BIOMETRICS 000, 000–000 DOI: 000
000 0000

Adaptive Contrast Weighted Learning for Multi-Stage Multi-Treatment Decision-Making

Yebin Tao and Lu Wang

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

**email:* yebintao@umich.edu; luwang@umich.edu

SUMMARY: Dynamic treatment regimes (DTRs) are sequential decision rules that focus simultaneously on treatment individualization and adaptation over time. To directly identify the optimal DTR in a multi-stage multi-treatment setting, we propose a dynamic statistical learning method, adaptive contrast weighted learning. We develop semiparametric regression-based contrasts with the adaptation of treatment effect ordering for each patient at each stage, and the adaptive contrasts simplify the problem of optimization with multiple treatment comparisons to a weighted classification problem that can be solved by existing machine learning techniques. The algorithm is implemented recursively using backward induction. By combining doubly robust semiparametric regression estimators with machine learning algorithms, the proposed method is robust and efficient for the identification of the optimal DTR, as shown in the simulation studies. We illustrate our method using observational data on esophageal cancer.

KEY WORDS: Dynamic treatment regime; Personalized medicine; Classification; Backward induction; Causal inference.

Author Manuscript

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Individualized treatment strategies (ITS) are decision rules that dictate treatment prescriptions based on a patient's specific characteristics (e.g., demographics, clinical outcomes and genetic makeup). Given the increasingly popular theme of personalized medicine, many clinical and intervention scientists have now become interested in the development of ITS. Treatment individualization is important due to the fact that many diseases, such as cancer and diabetes, have complex causes by the interplay among genetic, physiological and environmental factors that vary from person to person. The effectiveness of a given treatment is usually determined not only by a patient's current disease status but also by his/her past treatment and disease history and perhaps other concurrent medical conditions. Moreover, due to the progressive nature of many chronic diseases, treatment adaptation over time is also crucial to optimize treatment effects.

Author Manuscript

Dynamic treatment regimes (DTRs) (Robins, 1986, 1997, 2004; Murphy, 2003; Chakraborty et al., 2013) mathematically generalize personalized medicine to a time-varying treatment setting. They are sequential decision rules that focus simultaneously on treatment individualization and adaptation over time. Identifying the optimal DTRs offers an effective vehicle for personalized management of diseases, and helps physicians tailor the treatment strategies dynamically and individually based on clinical evidence, which provides a key foundation for better chronic care (Wagner et al., 2001). However, it is challenging to identify optimal DTRs in a multi-stage treatment setting due to the complex relationships between the alternating sequences of time-varying treatments and clinical outcomes. Recent research on estimating optimal DTRs has focused on sequential multiple assignment randomized trials (SMARTs) (Murphy, 2005), which are desirable for causal inference, as well as longitudinal observational studies (Murphy, 2003; Robins, 2004), which are the more common source of data. The observational data may restrict the set of DTRs that can be assessed due to possible violation of key causal assumptions and thus require careful thoughts and formulations in order to make valid inference (Robins and Hernán, 2009). Diverse statistical methods have been developed including marginal structural models with inverse probability weighting (IPW) (Robins, 2000; Hernán et al., 2001; Wang et al., 2012), G-estimation of structural nested mean models (Robins, 1986, 1989, 1997), generalized by Murphy (2003) and Robins (2004), targeted maximum likelihood estimators (van der Laan and Rubin, 2006), and likelihood-based approaches (Thall et al., 2007). However, susceptibility to model misspecification remains as a major limitation of many methods in this field due to the inherent difficulty of modeling high-dimensional information in a time-varying setting.

