASSO, or a nonparametric learning method, such as a regression tree (Breiman et al., 1984) or support vector regression (SVR, Vapnik (1995); Smola and Schölkopf (2004)). In the second step, the optimal treatment for a given covariate value is obtained as the treatment option that maximizes the predicted mean reward estimated from the regression model in the first step. Some alternative regression-based methods based on regret or contrasts between treatment responses have also been suggested for the first step (Robins, 2004; Moodie et al., 2009; Henderson et al., 2010; Schulte et al., 2014; Wallace and Moodie, 2015), which are equivalent to Q-learning in a randomized trial setting (Schulte et al., 2014).

In regards to the potential over-fitting in the first step of Q-learning, which can lead to a less than optimal ITR, direct methods have been proposed by Zhao et al. (2012). Specifically, Zhao et al. (2012) introduced the framework of outcome weighted learning (O-learning) to directly find the optimal binary treatment rule, where the problem of finding the optimal ITR is formulated as a weighted binary classification with the rewards as weights. The authors demonstrated superior performance of O-learning over indirect methods especially when the sample size is small, which is not uncommon in clinical trials. In other relevant work, Zhang et al. (2012a,b) proposed a robust semiparametric regression function for maximization to infer the optimal rule; comparatively, the O-learning approach in Zhao et al. (2012) is more robust and can handle higher dimensional covariates.

Extending the current designs and methods to personalized dose finding is not trivial because there can be an infinite number of treatment options per a given interval. To infer the optimal individualized dose rule (IDR), a good design should ensure each candidate dose level has a potential chance to be observed in the trial and the different outcomes from patients receiving different dose levels should only be attributed to dose level difference. Therefore, a randomized dose trial design, where each patient receives a dose level randomly chosen from a continuous distribution within the range of safe doses, should be adopted. Indeed, we will show that such a design can lead to a consistent optimal IDR asymptotically.

When analyzing data from such a design, the existing methods for handling finite treatment options can no longer be applied. First, for Q-learning, the performance can be sensitive to the regression model specification in the first step. To identify the IDR, both the main and interaction effect of a continuous dose variable and many covariates on the reward need to be correctly specified. Furthermore, since the regression model in the first step has a high potential for nonlinear interactions, the optimization in the second step will be a nonlinear optimization problem for each given set of covariate values, which can be very unstable and computationally intensive when the dimension of covariates is not small. On the other hand, existing direct methods, including for example outcome weighted learning, which maximizes the value associated with each IDR by using patients whose dose assignment follows this rule, are not applicable since only a few patients are likely to be given the dose level specified by the rule for each covariate value. This is because the dose level follows a continuous distribution, so that the probability of observing a dose equal to the rule-specified dose is zero. We will further elaborate on this in the next section.

In this paper, we advocate the randomized trial design for personalized dose finding and show that this design leads to a consistent optimal IDR. Using data from such a design, we propose a robust outcome weighted learning method to infer the optimal IDR. Our proposed method is a non-trivial extension of O-learning for binary treatments proposed by Zhao et al. (2012). Specifically, we show that the dose finding problem is a weighted regression with individual rewards as weights. We then propose a nonconvex loss function for optimization, and provide a difference of convex functions (DC) algorithm to solve the corresponding optimization problem. We show that this loss function can lead to consistent estimation of the optimal dose. As far as we know, this is the first direct method for estimating IDR where treatment options are on a continuum.

The outline of the rest of the paper is as follows: In Section 2, we first introduce appropriate background information on IDRs and provide an overview of O-learning for binary treatment options. We then discuss how to extend O-learning for estimating an optimal IDR using data from a randomized dose trial, such that there is no unmeasured confounding. Section 3 describes the DC algorithm for solving the optimization problem. The theoretical properties of our method are provided in Section 4. Since existing data comes mostly from observational studies, we extend our approach to analyze such data in Section 5. In Section

6, we demonstrate through simulation studies that our proposed method can identify optimal IDRs which lead to a better predicted clinical outcome than competing methods. In Section 7, the proposed method is applied to analyzing data from a Warfarin study. A concluding discussion is given in Section 8. Most of the proofs are deferred to the Appendix. Sample simulation codes are included as the web supplementary file.

2 Method

2.1 Individualized Dose Rule

As previously mentioned, we consider a randomized dose trial where each patient receives a dose level randomly selected from a continuous distribution in an interval within a safe dose range. Particularly, this assignment can be achieved using a random number generator. Thus, we assume the data are collected from a randomized trial with dose assignment A, which could have many dose levels even in this finite sample, and we assume patient-level covariates consist of a d-dimensional vector $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T \in \mathscr{X}$. The traditional dose finding trial is usually designed to identify the best dose level among some small number of fixed dose levels. In contrast, we allow A to be continuous over a bounded interval $\mathscr{A}$ (a safe dose range). Without loss of generality, we assume that $\mathscr{A} = [0, 1]$. The clinical beneficial outcome, the reward, is denoted as R and assumed to be bounded and positive. An IDR is a map $f: \mathscr{X} \to \mathscr{A}$ that outputs a dose suggestion based on a patient's covariate value. A value function corresponding to this IDR, as denoted by $\mathscr{V}(f)$, is defined as the expected reward in the population if the dose levels for the subjects follow the rule f, i.e., $A = f(\mathbf{X})$. Specifically, if we let $R^*(a)$ be the outcome that would be observed if the dose level a were given, then $\mathscr{V}(f) = E[R^*(f(\mathbf{X}))]$.

We assume the Stable Unite Treatment Value Assumption (SUTVA) (see Rubin (1978)). That is, $R = \Sigma_a I(A = a)R^*(a)$. Then under a randomized dose design, since A is independent of $R^*(a)$ given $\mathbf{X}$, we obtain

$$\begin{aligned}\mathscr{V}(f)&=E[R^*(f(\mathbf{X}))]\\&=E_{\mathbf{X}}[E\{R^*(f(\mathbf{X}))|\mathbf{X}\}]=E_{\mathbf{X}}[E\{R^*(f(\mathbf{X}))|A=f(\mathbf{X}),\mathbf{X}\}]\\&=E_{\mathbf{X}}[E\{R|A=f(\mathbf{X}),\mathbf{X}\}].\end{aligned} \tag{1}$$

As a result, the optimal rule $f_{opt} = \operatorname{argmax}_f \mathscr{V}(f)$. Note that f_{opt} does not change if R is replaced by $R + g(\mathbf{X})$ for any known function $g(\mathbf{X})$ (a common choice of $g(\mathbf{X})$ is a constant function). Also, f_{opt} is invariant in the scale of R. Thus, without loss of generality, we assume that R is positive by subtracting from R its lower bound.

Therefore, we conclude that for any IDR f, $\mathscr{V}(f)$ can be estimated consistently using the mean of the reward among subjects whose dose levels are the same as $f(\mathbf{X})$. In other words, a randomized dose design can consistently evaluate the values from all IDRs to lead to a consistent optimal IDR when the sample size is large enough. This justifies the use of such a design in personalized dose finding.

2.2 O-learning for personalized dose finding

In this section, we propose a learning method to estimate f_{opt} using data from a randomized dose trial which consist of n observations $(A_i, \mathbf{X}_i, R_i)$, $i = 1, \dots, n$. The essential idea is to use empirical data to estimate $\mathscr{V}(f)$ then directly optimize this estimated function for f_{opt}. Specifically, we extend the O-learning method in Zhao et al. (2012) for binary treatment options to our setting, but with non-trivial modifications.

When A is a binary treatment, both Qian and Murphy (2011) and Zhao et al. (2012) show that $\mathscr{V}(f)$ is equal to $E[RI(A = f(\mathbf{X}))/p(A|\mathbf{X})]$, where $p(a|\mathbf{X})$ is the randomization probability of $A = a$ given $\mathbf{X}$. Therefore, $\mathscr{V}(f)$ can be empirically approximated by $n^{-1}\sum_{i=1}^{n} R_i I(A_i = f(\mathbf{X}_i))/p(A_i|\mathbf{X}_i)$. Thus, estimating f_{opt} can be carried out by maximizing the above function, which is also equivalent to minimizing

$$\mathscr{R}_n(f) = n^{-1}\sum_{i=1}^{n} R_i I(A_i \neq f(\mathbf{X}_i))/p(A_i|\mathbf{X}_i).$$

However, since minimizing this function is infeasible due to the discontinuity of the indicator function, the O-learning method in Zhao et al. (2012) proposes to minimize a surrogate version of the above loss function by replacing the indicator function with a hinge loss function defined as

$$n^{-1}\sum_{i=1}^{n} R_i I(1 - A_i f(\mathbf{X}_i))_+/p(A_i|\mathbf{X}_i) + \lambda_n \|f\|^2,$$

where $x_+ = \max(x, 0)$, $\|f\|$ is the seminorm for f from a normed space (usually, a reproducing kernel Hilbert space) and λ_n is a tuning parameter. Equivalently, O-learning is a weighted version of a support vector machine (SVM), where each subject is weighted by his/her reward value R_i. For the binary treatment options, this minimization leads to a consistent optimal treatment rule.

Direct adaption of O-learning to our dose finding setting is not feasible when A is continuous. First, the probability $p(A|\mathbf{X})$ is always zero since A is continuous. Second, in the empirical approximation to $\mathscr{V}(f)$ in O-learning, only a few subjects satisfy $A_i = f(\mathbf{X}_i)$, so this approximation is very unstable. On the other hand, it may be tempting to modify $\mathscr{R}_n(f)$ to resolve these two issues by replacing $p(A_i|\mathbf{X}_i)$ with the density function of A_i given X_i and replacing $I(A_i \neq f(\mathbf{X}_i))$ with a smooth loss, such as $(A_i - f(\mathbf{X}_i))^2$ or $|A_i - f(\mathbf{X}_i)|$. However, our Lemma 1 in the Appendix shows that the estimated dose rule resulting from such a loss is not consistent for the optimal IDR. Therefore, this motivates us to find a different approximation to $\mathscr{V}(f)$ as detailed in the following.