Machine learning methods have become popular alternative approaches on estimating optimal DTRs. The commonly employed methods include Q-learning (Watkins and Dayan, 1992; Sutton and Barto, 1998) and A-learning (Murphy, 2003; Schulte et al., 2014), both of which use backward induction (Bather, 2000) to first optimize the treatment at the last stage and then sequentially optimize the treatment at each of the earlier stages. Q- and A- learning are both indirect approaches as they rely on maximizing or minimizing an objective function to infer the optimal DTRs and thus emphasize prediction accuracy of the clinical response model instead of directly optimizing the decision rule (Zhao et al., 2012). Zhang et al. (2012) propose a framework to transform the problem of estimating the optimal treatment regime into a weighted classification problem, and then directly estimate the optimal regime. Their proposed method is robust and efficient due to a combination of semiparametric regression estimators and nonparametric classification methods. However, their approach is limited to a single decision point with binary treatment options. For multi-stage decisions, Zhao et al. (2015) propose outcome weighted learning (OWL) to convert the optimal DTR problem into an either sequential or simultaneous classification problem. OWL utilizes existing machine learning techniques, such as support vector machines (SVM) (Cortes and Vapnik, 1995), to directly estimate the optimal DTR, which is flexible without the specification of outcome regression models. However, it is also not as efficient as model-based approaches if the models can be well approximated. As reviewed by Zhou et al. (2015), OWL is susceptible to

Author Manuscript

trying to retain the actually observed treatments and is also unstable in general since its estimated individualized treatment rule is affected by a simple shift of the outcome. Moreover, OWL is susceptible to the misspecification of propensity score models since it is based on IPW. To our knowledge, few research attempts exist that deal with more than two discrete treatment options at each stage and estimate the optimal DTR in a robust and efficient way.

In this article, we develop a dynamic statistical learning method, adaptive contrast weighted learning (ACWL), to directly estimate the optimal DTR through a sequence of weighted classification for multi-stage multi-treatment decision-making in observational studies. The algorithm is implemented recursively using backward induction. Our method has multiple strengths and novelties compared to existing methods. First of all, it can handle more than two treatments at each stage. Extending from two treatment options to more than two is nontrivial since one must account for multiple treatment comparisons without sacrificing too much on efficiency, especially when the number of treatment options is large. We achieve this by using contrasts with the adaptation of treatment effect ordering for each patient at each stage. The proposed adaptive contrasts stand for the minimum or maximum expected loss in the outcome given any sub-optimal treatment for each patient, and simplify the problem of optimization with multiple treatment comparisons to a weighted classification problem at each stage. Second, ACWL is robust and efficient by combining semiparametric regression estimators with machine learning methods. Following Zhang et al. (2012), we employ the doubly robust augmented inverse probability weighted (AIPW) estimator (Robins et al., 1994; Scharfstein et al., 1999) to estimate the treatment effect ordering and adaptive contrasts at each stage. Last but not least, ACWL can be easily implemented using existing regression and classification methods, and is also flexible given the capability of incorporating various modeling and machine learning techniques.

The remainder of this paper is organized as follows. In Section 2, we formalize the problem of estimating the optimal DTR in a multi-stage multi-treatment setting using the counterfactual

Author Manuscript

framework and transform it to a sequence of weighted classification using adaptive contrasts. The performance of our proposed method in various scenarios is evaluated by simulation studies in Section 3. We further illustrate our method in Section 4 using esophageal cancer data. Finally, we conclude with some discussions and suggestions for future research in Section 5.

2. Adaptive Contrast Weighted Learning (ACWL)

2.1 *Notation*

Consider a clinical trial or observational study with n subjects from a population of interest and T treatment stages. For brevity, we suppress the patient index i ($i = 1, \dots, n$) in the following text when no confusion exists. For $j = 1, \dots, T$, let A_j denote the multi-categorical treatment indicator at the j^{th} stage with observed value $a_j \in \mathcal{A}_j = \{1, \dots, K_j\}$ $(K_j \geq 2)$. Let $\mathbf{X}_j$ denote the vector of patient characteristics history just prior to treatment assignment A_j, containing both baseline and time-varying covariates, and $\mathbf{X}_{T+1}$ denote the entire characteristics history up to the end of stage T. Let R_j be the clinical outcome following A_j, also known as rewards, which depends on the precedent covariate history $\mathbf{X}_j$ and treatment history $A_1, \dots, A_j$, and is also a part of the covariate history $\mathbf{X}_{j+1}$. We consider the overall outcome of interest to be $Y = f(R_1, \dots, R_T)$, where $f(\cdot)$ is a prespecified function (e.g., sum), and assume that Y is bounded and preferable with larger values.

A DTR $\mathbf{g} = (g_1, \dots, g_T)$ is a set of rules for personalized treatment decisions at all T stages, where g_j is a map from the domain of covariate and treatment history $\mathbf{H}_j = (A_1, \dots, A_{j-1}, \mathbf{X}_j^\top)^\top$ to the domain of treatment assignment A_j, and we set $A_0 = \emptyset$. The optimal DTR is the one that maximizes the expectation of Y if used to assign treatments to all patients in the population of interest.

Author Manuscript

2.2 *ACWL with* $T = 1$

To facilitate the presentation of our method, we start with optimizing the treatment regime for a single stage and $K(\geq 2)$ treatment options. The method is essentially the same for optimizing the regime for the last stage in a multi-stage decision problem. We suppress the stage index in this section for brevity. To define and identify the optimal treatment regime, we consider the counterfactual framework for causal inference (Robins, 1986). Let $Y^*(a), a = 1, \ldots, K$, denote the counterfactual outcome had a subject received treatment a. We make the following three assumptions in order to estimate $E\{Y^*(a)\}$. First, we assume that the observed outcome is the same as the counterfactual outcome under the treatment a patient is actually given, i.e., $Y = \sum_{a=1}^{K} Y^*(a)I(A = a)$, where $I(\cdot)$ is the indicator function that takes the value 1 if $\cdot$ is true and 0 otherwise. This is referred to as the consistency assumption, which also implies that there is no interference between subjects. Second, we make the no unmeasured confounding assumption (NUCA); treatment A is randomly assigned with probability possibly dependent on $\mathbf{H}$, i.e., $\{Y^*(1), \ldots, Y^*(K)\} \perp\!\!\!\perp A|\mathbf{H}$, where $\perp\!\!\!\perp$ denotes statistical independence. Third, we assume that with probability one, the propensity score $\pi_a(\mathbf{H}) = Pr(A = a|\mathbf{H})$ is bounded away from zero, which is known as the positivity assumption.

We define the counterfactual outcome for a patient following regime g as

$$Y^*(g) = \sum_{a=1}^{K} Y^*(a)I\{g(\mathbf{H}) = a\},$$

and thus conditioning on $\mathbf{H}$, we have

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\sum_{a=1}^{K} E\left\{Y^*(a)|\mathbf{H}\right\} I\{g(\mathbf{H}) = a\}\right],$$

where $E_{\mathbf{H}}(\cdot)$ denotes expectation with respect to the marginal joint distribution of $\mathbf{H}$. Under NUCA, we can further show that

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\sum_{a=1}^{K} E\{Y^*(a)|A = a, \mathbf{H}\}I\{g(\mathbf{H}) = a\}\right],$$

and given the consistency assumption, we have

$$E\{Y^*(g)\} = E_{\mathbf{H}}\left[\sum_{a=1}^{K} E(Y|A = a, \mathbf{H})I\{g(\mathbf{H}) = a\}\right].$$

Author Manuscript

The positivity assumption assures the identifiability of $E(Y|A = a, \mathbf{H})$.

The optimal regime, g^{opt}, is the one that maximizes the expected counterfactual outcome among the class of all potential regimes, $\mathcal{G}$. If we denote the conditional mean $E(Y|A = a, \mathbf{H})$ as $\mu_a(\mathbf{H})$, we have

$$g^{opt} = arg \max_{g\in\mathcal{G}} E_{\mathbf{H}} \left[\sum_{a=1}^{K} \mu_a(\mathbf{H}) I\{g(\mathbf{H}) = a\} \right].$$

Let $\mu_{(1)}(\mathbf{H}) \leq \ldots \leq \mu_{(K)}(\mathbf{H})$ denote the order statistics of $\mu_1(\mathbf{H}), \ldots, \mu_K(\mathbf{H})$, and l_a denote the treatment effect order with $\mu_{(a)}(\mathbf{H}) = \mu_{l_a}(\mathbf{H})$. Note that l_a depends on $\mathbf{H}$. Therefore, we get

$$g^{opt} = arg \max_{g\in\mathcal{G}} E_{\mathbf{H}} \left[\sum_{a=1}^{K} \mu_{(a)}(\mathbf{H}) I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right].$$