First, assuming $E[R|A = a, \mathbf{X}]$ to be continuous in a, we note that

$$\lim_{\phi\to 0+}\frac{E[RI[A\in(a-\phi,a+\phi)]/p(A|\mathbf{X})|\mathbf{X}]}{2\phi}=E[R|A=a,\mathbf{X}],$$

where $p(a|\mathbf{X})$ is the conditional density of $A = a$ given $\mathbf{X}$, which is known and positive by design. As a result,

$$\begin{aligned}\lim_{\phi\to 0+}E\left\{\frac{RI[A\in(f(\mathbf{X})-\phi,f(\mathbf{X})+\phi)]}{2\phi p(A|\mathbf{X})}\right\}&=E_{\mathbf{X}}\left[E\{R|A=f(\mathbf{X}),\mathbf{X}\}\right]\\&=\mathscr{V}(f).\end{aligned}$$

If we let

$$\tilde{\mathscr{V}}_{\phi}(f)=E\left[\frac{RI[A\in(f(\mathbf{X})-\phi,f(\mathbf{X})+\phi)]}{2\phi p(A|\mathbf{X})}\right],$$

then $\tilde{\mathscr{V}}_{\phi}(f)$ approximates $\mathscr{V}(f)$ when ϕ is sufficiently small. Hence, an IDR maximizing $\tilde{\mathscr{V}}_{\phi}(f)$, or equivalently, minimizing

$$E\left[\frac{R}{2\phi p(A|\mathbf{X})}\right]-\tilde{\mathscr{V}}_{\phi}(f)=E\left[\frac{RI(|A-f(\mathbf{X})|>\phi)}{2\phi p(A|\mathbf{X})}\right], \tag{2}$$

will be close to the optimal IDR. The zero-one loss $I(|A - f(\mathbf{X})| > \phi)$ in the expression often causes difficulty when optimizing using empirical data (Zhang, 2004). Thus, a continuous surrogate loss to replace the zero-one loss is more desirable. Specifically, in our method, we choose such a surrogate loss to be

$$\ell_{\phi}(A-f(\mathbf{X}))=\min\left(\frac{|A-f(\mathbf{X})|}{\phi},1\right).$$

Note that $\ell_{\phi}(x)$ is the difference of two convex functions $|x|/\phi$ and $(x/\phi - 1)_{+}$. Figure 1 plots the curve of $\ell_{\phi}(x)$ and these two convex functions. Clearly, when ϕ goes to zero, $\ell_{\phi}(x)$ converges to the indicator function $I(x \neq 0)$. After replacing the zero-one loss by the new surrogate loss in (2), the objective function to be minimized for the optimal IDR becomes

$$\mathscr{R}_{\phi}(f)=E\left[\frac{R\ell_{\phi}(A-f(\mathbf{X}))}{\phi p(A|\mathbf{X})}\right],$$

and the counterpart to $\tilde{\mathscr{V}}_{\phi}(f)$ is

$$V_\phi(f)=E\left[\frac{R}{\phi p(A|\mathbf{X})}\right]-\mathscr{R}_\phi(f)=E\left[\frac{R\max(1-|A-f(\mathbf{X})|/\phi,0)}{\phi p(A|\mathbf{X})}\right].$$

Finally, using the data from a randomized dose trial, we propose to minimize the empirical version of $\mathscr{R}_{\phi_n}(f)$ for some constant ϕ_n, which will be chosen data adaptively, denoted by

$$\hat{\mathscr{R}}_{\phi_n}(f)=\frac{1}{n}\sum_{i=1}^{n}\frac{R_i\ell_{\phi_n}[A_i-f(\mathbf{X}_i)]}{\phi_n p(A_i|\mathbf{X}_i)},$$

to estimate the optima IDR. Furthermore, to prevent overfitting, we penalize the complexity of $f(\mathbf{X})$ as in the standard support vector machine. Consequently, our O-learning method solves the following optimization problem:

$$\min_f\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i\ell_{\phi_n}(A_i-f(\mathbf{X}_i))}{2\phi_n p(A_i|\mathbf{X}_i)}+\lambda_n\|f\|^2\right\}, \qquad (3)$$

where $\|f\|$ is some seminorm for f, and λ_n controls the severity of the penalty on f. For example, if we assume a linear decision rule, $f(\mathbf{X}) = \mathbf{X}^T\mathbf{w} + b$ and $\|f\|$ is the Euclidean norm of $\mathbf{w}$, kernel tricks can be used to obtain a nonlinear rule by assuming f is from a reproducing kernel Hilbert space and $\|\cdot\|$ is the corresponding norm.

3 Computational Algorithm

The objective function in (3) is nonconvex, and nonconvex optimization is known to be difficult. However, noting that ℓ_ϕ is the difference of the two convex functions defined earlier, we will adapt the difference of convex functions (DC) algorithm (An and Tao, 1997) to tackle this nonconvex optimization. We first discuss the algorithm for the linear learning rule, where $f(\mathbf{X})$ is a linear function of $\mathbf{X}$, and then extend it to a nonlinear learning rule where $f(\mathbf{X})$ is chosen from a Reproducing Kernel Hilbert Space (RKHS).

3.1 Learning a linear IDR

Consider $f(\mathbf{X}) = \mathbf{X}^T\mathbf{w} + b$. To simplify the notation, we absorb the $p(A|\mathbf{X})$ into R, as $p(A|\mathbf{X})$ is known in a randomized trial. We can formulate the objective function as follows:

$$S(\boldsymbol{\Theta})=\frac{\lambda_n}{2}\|\mathbf{w}\|_2^2+\frac{1}{n\phi_n}\sum_{i=1}^{n}R_i\min\left(\frac{|A_i-(\mathbf{X}_i^T\mathbf{w}+b)|}{\phi_n},1\right),$$

where λ_n is the tuning parameter and $\boldsymbol{\Theta} = (\mathbf{w}^T, b)^T$. From Figure 1, we express the objective function S as the difference of two convex functions, $S(\boldsymbol{\Theta}) = S_1(\boldsymbol{\Theta}) - S_2(\boldsymbol{\Theta})$, where

$$S_1(\mathbf{\Theta})=(\frac{\lambda_n}{2}\|\mathbf{w}\|_2^2+\frac{1}{n\phi_n}\sum_{i=1}^n R_i\frac{|A_i-(\mathbf{X}_i^T\mathbf{w}+b)|}{\phi_n})$$

and

$$S_2(\mathbf{\Theta})=\frac{1}{n\phi_n}\sum_{i=1}^n R_i(\frac{|A_i-(\mathbf{X}_i^T\mathbf{w}+b)|}{\phi_n}-1)_+.$$

Then, the DC algorithm is essentially an iterative sequence of convex minimization problems for solving the original nonconvex minimization problem. Specifically, we initialize $\mathbf{\Theta}^0$, then repeatedly update $\mathbf{\Theta}$ via

$$\mathbf{\Theta}^{t+1}=\text{argmin}_{\mathbf{\Theta}}(S_1(\mathbf{\Theta})-[\nabla S_2(\mathbf{\Theta}^t)]^T(\mathbf{\Theta}-\mathbf{\Theta}^t))$$

until convergence of $\mathbf{\Theta}$, where $\nabla S_2(\mathbf{\Theta}^t)$ is the gradient function of $S_2(\mathbf{\Theta})$ evaluated at $\mathbf{\Theta}^t$. For the initial value of $\mathbf{\Theta}$, we use a least squares estimator to predict A with $\mathbf{X}$ as predictors using the observations with a large observed reward, e.g. the observations with R_i in the upper 50th percentile of the training data.

The minimization step to obtain $\mathbf{\Theta}^{t+1}$ is a convex minimization, which we will detail in the following. In this step, if we define $Q_i^{(t)}=I(|a_i-\mathbf{X}_i^T\mathbf{w}^t-b^t| \le \phi_n)$, where $\mathbf{X}_i^T\mathbf{w}^t+b^t$ is the temporary predicted optimal dose with $\mathbf{w}^t$ and b^t being the solution from the t-th iteration, then after some algebra, the objective function for this step, denoted by $S^{(t+1)}(\mathbf{\Theta})$, equals

$$\frac{\lambda_n}{2}\|\mathbf{w}\|_2^2+\frac{1}{n\phi_n^2}\sum_{i=1}^n R_iQ_i^{(t)}|a_i-\mathbf{X}_i^T\mathbf{w}-b|.$$

Consequently, the convex subproblem is a weighted penalized median regression problem. Note that the t-th iteration result only impacts S^{t+1} through Q_i^t. Thus, if the observed patient receives a dose that is close to the surrogate optimal dose ($Q_i^{(t)}=1$), then that observation will contribute to the objective function of the $t+1$ step subproblem, otherwise it will not contribute. Let $\mathcal{T}=\{i{:}Q_i^{(t)}=1\}$. Divide the objective function by λ_n and introduce slack variables into $S^{(t+1)}(\mathbf{\Theta})$; then the primary optimization problem at the t+1-th iteration becomes

$$\min_{\mathbf{w},b,\xi,\tilde{\xi}}\frac{1}{2}\|\mathbf{w}\|_2^2+\frac{1}{n\lambda_n\phi_n^2}\sum_{i\in\mathcal{T}}(\xi_i+\tilde{\xi}_i)R_i, \qquad (4)$$

subject to ξ_i, $\tilde{\xi}_i$ ≥0, $a_i - \mathbf{X}_i^T \mathbf{w} - b \le \xi_i,\ -(a_i - \mathbf{X}_i^T \mathbf{w} - b) \le \tilde{\xi}_i$, $\forall i \in \mathscr{T}$. Using Lagrangian multipliers and employing some algebra, we end up minimizing

$$\begin{array}{c}\frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n\lambda_n\phi_n^2}\sum_{i\in\mathscr{T}}(\xi_i+\tilde{\xi}_i)R_i - \sum_{i\in\mathscr{T}}\alpha_i(\xi_i - a_i + \mathbf{X}_i^T\mathbf{w} + b) \\ -\sum_{i\in\mathscr{T}}\tilde{\alpha}_i(\tilde{\xi}_i + \alpha_i - \mathbf{X}_i^T\mathbf{w} - b) - \sum_{i\in\mathscr{T}}u_i\xi_i - \sum_{i\in\mathscr{T}}\tilde{u}_i\tilde{\xi}_i\end{array}$$

subject to

$$\begin{array}{c}\mathbf{w} - \sum_{i\in\mathscr{T}}\alpha_i\mathbf{X}_i + \sum_{i\in\mathscr{T}}\tilde{\alpha}_i\mathbf{X}_i = 0; -\sum_{i\in\mathscr{T}}\alpha_i + \sum_{i\in\mathscr{T}}\tilde{\alpha}_i = 0; \\ \frac{R_i}{n\lambda_n\phi_n^2} - \alpha_i - u_i = 0; \frac{R_i}{n\lambda_n\phi_n^2} - \tilde{\alpha}_i - \tilde{u}_i = 0.\end{array}$$

After plugging in the equations obtained from the above constraints, we obtain the following convex dual problem:

$$\min_{\alpha,\tilde{\alpha}} \frac{1}{2}\sum_{i\in\mathscr{T}}\sum_{j\in\mathscr{T}}(\alpha_i - \tilde{\alpha}_i)\langle\mathbf{X}_i, \mathbf{X}_j\rangle(\alpha_j - \tilde{\alpha}_j) - \sum_{i\in\mathscr{T}}(\alpha_i - \tilde{\alpha}_i)a_i$$

subject to

$$\sum_{i\in\mathscr{T}}(\alpha_i - \tilde{\alpha}_i) = 0; 0 \le \alpha_i \le \frac{R_i}{n\lambda_n\phi_n^2},\ 0 \le \tilde{\alpha}_i \le \frac{R_i}{n\lambda_n\phi_n^2}, \forall i \in \mathscr{T},$$

where ⟨·, ·⟩ denotes inner product, i.e. $\mathbf{X}_i^T\mathbf{X}_j$. This dual problem is a quadratic programming (QP) problem and is solvable via many standard optimization packages. Once its solution is obtained, the coefficients **w** can be recovered by the relation $\mathbf{w} = \Sigma_{i\in\mathscr{T}}(a_i - \tilde{a_i})\mathbf{X}_i$. After the solution of **w** is derived, b can be obtained by solving either a sequence of Karush-Kuhn-Tucker conditions or linear programming (Boyd and Vandenberghe, 2004).

In the DC algorithm, the iterations stop when $\|\mathbf{w}^{t+1} - \mathbf{w}^t\|$ is smaller than a pre-specified small constant (10^{-8} in our simulations). Note that since the convex function $S_2(\Theta)$ is replaced by its affine majorization, the DC algorithm is also a special case of the minorize-maximize or majorize-minimize (MM) algorithm (Hunter and Lange, 2004). Additionally, the objective function $S(\Theta)$ is bounded below by 0, and $S^{(t)}$ is descending after each iteration, and it is thus guaranteed that the DC algorithm converges to a local minimizer in finite steps (An and Tao, 1997).

3.2 Learning a nonlinear IDR

To allow more flexible functional forms for the decision, we define the kernel $K(\cdot, \cdot)$ as a symmetric, continuous and positive semidefinite function mapping from $\mathscr{X} \times \mathscr{X}$ to $\mathbb{R}$. A Reproducing Kernel Hilbert Space (RKHS) H_K associated with K is the completion of the linear span of all functions $K(\cdot, \mathbf{X})$, $\mathbf{X} \in \mathscr{X}$. In addition, there always exists a transform of $\mathbf{X}$, $\Phi(\cdot)$ such that $\Phi(\mathbf{X}_i)^T \Phi(\mathbf{X}_j) = K(\mathbf{X}_i, \mathbf{X}_j)$. We further assume that the decision function for optimal dose $f(\mathbf{X})$ is from H_K, i.e., $f(\mathbf{X}) = \mathbf{w}^T \Phi(\mathbf{X}) + b$ (Vapnik (1995); Smola and Schölkopf (2004)). Using similar derivations to that done in the linear learning setting, we obtain the following dual problem for nonlinear learning:

$$\min_{\alpha,\tilde{\alpha}} \frac{1}{2} \sum_{i\in\mathscr{T}} \sum_{j\in\mathscr{T}} (\alpha_i - \tilde{\alpha}_i) K(\mathbf{X}_i, \mathbf{X}_j)(\alpha_j - \tilde{\alpha}_j) - \sum_{i\in\mathscr{T}} (\alpha_i - \tilde{\alpha}_i) a_i$$

subject to

$$\sum_{i\in\mathscr{T}} (\alpha_i - \tilde{\alpha}_i) = 0; 0 \le \alpha_i \le \frac{R_i}{n\lambda_n \phi_n^2}, 0 \le \tilde{\alpha}_i \le \frac{R_i}{n\lambda_n \phi_n^2}, \forall i \in \mathscr{T}.$$

After solving the above QP problem, we can recover the coefficients $\mathbf{w}$ via

$$\mathbf{w} = \sum_{i=1}^{n} I(i \in \mathscr{T})(\alpha_i - \tilde{\alpha}_i).$$

Then we obtain the intercept b using the same methods as in Section 3.1. According to the representation theorem of Kimeldorf and Wahba (1971), we obtain that at each iteration $f(\mathbf{X}) = \sum_{i=1}^{n} I(i \in \mathscr{T})(\alpha_i - \tilde{\alpha}_i) K(\mathbf{X}, \mathbf{X}_i) + b$. In our paper, we implement nonlinear learning via the Gaussian kernel, i.e. $K(\mathbf{X}_i, \mathbf{X}_j) = \exp(-\gamma^{-2} \|\mathbf{X}_i - \mathbf{X}_j\|_2^2)$, where $\gamma > 0$ is the parameter for $K(., .)$. We denote the RKHS induced by the Gaussian kernel as H_γ.

3.3 Tuning parameter selection

We choose the tuning parameters by cross validation. Ideally, we wish to find an unbiased estimate for the true value function so that one can use it to evaluate the prediction performance associated with each tuning parameter using tuning data. However, since the true value function, $\mathscr{V}(f)$, cannot be well estimated in the current setting with continuous dose levels, we suggest the use of an approximate estimate $\mathscr{V}_\varepsilon(f)$ where ε is a very small parameter as the criterion function (we set $\varepsilon = 0.01$ in our numerical studies). Therefore, we propose the following procedure for tuning ϕ_n: We divide the data into training and tuning sets and consider a sequence of candidate values for ϕ_n, say $\Phi = \{0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05\}$. For each ϕ_n, we estimate the optimal IDR, denoted by $\hat{f}_{\phi_n}$, using the training data where λ_n is selected using another cross-validation with the training data only. We then evaluate the predicted value function as the empirical estimate of $\mathscr{V}_\varepsilon(\hat{f}_{\phi_n})$ using the tuning data. Then, the optimal choice of ϕ_n is the one maximizing the average of the predicted values across all the tuning sets.

4 Theoretical Results

In this section, we study the asymptotic behavior of the minimizer ($\hat{f}_n$), which solves the optimization problem (3). In particular, we will show that $\mathscr{V}(\hat{f}_n)$ converges to $\mathscr{V}(f_{opt})$ with a certain rate. This result implies that the estimated rule is an optimal rule asymptotically. As an interesting supplement to this argument, we prove in the Appendix that it is not appropriate to use convex losses, such as least squares loss or absolute deviations loss, to replace the nonconvex loss ℓ_ϕ under the O-learning framework. Our first result shows that our approximation of the value function is valid.

Theorem 1—For any measurable function $f: \mathscr{X} \to \mathbb{R}$, if $\sup_{(a,\mathbf{x})\in\mathscr{A}\times\mathscr{X}} |\partial E(R|A = a,\mathbf{X} = \mathbf{x})/\partial a| \le C$ for a bounded constant C, then $|\mathscr{V}_\phi(f) - \mathscr{V}(f)| \le C\phi_n$.

<u>Proof:</u> First, we have

$$\begin{aligned}\mathscr{V}_\phi(f)=&E\left\{\frac{R}{\phi_n p(A|\mathbf{X})}\max\left(1-\frac{|A-f(\mathbf{x})|}{\phi_n},0\right)\right\}\\ =&E_{\mathbf{x}}\left\{\frac{1}{\phi_n^2}\int_{|f(x)-a|\le\phi_n}[\phi_n-|a-f(x)|]E(R|A=a,\mathbf{X}=\mathbf{x})da\right\}\\ =&E_{\mathbf{x}}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]E(R|A=z+f(\mathbf{x}),\mathbf{X}=\mathbf{x})dz\right\}.\end{aligned}$$

After Taylor expansion, we obtain

$$\begin{aligned}\mathscr{V}_\phi(f)=&E_{\mathbf{x}}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\left[E(R|A=a,\mathbf{X}=\mathbf{x})+z\frac{\partial E(R|A=a,\mathbf{X}=\mathbf{x})}{\partial a}\Big|_{a=f(\mathbf{x})}+o(z)\right]dz\right\}.\\ =&V(f)+E_{\mathbf{x}}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\left[z\frac{\partial E(R|A=a,\mathbf{X}=\mathbf{x})}{\partial a}\Big|_{a=f(\mathbf{x})}+o(z)\right]dz\right\}.\end{aligned}$$

Therefore,

$$|\mathscr{V}_\phi(f)-\mathscr{V}(f)|=\left|E_{\mathbf{x}}\left\{\frac{1}{\phi_n^2}\int_{|z|\le\phi_n}[\phi_n-|z|]\left[z\frac{\partial E(R|A=a,\mathbf{X}=\mathbf{x})}{\partial a}\Big|_{a=f(\mathbf{x})}+o(z)\right]dz\right\}\right|.$$

By the condition in Theorem 1, we conclude that

$$|\mathscr{V}_\phi(f)-\mathscr{V}(f)|\le C\int_{|z|\le\phi_n}\frac{\phi_n|z|-z^2}{\phi_n^2}dz=C\phi_n.$$

The theorem thus holds.

Assume that the kernel function in the reproducing kernel Hilbert space in (3) is a Gaussian kernel. Then our main result establishes the convergence rate of $\mathscr{V}(\hat{f}_n) - \mathscr{V}(f_{opt})$, as given in the following theorem.