By subtracting $\mu_{(K)}(\mathbf{H})$ and reversing the sign, we have

$$g^{opt} = arg \min_{g\in\mathcal{G}} E_{\mathbf{H}} \left[\sum_{a=1}^{K-1} \{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\} I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right]. \tag{1}$$

According to (1), g^{opt} minimizes the expected loss in the outcome due to sub-optimal treatments in the entire population of interest. It would classify as many patients as possible to their corresponding treatment l_K (i.e., letting $I\{g(\mathbf{H}) = l_a(\mathbf{H})\} = 0, a = 1, \ldots, K-1$) while putting more emphasis on patients with larger contrasts (i.e., larger values of $\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})$) if misclassification is inevitable. Ideally, for each patient, we would utilize all $K-1$ contrasts as weights to conduct treatment classification, which, however, is challenging in practice. Meanwhile, given the inequality

$$0 \leq \mu_{(K)}(\mathbf{H}) - \mu_{(K-1)}(\mathbf{H}) \leq \mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H}) \leq \mu_{(K)}(\mathbf{H}) - \mu_{(1)}(\mathbf{H}),$$

it is easy to show

$$\begin{aligned} E_{\mathbf{H}} \left[\sum_{a=1}^{K-1} \{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\} I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right] &\geq E_{\mathbf{H}} \left[\sum_{a=1}^{K-1} \{C_1(\mathbf{H}) I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right] \\ &= E_{\mathbf{H}} \left[C_1(\mathbf{H}) I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\} \right] \end{aligned}$$

and

$$\begin{aligned} E_{\mathbf{H}} \left[\sum_{a=1}^{K-1} \{\mu_{(K)}(\mathbf{H}) - \mu_{(a)}(\mathbf{H})\} I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right] &\leq E_{\mathbf{H}} \left[\sum_{a=1}^{K-1} \{C_2(\mathbf{H}) I\{g(\mathbf{H}) = l_a(\mathbf{H})\} \right] \\ &= E_{\mathbf{H}} \left[C_2(\mathbf{H}) I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\} \right], \end{aligned}$$

Author Manuscript

where $C_1(\mathbf{H}) = \mu_{(K)}(\mathbf{H}) - \mu_{(K-1)}(\mathbf{H})$ and $C_2(\mathbf{H}) = \mu_{(K)}(\mathbf{H}) - \mu_{(1)}(\mathbf{H})$. These two contrasts indicate the minimum and maximum expected losses in the outcome, respectively, if a subject does not receive the optimal treatment, and thus are adaptive to each patient's own treatment effect ordering.

In the best (least conservative) case where sub-optimal treatments only lead to minimal expected losses in the outcome, g^{opt} is equal to

$$arg \min_{g \in \mathcal{G}} E_{\mathbf{H}} \left[C_1(\mathbf{H}) I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\} \right], \tag{2}$$

while in the worst (most conservative) case where sub-optimal treatments all lead to maximal expected losses in the outcome, g^{opt} is equal to

$$arg \min_{g \in \mathcal{G}} E_{\mathbf{H}} \left[C_2(\mathbf{H}) I\{g(\mathbf{H}) \neq l_K(\mathbf{H})\} \right]. \tag{3}$$

We propose to estimate g^{opt} via (2) and (3) for the following reasons. By using the adaptive contrasts $C_1(\mathbf{H})$ and $C_2(\mathbf{H})$, (2) and (3) minimize, respectively, the lower and the upper bounds of the expected loss in the outcome due to sub-optimal treatments in the entire population of interest. Note that both the lower and the upper bounds of the expected loss have a limiting value of zero that can be reached with perfect classification, implying that (2) and (3) tend to g^{opt} as the expected loss goes to zero. Even when the classification is far from perfect, by minimizing the expected weighted misclassification error, (2) and (3) tend to classify as many patients as possible to their optimal treatment l_K with more focus on subjects with larger contrasts, which is consistent with g^{opt}. Therefore, we expect (2) and (3) to yield an optimal treatment regime similar, if not identical, to g^{opt}. Moreover, using the adaptive contrasts $C_1(\mathbf{H})$ and $C_2(\mathbf{H})$ simplifies the problem of optimization with multiple treatment comparisons to a weighted classification problem that many existing statistical learning methods can handle, for example, classification and regression tree (CART) (Breiman et al., 1984) and SVM. These classification methods aim to reduce the difference between the true and the estimated classes by minimizing an objective function, which is the expected weighted misclassification error in our case.