Theorem 2—Assume that the optimal rule $f_{opt} \in B^{\alpha}_{1,\infty}(\mathbb{R}^d)$, a Besov space, i.e. $B^{\alpha}_{1,\infty}(\mathbb{R}^d)=\{f \in L_{\infty}((\mathbb{R}^d)) : sup_{t>0}(t^{-\alpha}\omega_{r,L_1((\mathbb{R}^d))}(f,t))<\infty\}$, where ω is the modulus of continuity. Then, for any $\varepsilon > 0$, $d/(d+\tau) < p < 1$, $\tau > 0$, and parameter γ_n for the Gaussian kernel,

$$\mathscr{V}(f_{opt})-\mathscr{V}(\hat{f}_n) \leq c_1\left[\frac{1}{\gamma_n^{(1-p)(1+\varepsilon)d}\lambda_n^p\phi_n^2 n}\right]^{\frac{1}{2-p}}+c_2\frac{\tau^{1/2}}{\phi_n n^{1/2}}+c_3\frac{\tau}{\phi_n n}+c_4\frac{\lambda_n}{\gamma_n^d}+c_5\frac{\gamma_n^{\alpha}}{\phi_n^2}+c_6\phi_n$$

with probability P_n not less than $1 - 3e^{-\tau}$. The constants c_1 to c_6 are independent of n.

With properly chosen parameters γ_n, λ_n, ϕ_n, the right hand side of the inequality will go to 0 as n goes to infinity. The last term is due to the use of $\mathscr{V}_{\phi}$ to approximate $\mathscr{V}$. The other terms are from the approximation error due to H_{γ} and the stochastic error due to the finite sample size. We can choose λ_n, γ_n, and ϕ_n to balance the approximation accuracy and the stochastic variability as follows:

$$\lambda_n=\left(\frac{1}{n}\right)^{1/4}, \quad \gamma_n=\left(\frac{1}{n}\right)^{\frac{3/\alpha}{4+3d/\alpha}}, \quad \phi_n=\left(\frac{1}{n}\right)^{\frac{1}{4+3d/\alpha}}.$$

Then, the optimal rate for the value function approximation using the estimated rule is

$$\mathscr{V}(f_{opt})-\mathscr{V}(\hat{f}_n)=O_p\left(\left(\frac{1}{n}\right)^{\frac{1}{4+3d/\alpha}}\right).$$

Theorem 2 implies that the value of the estimated rule $\hat{f}_n$ from O-learning converges to the optimal value function. Clearly, the convergence rate decreases as the dimension of prognostic variables increases. Moreover, if f_{opt} is smooth enough, i.e. α goes to infinity, the optimal convergence rate of O-learning with Gaussian kernel is close to the rate $n^{-1/4}$. Note that the rate cannot be close to n^{-1}, the rate proved by Zhao et al. (2012) for a binary treatment rule. The main reason is due to the continuous nature of the dose, i.e. the data used for learning are only from subjects who have received similar dose levels.

5 Extension to Observational Studies

In the above development, we assumed that the training data was from a randomized dose trial. However, in practice, the training data can also come from observational data, where the distribution of receiving a particular dose given the covariates is unknown and therefore needs to be estimated. More importantly, similar to causal inference using observational data, we make the no unobserved confounders assumption. That is, conditional on all covariates $\mathbf{X}$, the observed dose level, A, is independent of all potential outcomes $R^*(a)$. Under this assumption, the proposed O-learning approach remains valid except that the density $p(a/\mathbf{X})$ needs to be estimated by the observed data.

A simple approach to estimate $p(a|\mathbf{X})$ is based on some parametric model, for example, the logarithm of the observed dose follows a normal distribution with mean as a linear function of the covariates and splines of the covariates (Imai and Van Dyk, 2004). In addition, more nonparametric methods such as boosting (Bühlmann and Hothorn, 2007; Zhu et al., 2015), Random Forests (Breiman, 2001), and SVR can be used to estimate the mean function in the above model. In our numerical studies, we will utilize the latter to estimate the mean of the normal distribution for the logarithm of A given $\mathbf{X}$.

6 Simulation Study

6.1 Simulation for a randomized trial

We have conducted extensive simulations to assess the performance of the proposed method with various training sample sizes. In these simulations, we generate d-dimensional vectors of prognostic variables, $\mathbf{X} = (X_1, \dots, X_d)^T$, independently from Uniform[−1, 1]. Treatment A is generated from Uniform[0, 2] independently of $\mathbf{X}$. The response R is normally distributed in $N(Q_0(\mathbf{X},A), 1)$, where $Q_0(\mathbf{X},A)$ reflects the interaction between the treatment and the prognostic variables and is chosen to vary according to the following scenarios:

Scenario 1:

$$Q_0(\mathbf{X}, A) = 8 + 4X_1 - 2X_2 - 2X_3 - 25 \times (f_{opt}(\mathbf{X}) - A)^2,$$
$$f_{opt}(\mathbf{X}) = 1 + 0.5X_1 + 0.5X_2.$$

Scenario 2:

$$Q_0(\mathbf{X}, A) = 8 + 4\cos(2\pi X_2) - 2X_4 - 8X_5^3 - 15 \times |f_{opt}(\mathbf{X}) - A|,$$
$$f_{opt}(\mathbf{X}) = 0.6(-0.5 < X_1 < 0.5) + 1.2(X_1 > 0.5) + 1.2(X_1 < -0.5) + X_4^2 + 0.5\log(|X_7| + 1) - 0.6.$$

In Scenario 1, we set $d = 30$, and the optimal IDR is a linear function of $\mathbf{X}$, and in Scenario 2, $d = 10$, and the optimal IDR is a nonlinear function of $\mathbf{X}$.

We apply the proposed method to estimate the optimal IDR for each simulated data set. We estimate both a linear IDR (called L-O-learning) and a non-linear IDR using the Gaussian kernel (called K-O-learning). In our O-learning methods, we fix ϕ_n to be 0.1, and the tuning parameters λ_n and γ_n are selected by 5-fold cross validation. For comparison, two competing regression-based methods (the LASSO and SVR) are also considered for estimation. Both methods are two-step procedures similar to Q-learning: in the first step, we estimate the conditional mean of R given $(A,\mathbf{X})$ using either a linear regression model with LASSO penalty (Tibshirani, 1996) or support vector regression (Vapnik, 1995; Smola and Schölkopf, 2004). Then in the second step, for each $\mathbf{X}$, we search for the optimal dose level for A which maximizes the estimated regression mean. Specifically, the covariates used in the LASSO model are $(\mathbf{X}, A, \mathbf{X}^2, \mathbf{X}A, A^2)$. In other words, we assume that given $\mathbf{X}$, the treatment dose and the reward have a quadratic relationship. Therefore, the second step for finding the optimal dose can be solved analytically. In the SVR method, a Gaussian kernel is used to estimate the non-linear relationship between R and A, as well as the interaction between A and $\mathbf{X}$ (Zhao et al., 2009). However, the closed form of the optimal dose in the second step does not exist. We use a grid search procedure to find the optimal treatment dose. In particular, we choose 400 equally spaced grids within the interval (0, 2). The tuning parameters in both LASSO and SVR are selected using 5-fold cross validation. We utilize the R (R Core Team, 2013) package "glmnet" for the LASSO (Friedman et al., 2010) and the "kernlab" package for SVR (Karatzoglou et al., 2004). We evaluate the performance of all methods by comparing the expected values under the estimated rules. The expected values are calculated from the average values of a testing set with 5000 observations. The estimated reward is calculated by plugging the estimated optimal dose into the true underlying value function.

The results from 200 replicates are summarized in Table 1. Each column is the average value function of a method evaluated from the testing set. The standard deviation of the estimated values are given in parentheses. Our methods clearly outperform the competing regression-based methods in most cases, and the difference is more dramatic for small sample size situations. When the sample size is large, L-O-learning and the LASSO work better in Scenario 1, where the true optimal IDR is linear; while K-O-learning and SVR work better in Scenario 2 with a nonlinear true optimal rule. In Figure 2, we also show the mean absolute deviation between the predicted optimal doses and the true optimal doses for all the simulations. The results in the figure show a similar pattern as that in Table 1: the proposed O-learning methods perform best in scenarios with small sample sizes.

Furthermore, regression-based methods tend to be more sensitive to covariate dimension. To demonstrate this, we conduct additional simulations for Scenario 1 with covariate dimension $d = 50$ and training sample size $n = 800$. The estimated value of SVR is 5.51, which is much smaller than the result of the SVR in Table 1 with $d = 30$. On the other hand, the average value based on our K-O-learning is 7.45, which is comparable to its value when $d = 30$ and $n = 800$.

Finally, we also compare the values from the proposed methods to the ones corresponding to fixed dose levels. Specifically, the respective optimal fixed doses are 1.00 for Scenario 1 and 0.85 for Scenario 2, with the respective values equal to 3.8 and 2.7. From Table 1, when the sample size is greater than 100, our methods already yield a larger value function than the optimal value based on either fixed dose. This result indicates that our proposed randomized trial design with our proposed O-learning method can lead to a better dose rule than the fixed dose rule identified through a traditional dose finding trial.

6.2 Simulation for an observational study

In this section, we study three scenarios where the training data are from observational studies. In particular, we want to quantify the performance change of our methods when the propensity score estimation is biased. The simulation setting is the same as in Scenario 2, except the distribution of A may depend on $\mathbf{X}$. Specifically, let TruncN(μ, a, b, σ) denote the truncated normal distribution with mean μ, lower bound a, upper bound b, and standard deviation σ. The dose assignment for Scenario 3–5 follows:

Scenario 3:

$$A \sim \begin{cases} \text{TruncN}(-0.5+0.5X_1+0.5X_2, 0, 2, 0.5), \text{when } X_3<0; \\ \text{TruncN}(|0.5+1.5X_2|, 0, 2, 1), \text{when } X_3>0. \end{cases}$$

Scenario 4:

$$A \sim \text{TruncN}(f_{opt}(\mathbf{X}), 0, 2, 0.5).$$

For both scenarios, dose levels are skewed as compared to the randomized scenarios. When implementing the proposed methods, we first estimate $p(a|\mathbf{X})$ using a log-normal model where the mean is estimated using the boosting algorithm from Zhu et al. (2015). Then, we estimate the optimal IDR using the O-learning procedure. For comparison, we also evaluate the estimated optimal IDR from the proposed method which ignores the propensity score adjustment, i.e., setting $p(a|\mathbf{X})$ to be constant, and from SVR (a regression-based approach).

From the results in Table 2, we observe that the effect of using propensity score adjustment is small for the proposed O-learning methods. In Scenario 3, the O-learning method ignoring the propensity score adjustment performs slightly worse as compared to the one using the propensity scores, yet it is slightly better in Scenario 4. In both scenarios, our methods still outperform SVR. The current simulation scenarios show that the proposed method appears to not be sensitive to misspecified propensity score models, in contrast to the usual observation when treatment options are discrete or even binary. The apparent robustness of our method for the mis-specification of the propensity score is probably due to the choice of the loss function ℓ_ϕ and the use of the DC algorithm. We have shown in Section 3 that in each subproblem of the DC algorithm, only observations whose observed dose is close to the predicted optimal dose will contribute to the loss function. It is reasonable to assume that observations with large observed rewards are more likely to receive the dose close to the optimal. Hence, the observed reward will largely determine which observations contribute most to the loss function and the propensity score will only impact the scale of such contributions. Furthermore, it appears that the errors in the propensity score are largely diluted when treatment is continuous, i.e. patients with large propensity score receiving nearly optimal treatment is a rare occurrence within a simulation. For these reasons, the performance of our methods does not change much with or without using the propensity score. This is an issue we plan on investigating more in the future, along with deriving an accompanying theoretical explanation. Note that for Scenario 4, patients receive doses near the optimal dose level. In particular, the value function of the observed dose assignment rule is about 2.7. As compared to Scenario 2, this suggests our method may have advantages when the observed dose is close to the optimal dose level; however, this is not observed for the SVR approach.

7 Warfarin Dosing

Warfarin is a commonly used medicine for preventing thrombosis and thromboembolism. Proper dosing of Warfarin is vitally important, as overdosing predisposes patients to a high risk of bleeding, while underdosing diminishes the drug's preemptive protection against thrombosis. The international normalized ratio (INR) is a measure of how rapidly the blood can clot, and the INR is actively monitored to ensure the dose of Warfarin is safe and effective. For normal people, the INR is typically 1, but for patients prescribed Warfarin, the therapeutic INR range is typically between 2 to 3 (The International Warfarin Pharmacogenetics Consortium, 2009). Predicting the optimal dose for Warfarin remains an open problem in the medical community, with many methods having been proposed to refine the optimal dose rule (The International Warfarin Pharmacogenetics Consortium, 2009; Hu et al., 2012). The InternationalWarfarin Pharmacogenetics Consortium (2009) compared three methods for predicting Warfarin dose: models by clinical data, models by pharmacogenetic data, and a single fixed dose rule. The paper concluded that pharmacogenetic data yields a better performance for predicting the optimal dose. In The International Warfarin Pharmacogenetics Consortium (2009), patient samples were used for training the prediction model only if the patient's INR was therapeutically stable and between the 2 to 3 range upon habitually taking Warfarin. The authors fitted a linear model with the received dose as the response and the pharmacogenetic data result as the predictor. Such an approach is valid when the doses received by the patients in the training data are optimal (optimal dose assumption). Later studies found that the pharmacogenetic model proposed by The International Warfarin Pharmacogenetics Consortium (2009) for optimal Warfarin dose identification is suboptimal for elderly patients. Hence, it is reasonable to assume that the optimal dose assumption may be violated in some settings.

For the following analysis, we use the data set from The International Warfarin Pharmacogenetics Consortium (2009). To estimate the optimal dose, we utilize both pharmacogenetic and clinical variables, including age, height, weight, race, CYP2C9 genotype, and VKORC1 genotype, and the use of two classes of medications: Cytochrome P450 enzyme (including phenytoin, carbamazepine, and rifampin) and Amiodarone (Cordarone)). After removing observations with missing data in these covariates, there remained a total of 1732 patients with 189 patients having INRs not in the 2 to 3 range. Instead of using patients with INRs between 2 to 3 after treatment and making the optimal dose assumption, we include all of these 1732 patients in our analysis. To convert the INR to a direct measure of reward, we code $R_i = -|INR_i - 2.5|$ for the i-th individual, as INR = 2.5 is in the ideal range. Note that the study consists of observational data rather than a randomized trial, hence we need to estimate the propensity score for O-learning. Particularly, we use the method described in Section 5 to estimate the optimal IDR.

We randomly split the data into training and testing sets 100 times, independently. We consider the scenario that the training set contains 800 samples, and the testing set contains the rest. The performances of different methods are evaluated by comparing the predicted doses and observed doses across these 100 splits. We choose the interval 10 to 100 for the grids since the 5th and 95th quantiles are $14.77 mg/week$ and $70 mg/week$ in the observed data, respectively.

Usually, in practice, the true relationship between dose and reward is unknown, hence we need to estimate the value function for a given dose rule. A potential criterion is $\mathscr{V}_\phi(\cdot)$, which will involve tuning parameters. We examine whether the methods can yield a reasonable estimated optimal dose as an alternative criterion. The LASSO method predicts all optimal doses to be 10*mg/week* or 100*mg/week* due to a vanishing quadratic term for all 100 random splits. Instead, the SVR method predicts that about 70 percent of the patients have optimal doses at extreme values: 20 percent with 10*mg/week* and 50 percent with 100*mg/week*. On the other hand, the dose prediction from O-learning is more reasonable, as shown in the density plots from Figure 3 (from one random split for illustration). In the dataset, the majority of the observed doses yield an INR within 2 to 3, hence the observed values should not be far away from optimal. For this reason, we expect the correlation of predicted optimal dose and observed dose to be relatively high if the prediction model performs well. The correlation between predicted dose and observed dose (*Corr*) for L-O-learning is about 0.60 (sd = 0.08), and the *Corr* for K-O-learning is 0.32 (sd = 0.06), *Corr* for SVR is −0.06 (sd = 0.02). In general, L-O-learning performs the best. In addition, L-O-learning suggests decreasing the dose if patients are taking Amiodarone and increasing the dose if patients are taking Cytochrome P450 enzyme (including phenytoin, carbamazepine, and rifampin). This dose suggestion is consistent with what is reported in the literature (Holbrook et al., 2005; The International Warfarin Pharmacogenetics Consortium, 2009). As a remark, our results only reveal some potential predictive biomarkers which the optimal dose levels should depend on. Future trials based on the proposed randomized dose designs are warranted to confirm these findings.

8 Discussion

The proposed O-learning method appears to be more effective than alternative approaches in both simulation studies and in the Warfarin example, especially when the training sample size is relatively small. Our method has advantages over regression-based methods through the direct estimation of the optimal dose. As a result, our method is more robust to model specification of the reward. In contrast to O-learning, one needs to correctly specify the model between reward (outcome) and treatment together with the covariates to successfully identify the optimal dose using the indirect methods. Note that the loss function proposed in this paper can be further generalized to $|A - f(\mathbf{X})|_{\phi 1} - |A - f(\mathbf{X})|_{\phi 2}$ with $0 \le \phi_1 < \phi_2$. In particular, the current loss function is a special case of this general loss with $\phi_1 = 0$. Such a generalization provides further robustness to our method at the cost of adding additional tuning parameters in the implementation. Future investigation will also include a comparison between our approaches and other methods based on regrets or contrasts (Murphy, 2003; Robins, 2004).

From the formula in Section 2.1, it is clear that the theoretical solution of $f_{opt}(\mathbf{X})$ is invariant if we replace the R by $R + g(\mathbf{X})$, where $g(\mathbf{X})$ can be any known function of $\mathbf{X}$. However, analogous to the phenomenon illustrated in Zhou et al. (2015) for personalized binary treatment, the choice of $g(\mathbf{X})$ will impact the performance of O-learning especially when the sample size is small. Choosing a function that can minimize the variance for the dose finding problem would be of great interest. Zhou et al. (2015) recommended a choice of $g(\mathbf{X})$ such that $R + g(\mathbf{X})$ can be interpreted as the residual of a regression model. The authors argued that the residual better reflected the net treatment benefit than the original outcome. In Section 6, we have shown that O-learning and regression-based approaches have different strengths and weaknesses, hence a potential choice of $g(\mathbf{X})$ could come from the regression-based approach.

In practice, the reward could be censored. Techniques such as the inverse probability of censoring weighting could be used to weight the observations; however, such a procedure can yield a less efficient rule. A similar problem can occur when the training data comes from an observational study, as in the Warfarin example. For both situations, it would be valuable to develop doubly robust estimators for IDR. Additional considerations can include accounting for drug toxicity (Thall and Russell, 1998; Laber et al., 2014) and variable selection when estimating the optimal treatment dose.

For some complex diseases, sequential treatments are needed, hence dynamic treatment regimes in place of single stage treatment rules are more useful. For Warfarin dosing, patients need to take the medicine for a set amount of time to become therapeutic, possibly causing the optimal dose to vary over time. Recently, Rich et al. (2014) proposed an adaptive strategy under the framework of structured nested mean models. In addition, other methods for estimating dynamic treatment regimes under a reinforcement learning framework (Sutton and Barto, 1998) have also been proposed in several papers, including Murphy (2003), Zhao et al. (2009) and Moodie et al. (2012), to solve multiple stage optimal treatment problems. Extensions of our proposed method for dynamic treatment regimes would be of great interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

All authors were partially funded by NCI Grant P01 CA142538. Guanhua Chen was also supported in part by an appointment to the Research Participation Program at the Center for Drug Evaluation and Research administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the Department of Energy and the US Food and Drug Administration. Donglin Zeng was also partially supported by U01-NS082062. The authors thank the co-editors, associated editor and two referees for their constructive comments which led to a significantly improved paper. The authors also thank Dr. Holloway from NCSU for helping implementing the weighted SVR.

References

An TH, Tao PD. Solving a Class of Linearly Constrained Indefinite Quadratic Problems by D. C. Algorithms. Journal of Global Optimization. 1997; 11:253–285.

Boyd, S., Vandenberghe, L. Convex Optimization. Vol. 25. Cambridge University Press; 2004.

Breiman L. Random forests. Machine Learning. 2001; 45:5–32.