Author Manuscript

The key to identifying the optimal treatment regime lies in the estimation of $\mu_A(\mathbf{H})$ and $l_A(\mathbf{H})$. Wang et al. (2016) show that given root-n consistent estimators $\hat{\mu}_k(\mathbf{H}), k = 1, \ldots, K$, the corresponding orders $\hat{l}_k(\mathbf{H})$ are also consistent. An intuitive approach is to posit a parametric regression model for $\mu_A(\mathbf{H}) = E(Y|A, \mathbf{H})$ to get the regression estimator $\hat{\mu}_A^{RG}(\mathbf{H})$, and then we can obtain $\hat{g}^{opt}(\mathbf{H}) = \hat{l}_K^{RG}(\mathbf{H})$ directly from $\hat{\mu}_A^{RG}(\mathbf{H})$. Alternatively, instead of using solely the regression model to infer g^{opt}, we could use it as the working model to estimate treatment effect ordering and adaptive contrasts, and then solve the weighted classification problems (2) and (3). However, both methods are susceptible to the misspecification of $\mu_A(\mathbf{H})$ by using $\hat{\mu}_A^{RG}(\mathbf{H})$. If sample size is sufficiently large, one may estimate $\mu_A(\mathbf{H})$ using nonparametric methods, for example, random forest (Breiman, 2001). To balance robustness and efficiency, we propose to apply the AIPW estimator (Robins et al., 1994; Scharfstein et al., 1999). The K treatment options can be regarded as K arbitrary missing data patterns as in Rotnitzky et al. (1998). Given the estimated propensity score $\hat{\pi}_a(\mathbf{H})$, the AIPW estimator $\hat{\mu}_a^{AIPW}$ for $\mu_a = E(Y|A = a)$ is calculated by solving

$$\mathbb{P}_n\left\{\frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}(Y - \mu_a) + U(\mathbf{H})\right\} = 0$$

with the augmentation term

$$U(\mathbf{H}) = \sum_{k \neq a}\left\{I(A = k) - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}\hat{\pi}_k(\mathbf{H})\right\}\phi_k(\mathbf{H}).$$

Here $\phi_k(\mathbf{H})$ is an arbitrary function for treatment k, which could potentially improve the efficiency of the AIPW estimator and meanwhile does not affect the consistency of the AIPW estimator as long as the model for $\pi_a(\mathbf{H})$ is correctly specified. To incorporate the doubly robust property, we propose to set $\phi_k(\mathbf{H}) = \hat{\mu}_a(\mathbf{H}) - \mu_a$ for all $k \neq a$, and then it is straightforward to show that

$$\hat{\mu}_a^{AIPW} = \mathbb{P}_n\left[\frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}Y + \left\{1 - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}\right\}\hat{\mu}_a(\mathbf{H})\right].$$

Notice $\mu_a = E_{\mathbf{H}}\{\mu_a(\mathbf{H})\}$ and thus we define

$$\hat{\mu}_a^{AIPW}(\mathbf{H}) = \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}Y + \left\{1 - \frac{I(A=a)}{\hat{\pi}_a(\mathbf{H})}\right\}\hat{\mu}_a(\mathbf{H}). \tag{4}$$

$\mathbb{P}_n\{\hat{\mu}_a^{AIPW}(\mathbf{H})\}$ converges to μ_a if either the model for $\pi_a(\mathbf{H})$ or the model for $\mu_a(\mathbf{H})$ is correctly specified, and thus the method is doubly robust. To apply the weighted classification problems