Breiman, L., Friedman, J., Stone, CJ., Olshen, RA. Classi_cation and Regression Trees. CRC press; 1984.

Bühlmann P, Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. Statistical Science. 2007:477–505.

Cai T, Tian L, Uno H, Solomon SD. Calibrating Parametric Subject-specific Risk Estimation. Biometrika. 2010; 97:389–404. [PubMed: 23049123]

Chakraborty, B., Moodie, EEM. Statistical Methods for Dynamic Treatment Regimes, Statistics for Biology and Health. New York, NY: Springer New York; 2013.

Chevret, S. Statistical Methods for Dose Finding Experiments. John Wiley-Sons; New York: 2006.

Eagle KA, Lim MJ, Dabbous OH, Pieper KS, Goldberg RJ, de Werf FV, Goodman SG, Granger CB, Steg PG, Joel M, Gore M, Budaj A, Avezum A, Flather MD, Fox KAA, Investigators G. A Validated Prediction Model for All Forms of Acute Coronary Syndrome: Estimating the Risk of 6-

Month Postdischarge Death in An International Registry. The Journal of the American Medical Association. 2004; 291:2727–33. [PubMed: 15187054]

Eberts M, Steinwart I. Optimal Regression Rates for SVMs using Gaussian Kernels. Electronic Journal of Statistics. 2013; 7:1–42.

Friedman J, Hastie H, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010; 33:1–22. [PubMed: 20808728]

Henderson R, Ansell P, Alshibani D. Regret-Regression for Optimal Dynamic Treatment Regimes. Biometrics. 2010; 66:1192–1201. [PubMed: 20002404]

Holbrook AM, Pereira JA, Labiris R, McDonald H, Douketis JD, Crowther M, Wells PS. Systematic Overview of Warfarin and Its Drug and Food Interactions. Archives of Internal Medicine. 2005; 165:1095–1106. [PubMed: 15911722]

Hu Y-H, Wu F, Lo C-L, Tai C-T. Predicting Warfarin Dosage from Clinical Data: A Supervised Learning Approach. Arti_cial Intelligence in Medicine. 2012; 56:27–34.

Hunter D, Lange K. A Tutorial on MM Algorithms. American Statistician. 2004; 58:30–37.

Imai K, Van Dyk DA. Causal Inference with General Treatment Treatment Regimes: Generalizing the Propensity Score. Journal of the American Statistical Association. 2004; 99:854–866.

Karatzoglou A, Smola A, Hornik K, Zeileis A. kernlab – An S4 Package for Kernel Methods in R. Journal of Statistical Software. 2004; 11:1–20.

Kimeldorf G, Wahba G. Some Results on Tchebycheffian Spline Functions. Journal of Mathematical Analysis and Applications. 1971; 33:82–95.

Laber EB, Lizotte DJ, Ferguson B. Set-Valued Dynamic Treatment Regimes for Competing Outcomes. Biometrics. 2014; 70:53–61. [PubMed: 24400912]

Marlowe DB, Festinger DS, Dugosh KL, Lee PA, Benasutti KM. Adapting Judicial Supervision to the Risk Level of Drug Offenders: Discharge and 6-month Outcomes from a Prospective Matching Study. Drug and Alcohol Dependence. 2007; 88(Suppl 2):S4–S13.

Moodie EEM, Chakraborty B, Kramer MS. Q-learning for estimating optimal dynamic treatment rules from observational data. Canadian Journal of Statistics. 2012; 40:629–645. [PubMed: 23355757]

Moodie EEM, Dean N, Sun YR. Q-learning: Flexible learning about useful utilities. Statistics in Biosciences. 2014; 6:223–243.

Moodie EEM, Platt RW, Kramer MS. Estimating Response-Maximized Decision Rules With Applications to Breastfeeding. Journal of the American Statistical Association. 2009; 104:155–165.

Murphy SA. Optimal Dynamic Treatment Regimes. Journal of the Royal Statistical Society - Series B. 2003; 65:331–355.

Qian M, Murphy SA. Performance Guarantees for Individualized Treatment Rules. The Annals of Statistics. 2011; 39:1180–1210. [PubMed: 21666835]

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013.

Rich B, Moodie EEM, Stephens DA. Simulating Sequential Multiple Assignment Randomized Trials to Generate Optimal Personalized Warfarin Dosing Strategies. Clinical Trials. 2014; 11:435–444. [PubMed: 24464036]

Robins, JM. Optimal Structural Nested Models for Optimal Sequential Decisions. Proceedings of the Second Seattle Symposium in Biostatistics Analysis of Correlated Data; 2004.

Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. The Annals of Statistics. 1978; 6:34–58.

Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. Statistical Science. 2014; 29:640–661. [PubMed: 25620840]

Smola AJ, Schölkopf B. A Tutorial on Support Vector Regression. Statistics and Computing. 2004; 14:199–222.

Steinwart, I., Christmann, A. Support Vector Machines, Information Science and Statistics. Springer-Verlag; New York: 2008.

Sutton, RS., Barto, AG. Reinforcement Learning: An Introduction. Vol. 28. MIT press; 1998.

Thall PF, Russell KE. A Strategy for Dose-finding and Safety Monitoring Based on Efficacy and Adverse Outcomes in Phase I/II Clinical Trials. Biometrics. 1998; 54:251–264. [PubMed: 9544520]

The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. The New England Journal of Medicine. 2009; 360:753–764. [PubMed: 19228618]

Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society - Series B. 1996; 58:267–288.

Vapnik, VN. The Nature of Statistical Learning Theory, vol. 8 of Statistics for Engineering and Information Science. Springer; 1995.

Wallace MP, Moodie EEM. Doubly-Robust Dynamic Treatment Regimen Estimation via Weighted Least Squares. Biometrics. 2015; 71:636–644. [PubMed: 25854539]

Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating Optimal Treatment Regimes from a Classification Perspective. Stat. 2012a; 1:103–114. [PubMed: 23645940]

Zhang B, Tsiatis AA, Laber EB, Davidian M. A Robust Method for Estimating Optimal Treatment Regimes. Biometrics. 2012b; 68:1010–1018. [PubMed: 22550953]

Zhang T. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. The Annals of Statistics. 2004; 32:56–85.

Zhao Y, Kosorok MR, Zeng D. Reinforcement Learning Design for Cancer Clinical Trials. Statistics in Medicine. 2009; 28:3294–3315. [PubMed: 19750510]

Zhao Y, Zeng D, Rush J, Kosorok MR. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. Journal of the American Statistical Association. 2012; 107:1106–1118. [PubMed: 23630406]

Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual Weighted Learning for Estimating Individualized Treatment Rules. Journal of the American Statistical Association. 2015; 0 in press.

Zhu Y, Coffman DL, Ghosh D. A Boosting Algorithm for Estimating Generalized Propensity Scores with Continuous Treatments. Journal of Causal Inference. 2015; 3:25–40. [PubMed: 26877909]

APPENDIX

A Inconsistency of weighted SVR for IDR estimation

The following Lemma 1 demonstrates that the theoretical minimizer for weighted support vector regression with absolute deviations loss (a special case of ε-insensitive loss used in SVR) is not consistent, i.e. it does not yield a function that maximizes $\mathscr{V}(f)$.

Lemma 1

Let $f_{abs}(\mathbf{X})=argmin_f E\left[\frac{R|A-f(\mathbf{X})|}{p(A|\mathbf{X})}\right]$, then $f_{abs}(\mathbf{X}) \neq f_{opt}(\mathbf{X})$.

Proof—By definition, $E[\frac{R|A-f(\mathbf{X})|}{p(A|\mathbf{X})}]=\int E(R|a,\mathbf{x})|a-f(\mathbf{x})|p(\mathbf{x})dadx$. Let $\tilde{p}(a|\mathbf{x}) \propto E(R|a, \mathbf{x})p(\mathbf{x})$, then for any given $\mathbf{x}$, $f_{abs}(\mathbf{x})$ is the median of a with respect to density $\tilde{p}(a|\mathbf{x})$. On the other hand, we have proved that $f_{opt} = \text{argmax} E_X[E\{R|A = f(\mathbf{X}), \mathbf{X}\}]$. This implies that for any given x, $f_{opt}(\mathbf{x})$ is the mode of a with respect to density $\tilde{p}(a|\mathbf{x})$. If $\tilde{p}(a|\mathbf{x})$ is not symmetric, then $f_{abs}(\mathbf{x}) \neq f_{opt}(\mathbf{x})$. As a result, $f_{abs}(\mathbf{X}) \neq f_{opt}(\mathbf{X})$.

Hence, it is not proper to use the absolute deviance loss in O-learning for dose finding. Similarly, we can show that using certain other loss functions, such as quadratic loss, is not consistent either. This follows by the argument that $f_{quad}(\mathbf{X})=\text{argmin}E\left[\frac{R(A-f(\mathbf{X}))^2}{p(A|\mathbf{X})}\right]$, and for given x, $f_{quad}(\mathbf{x})$ is the mean of a with respect to density $\tilde{p}(a|\mathbf{x})$.

B Proof of Theorem 2

Let f_ϕ^* be the minimizer of $\mathcal{R}_\phi(f)$ and by definition $\mathscr{R}_\phi(f)=E\left(\frac{R}{\phi_n p(A|X)}\right)-\mathscr{V}_\phi(f)$. Then

$$\begin{aligned}\mathscr{V}(f_{opt})-\mathscr{V}(\hat{f}_n)=E(R|A=f_{opt})-E(R|A=\hat{f}_n)\\ \leq \mathscr{V}_\phi(f_{opt})-\mathscr{V}_\phi(\hat{f}_n)+2C\phi_n \leq \mathscr{R}_\phi(\hat{f}_n)-\mathscr{R}_\phi(f_{opt})+2C\phi_n\\ \leq \mathscr{R}_\phi(\hat{f}_n)-\mathscr{R}_\phi(f_\phi^*)+c_6\phi_n.\end{aligned} \quad (5)$$

The first inequality is due to Theorem 1, and the second follows by the definition of $\mathcal{R}_\phi$. C is the same constant used in Theorem 1 and $c_6 = 2C$. Denoting $\mathscr{R}_\phi^*=\mathscr{R}_\phi(f_\phi^*)$, we can see that

$$\mathscr{R}_\phi(\hat{f}_n)-\mathscr{R}_\phi^* \leq \lambda_n\|\hat{f}_n\|_k^2+\mathscr{R}_\phi(\hat{f}_n)-\mathscr{R}_\phi^*=\frac{1}{\phi_n}\left[\lambda_n\phi_n\|\hat{f}_n\|_k^2+\phi_n\mathscr{R}_\phi(\hat{f}_n)-\phi_n\mathscr{R}_\phi^*\right]. \quad (6)$$

Next, we want to bound $=\lambda_n\phi_n\|\hat{f}_n\|_k^2+\phi_n\mathscr{R}_\phi(\hat{f}_n)-\phi_n\mathscr{R}_\phi^*$. The reason for scaling by the factor ϕ_n is that the corresponding loss function after scaling becomes $L_\phi(\mathbf{X},A,f(\mathbf{X}))=\frac{R\ell_\phi(A-f(\mathbf{X}))}{p(A|\mathbf{X})}$. $L_\phi(\mathbf{X}, A, f(\mathbf{X}))$ is a bounded loss function when $\phi_n \to 0$, and such a property will facilitate the proof. In the following theorem, let $\hat{f}_n = f_{D,\lambda_n}$, $\mathcal{R}_{L,P} = \phi_n\mathcal{R}_\phi$, and $\mathscr{R}_{L,P}^*=\phi_n\mathscr{R}_\phi(f_\phi^*)$. Hence, f_ϕ^* is also the minimizer for $\mathcal{R}_{L,P}$. In the following steps, we rely on the theorem proved by Steinwart and Christmann (2008), given as follows:

Theorem 7.23 (Oracle inequality for SVMs using benign kernels, Steinwart and Christmann, 2008)

Let $L : \mathbf{X} \times A \times \mathbb{R} \to [0, \infty)$ be a loss function. Also, let H be a separable RKHS of a measurable kernel over $\mathbf{X}$ and P be a distribution on $\mathbf{X} \times A$. If the following conditions are satisfied:

(A1) *L satisfies the supremum bound $L(.) \leq B$ for a $B > 0$.*

(A2) *L is a locally Lipschitz continuous loss that can be clipped at $M > 0$.*

(A3) *The variance bound $\mathbb{E}_P(L\circ\tilde{f}-L\circ f_{L,P}^*)^2 \leq V\cdot(\mathbb{E}_P(L\circ\tilde{f}-L\circ f_{L,P}^*))^v$ is satisfied for constants $\nu \in [0, 1]$, $V \geq B^{2-\nu}$, and all $f \in H$.*

(A4) For fixed n ≥ 1, there exist constants p $\in (0, 1)$ and a $\geq$B such that the entropy number $\mathbb{E}_{D_X\sim P_X^n}\mathfrak{e}_i(id:H \to L_2(D_X)) \leq ai^{-\frac{1}{2p}}$, $i \geq 1$.

Fix an $f_0 \in H$ and a constant $B_0 \geq B$ such that $L \circ f_0 \leq B_0$. Then, for all fixed $\tau > 0$ and $\lambda_n > 0$, the SVM using H and L satisfies

$$
\begin{aligned}
&\lambda_n\phi_n\|f_{D,\lambda_n}\|_H^2 \\
&\quad +\mathscr{R}_{L,P}(\tilde{f}_{D,\lambda_n}) \\
&\quad -\mathscr{R}^*_{L,P} \le 9(\lambda_n\phi_n\|f_0\|_H^2+\mathscr{R}_{L,P}(f_0) \\
&\qquad -\mathscr{R}^*_{L,P})+K_0(\frac{a^{2p}}{(\lambda_n\phi_n)^p n})^{\frac{1}{2-p-v+vp}} \\
&\quad +3(\frac{72V\tau}{n})^{\frac{1}{2-v}}+\frac{15B_0\tau}{n}
\end{aligned}
$$

with probability P^n not less than $1-3e^{-\tau}$, where $K_0 \geq 1$ is a constant only depending on p, M, B, ν, and V.

We will verify conditions (A1)–(A4) for our setting as follows: The loss function in our problem is $L_\phi(\mathbf{X},A,f(\mathbf{X}))=\frac{R\ell_\phi(A-f(\mathbf{X}))}{p(A|\mathbf{X})}$. For our problem, it is reasonable to assume the rewards are bounded and $p(A/\mathbf{X})$ is bounded such that $R/p(A/\mathbf{X}) \in [0, B]$, where B is some finite constant. Hence, we have $L_\phi(.) \leq B$ for (A1).

Note that $\tilde{f}$ is a clipped version of f (via Winsorization) for some value M, such that $\tilde{t} = I(/t/ \leq M)t + I(/t/ > M)\text{sign}(t)M$. The risks of $L_\phi(.)$ loss satisfy $\mathscr{R}(\tilde{f}) \leq \mathscr{R}(f)$, if we set M to be some large value, i.e. larger than the range of the dose. Hence, $L_\phi(\cdot)$ can be clipped. This implies we can investigate the clipped version of the loss instead of the original loss function without loss of generality (Steinwart and Christmann, 2008). Furthermore, $L_\phi(.)$ is Lipschitz continuous with Lipschitz constant equal to B/ϕ_n, hence it is also locally Lipschitz continuous. As a result, condition (A2) is satisfied.

Furthermore, (A3) is true since

$$\mathbb{E}_P(L_\phi\circ\tilde{f}-L_\phi\circ f^*_\phi)^2 \le 2\mathbb{E}_P[(L_\phi\circ\tilde{f})^2+(L_\phi\circ f^*_\phi)^2] \le 4B^2.$$

The benign kernel we implement in the algorithm is the Gaussian kernel. By Theorem 7.34 of Steinwart and Christmann (2008), we have that (A4) holds with the constant a being set equal to $c_{\varepsilon,p}\gamma_n^{-\frac{(1-p)(1-\varepsilon)d}{2p}}$.

So far, all the conditions needed for Theorem 7.23 are satisfied, hence by plugging in $a=c_{\varepsilon,p}\gamma_n^{-\frac{(1-p)(1+\varepsilon)d}{2p}}$, $\nu = 0$, $V = 4B^2$, and $B_0 = B$, we obtain

$$\le 9A(\lambda_n)+K_0[\frac{1}{\gamma_n^{(1-p)(1+\varepsilon)d}\lambda_n^p\phi_n^p n}]^{\frac{1}{2-p}}+36\sqrt{2}B(\tau/n)^{\frac{1}{2}}+15B(\tau/n), \tag{7}$$

where $A(\lambda_n)=\lambda_n\phi_n\|f_0\|^2_{H_{\gamma_n}}+\mathscr{R}_{L,P}(f_0)-\mathscr{R}^*_{L,P}$.

The next step is to bound the approximation error $A(\lambda_n)$. Since any $f_0 \in H_\gamma$ is valid for Equation (7), we can study a specific choice of f_0 and the corresponding bound for $A(\lambda_n)$. To construct this f_0, for $r \in \mathbb{N}$ and $\gamma > 0$, we define the function $\mathscr{K} : \mathbb{R}^d \to \mathbb{R}$ as $\mathscr{K}(x) = \sum_{j=1}^{r} \begin{pmatrix} r \\ j \end{pmatrix} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\gamma^2 \phi}\right)^{d/2} \mathscr{K}_{j\gamma/\sqrt{2}}(x)$, where $\mathscr{K}_\gamma = \exp(-\gamma^2 \|x\|_2^2)$ for all $x \in R^d$. Assuming that $f_\phi^* \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, we can define f_0 by convolving $\mathscr{K}$ with this optimal decision function (Eberts and Steinwart, 2013), that is

$$f_0 := \mathscr{K} * f_\phi^* = \int_{\mathbb{R}^d} \mathscr{K}(X - t) f_\phi^*(t) dt,\ x \in \mathbb{R}^d. \quad (8)$$

With the help of two theorems in Eberts and Steinwart (2013), we can show that f_0 is contained in H_γ, and that it is a suitable function for bounding $A(\lambda_n)$.

By the construction of f_0, the approximation error for our problem is written as:

$$A(\lambda_n) = \lambda_n \phi_n \|f_0\|_{H_{\gamma n}}^2 + \mathscr{R}_{L,P}(f_0) - \mathscr{R}_{L,P}^* = \lambda_n \phi_n \|\mathscr{K} * f_\phi^*\|_{H_{\gamma n}}^2 + \mathscr{R}_{L,P}(\mathscr{K} * f_\phi^*) - \mathscr{R}_{L,P}^*.$$

By Theorem 2.3 of Eberts and Steinwart (2013), we have:

$$A(\lambda_n) \le \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|f_\phi^*\|_{L_2(\mathbb{R}^d)}^2 + \mathscr{R}_{L,P}(\mathscr{K} * f_\phi^*) - \mathscr{R}_{L,P}^*.$$

By the Lipschitz continuity property of the loss function L_ϕ,

$$A(\lambda_n) \le \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|f_\phi^*\|_{L_2(\mathbb{R}^d)}^2 + \frac{B}{\phi_n} |\mathscr{K} * f_\phi^* - f_\phi^*|_{L_1(P_X)}.$$

By Theorem 2.2 of Eberts and Steinwart (2013),

$$A(\lambda_n) \le \lambda_n \phi_n (\gamma_n \sqrt{\pi})^{-d} (2^r - 1)^2 \|f_\phi^*\|_{L_2(\mathbb{R}^d)}^2 + \frac{B}{\phi_n} C_{r,1} \|g\| L_p(P_X) \omega_{r, L_1(\mathbb{R}^d)}(f_\phi^*, \gamma_n/2). \quad (9)$$

Recall that we assume $f_{opt} \in B_{1,\infty}^\alpha(\mathbb{R}^d)$, a Besov space, and if we further assume $f_\phi^* \in B_{1,\infty}^\alpha(\mathbb{R}^d)$, i.e. $B_{1,\infty}^\alpha(\mathbb{R}^d) = \{f \in L_\infty((\mathbb{R}^d)) : sup_{t>0}(t^{-\alpha} \omega_{r, L_1((\mathbb{R}^d))}(f, t)) < \infty\}$, then $\omega_{r, L_1(\mathbb{R}^d)}(f_\phi^*, \gamma_n/2) < c_0 \gamma_n^\alpha$, where c_0 is a constant. Plugging this into inequality (9), and combining with the assumption $f_\phi^* \in L_2(\mathbb{R}^d) \cap L_\infty(\mathbb{R}^d)$, we obtain

$$A(\lambda_n) \le c_1 \lambda_n \phi_n \gamma_n^{-d} + c_2 \gamma_n^{\alpha} \phi_n^{-1}. \quad (10)$$

Combining Equation (6), Equation (7) and Equation (10), we obtain that:

$$\mathscr{R}_\phi(\hat{f}_n) - \mathscr{R}_\phi(f_\phi^*) \le c_1 \left[\frac{1}{\gamma_n^{(1-p)(1+\varepsilon)d} \lambda_n^p \phi_n^2 n} \right]^{\frac{1}{2-p}} + c_2 \frac{\tau^{1/2}}{\phi_n n^{1/2}} + c_3 \frac{\tau}{\phi_n n} + c_4 \frac{\lambda_n}{\gamma_n^d} + c_5 \frac{\gamma_n^{\alpha}}{\phi_n^2}. \quad (11)$$

Combining Equation (5) and Equation (11) now yields the desired result.