Author Manuscript

(2) and (3), we obtain the working orders $\hat{l}_a^{AIPW}(\mathbf{H})$ by sorting $\hat{\mu}_1^{AIPW}(\mathbf{H}), \ldots, \hat{\mu}_K^{AIPW}(\mathbf{H})$ and calculate the AIPW adaptive contrasts $\hat{C}_1^{AIPW}(\mathbf{H}) = \hat{\mu}_{(K)}^{AIPW}(\mathbf{H}) - \hat{\mu}_{(K-1)}^{AIPW}(\mathbf{H})$ and $\hat{C}_2^{AIPW}(\mathbf{H}) = \hat{\mu}_{(K)}^{AIPW}(\mathbf{H}) - \hat{\mu}_{(1)}^{AIPW}(\mathbf{H})$.

For continuous outcomes, a simple and oftentimes reasonable $\hat{\mu}_a(\mathbf{H})$ can be obtained as the regression estimator $\hat{\mu}_a^{RG}(\mathbf{H})$ from a parametric linear model with coefficients dependent on treatment:

$$E(Y|A, \mathbf{H}) = \sum_{a=1}^{K} (\beta_a^\top \mathbf{H}^a) I(A = a), \tag{5}$$

where $\mathbf{H}^a, a = 1, \ldots, K$, are (potentially treatment dependent) summaries of the history $\mathbf{H}$ with the addition of a constant, or intercept, term, and β_a is a parameter vector for $\mathbf{H}^a$ under treatment a. For binary and count outcomes, it is straightforward to extend the method by using generalized linear models. For survival outcomes with non-informative censoring, one may use an accelerated failure time model to predict survival time for all patients. Survival outcomes with more complex censoring issues are beyond the scope of this study. The propensity score can be estimated via multinomial logistic regression (Menard, 2002). A working model could include all variables in $\mathbf{H}$ as linear main effect terms. Summary variables or interaction terms may also be included based on scientific knowledge.

Author Manuscript

2.3 *ACWL with* $T > 1$

The method proposed in Section 2.2 can be generalized to a multi-stage situation by estimating the treatment effect ordering and adaptive contrasts and applying weighted classification at each stage. Based on the idea of backward induction, we develop the following dynamic statistical learning procedure of ACWL.

For stage T, the assumptions and the way to derive the method are the same as in Section 2.2, except that we redefine the counterfactual outcome for a patient following regime g_T as

$$Y^*(A_1, \ldots, A_{T-1}, g_T) = \sum_{a_T=1}^{K_T} Y^*(A_1, \ldots, A_{T-1}, a_T) I\{g_T(H_T) = a_T\},$$

where $Y^*(A_1,\ldots,A_{T-1},a_T)$ is the counterfactual outcome for a patient treated with a_T conditional on previous treatments $(A_1,\ldots,A_{T-1})$. Let $\mu_{T,a_T}(\mathbf{H}_T)$ denote $E(Y|A_T = a_T,\mathbf{H}_T)$, we have

$$g_T^{opt} = arg \max_{g_T\in\mathcal{G}_T} E_{\mathbf{H_T}}\left[\sum_{a_T=1}^{K_T}\mu_{T,a_T}(\mathbf{H}_T)I\{g_T(\mathbf{H}_T)=a_T\}\right].$$

For stage j, $T-1\geq j\geq 1$, we combined the method in Section 2.2 with machine learning methods to conduct backward induction. Following Moodie et al. (2012), the stage-specific pseudo-outcome PO_j for estimating treatment effect ordering and adaptive contrasts is a predicted counterfactual outcome under optimal treatments at all future stages, also known as the ”optimal benefit-to-go” in Murphy (2005). Specifically, we have

$$PO_j = E\left\{Y^*(A_1,\ldots,A_j,g_{j+1}^{opt},\ldots,g_T^{opt})\right\},$$

or in a recursive form,

$$PO_j = E\{PO_{j+1}|A_{j+1}=g_{j+1}^{opt}(\mathbf{H}_{j+1}),\mathbf{H}_{j+1}\}$$