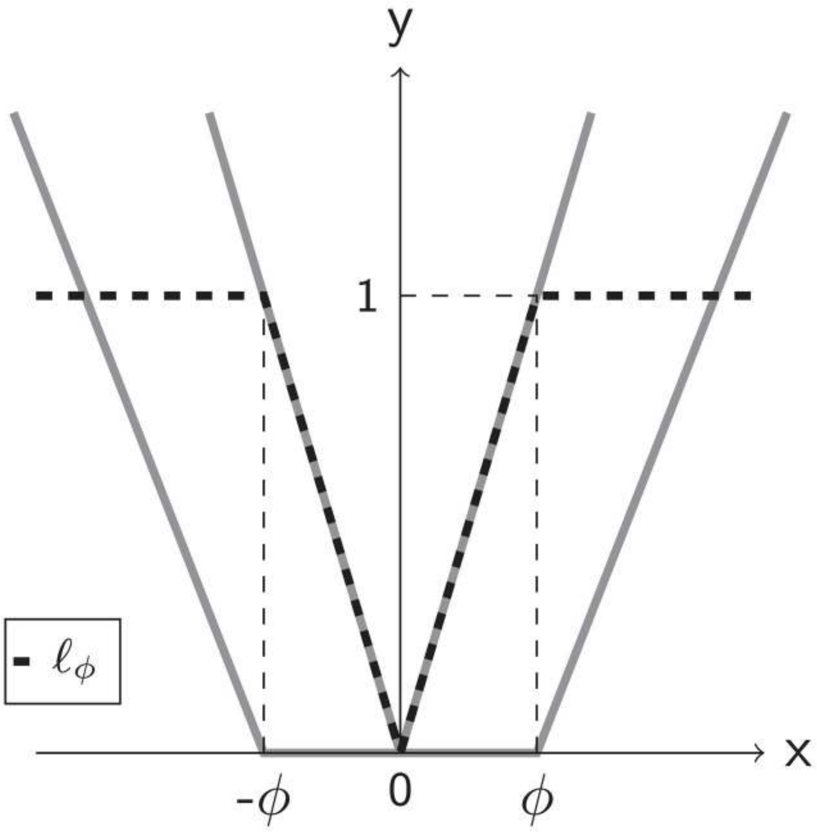

Figure 1.

Loss function ℓ_ϕ for dose finding in an IDR. ℓ_ϕ (black, dash) is the difference between the V shaped function (gray, solid) and U shaped function (gray, solid).

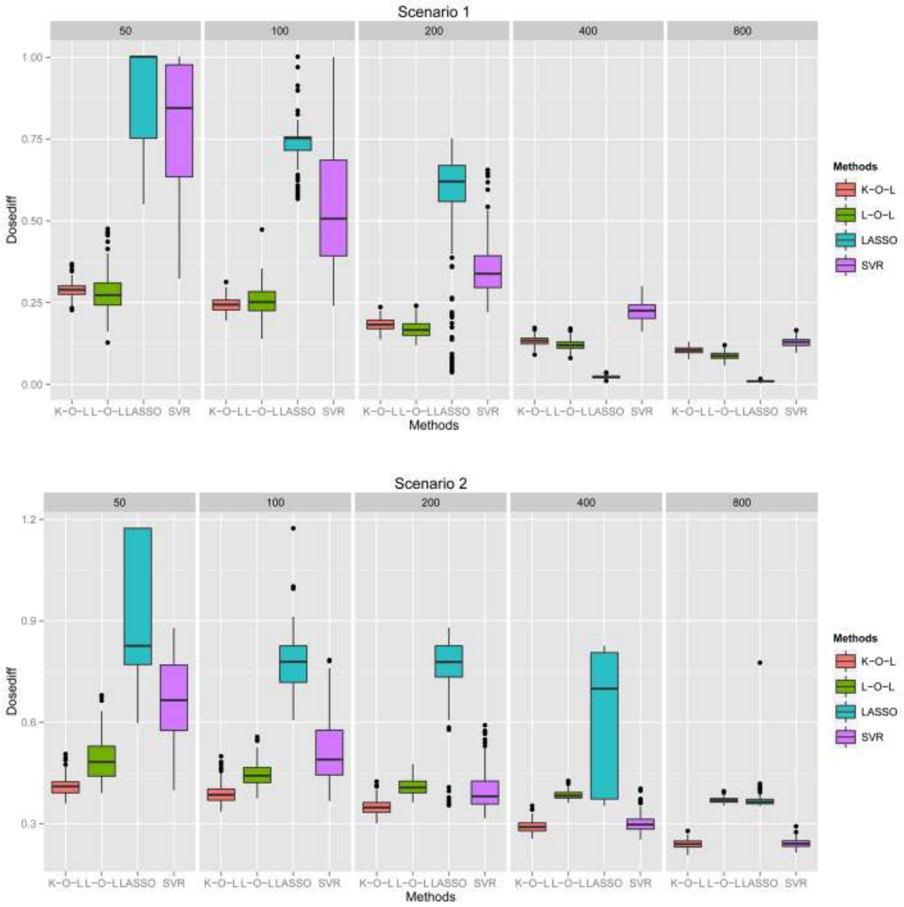

Figure 2.

Boxplots of the absolute difference of the predicted optimal doses from the truth.

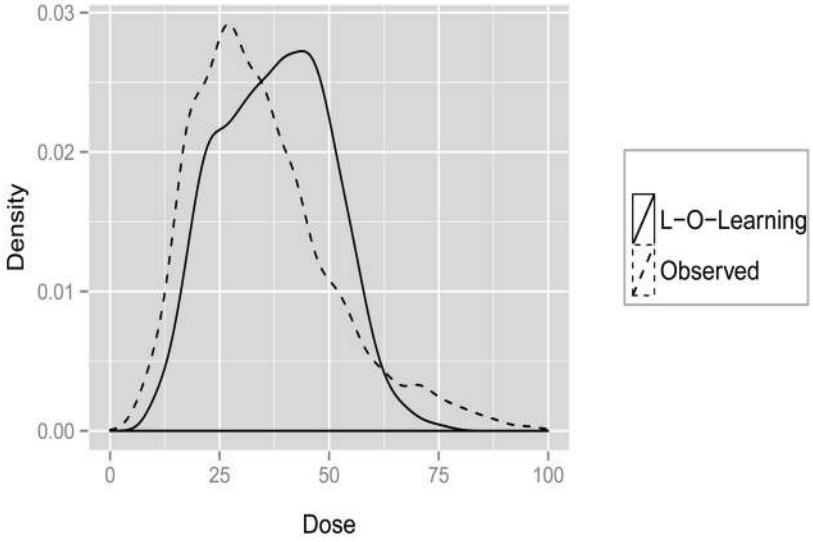

Figure 3.
Distribution of observed and predicted Doses.

Table 1

Average $\hat{\mathscr{V}}(f)$ from 200 replicates from randomized trials

	n	K-O-learning	L-O-learning	SVR	LASSO
Scenario 1	50	4.78 (0.48)	**4.83 (1.40)**	−12.21 (7.53)	−15.62 (6.61)
	100	**5.69 (0.40)**	5.39 (0.93)	−2.57 (6.34)	−9.76 (5.21)
	200	6.68 (0.26)	**6.85 (0.34)**	3.46 (1.97)	−1.91 (4.65)
	400	7.28 (0.15)	7.41 (0.14)	6.13 (0.47)	**7.95 (0.01)**
	800	7.54 (0.08)	7.67 (0.08)	7.36 (0.12)	**7.97 (0.01)**
Scenario 2	50	**2.00 (0.29)**	1.16 (0.71)	−1.96 (1.70)	−5.58 (2.79)
	100	**2.19 (0.43)**	1.57 (0.52)	0.24 (1.42)	−4.12 (2.17)
	200	**2.84 (0.37)**	2.02 (0.30)	2.01 (0.84)	−3.37 (1.38)
	400	**3.69 (0.27)**	2.30 (0.18)	3.47 (0.37)	−0.92 (3.12)
	800	**4.41 (0.19)**	2.49 (0.10)	4.35 (0.19)	1.92 (0.69)

Note: the numbers in boldface are the largest for each row.

Table 2

Average $\hat{\mathscr{V}}(f)$ from 200 replicates for observational studies

	n	K-O-learning	K-O-learning-ps	SVR
Scenario 3	200	2.68 (0.30)	**2.74 (0.29)**	1.99 (0.83)
	800	4.06 (0.30)	**4.19 (0.20)**	4.09 (0.28)
Scenario 4	200	**3.29 (0.28)**	3.23 (0.28)	−0.95 (1.57)
	800	**4.91 (0.14)**	4.73 (0.17)	3.04 (0.52)

Note: "K-O-learning" is the proposed method for a nonlinear IDR, but treating $p(A|\mathbf{X})$ as constant; "K-O-learning-ps" is the proposed method for a linear IDR using the estimated $p(A|\mathbf{X})$. The numbers in boldface are the largest in each row.