and we set $PO_T = Y$. For $a_j = 1,\ldots,K_j$, let $\mu_{j,a_j}(\mathbf{H}_j)$ denote the conditional mean $E[PO_j|A_j = a_j,\mathbf{H}_j]$, and we have $PO_j = \mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1})$. We replace Y with PO_j to apply the method in Section 2.2 at stage j. Specifically, let $PO_j^*(a_j)$ denote the counterfactual pseudo-outcome for a patient with treatment a_j at stage j. We have the consistency assumption as $PO_j = \sum_{a_j=1}^{K_j} PO_j^*(a_j)I(A_j = a_j)$, NUCA as $\{PO_j^*(1),\ldots,PO_j^*(K_j)\}\perp\!\!\!\perp\mathbf{H}_j$ and the positivity assumption as $\pi_{a_j}(\mathbf{H}_j) = Pr(A_j = a_j|\mathbf{H}_j)$ being bounded away from zero. With these three assumptions, we identify the optimal regime directly following Section 2.2 and get g_j^{opt} among all potential regimes $\mathcal{G}_j$ as

$$g_j^{opt} = arg \max_{g_j\in\mathcal{G}_j} E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j}\mu_{j,a_j}(\mathbf{H}_j)I\{g_j(\mathbf{H}_j)=a_j\}\right],$$

or equivalently,

$$g_j^{opt} = arg \min_{g_j\in\mathcal{G}_j} E_{\mathbf{H}_j}\left[\sum_{a_j=1}^{K_j-1}\{\mu_{j,(K)}(\mathbf{H}_j)-\mu_{j,(a)}(\mathbf{H}_j)\}I\{g_j(\mathbf{H}_j)=l_{a_j}(\mathbf{H}_j)\}\right], \tag{6}$$

where $\mu_{j,(1)}(\mathbf{H}_j)\leq\ldots\leq\mu_{j,(K)}(\mathbf{H}_j)$ denote the treatment effect ordering and the order $l_{a_j}(\mathbf{H}_j)$ means $\mu_{j,(a_j)}(\mathbf{H}_j) = \mu_{j,l_{a_j}}(\mathbf{H}_j)$.

Author Manuscript

Again, the optimization problem (6) is complicated by the multiple treatment comparisons. Therefore, we incorporate the adaptive contrasts as in Section 2.2 for each stage. Specifically, the adaptive contrasts are $C_{j,1}(\mathbf{H}_j) = \mu_{j,(K)}(\mathbf{H}_j) - \mu_{j,(K-1)}(\mathbf{H}_j)$ and $C_{j,2}(\mathbf{H}_j) = \mu_{j,(K)}(\mathbf{H}_j) - \mu_{j,(1)}(\mathbf{H}_j)$, which indicate respectively, the minimum and the maximum expected losses in the pseudo-outcome, if a patient does not receive the optimal treatment at stage j. Via the adaptive contrasts, we transform the problem of optimization with multiple treatment comparisons to a simpler weighted classification problem.

We start the estimation with stage T and conduct backward induction. Our ACWL algorithm starting with stage $j = T$ is carried out as follows:

Step 1: Fit regression model (5) with pseudo-outcome PO_j to obtain regression-based conditional mean estimator $\hat{\mu}^{RG}_{j,a_j}(\mathbf{H}_j)$.

Step 2: Fit the propensity score model to obtain $\hat{\pi}_{j,a_j}(\mathbf{H}_j)$.

Step 3: Calculate AIPW-based conditional mean estimator $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j)$ using $\hat{\mu}^{RG}_{j,a_j}(\mathbf{H}_j)$ and $\hat{\pi}_{j,a_j}(\mathbf{H}_j)$ as in (4).

Step 4: Calculate the AIPW-based working orders $\hat{l}^{AIPW}_{j,a_j}(\mathbf{H}_j)$ and adaptive contrasts $\hat{C}^{AIPW}_{j,1}(\mathbf{H}_j)$ and $\hat{C}^{AIPW}_{j,2}(\mathbf{H}_j)$ using $\hat{\mu}^{AIPW}_{j,a_j}(\mathbf{H}_j)$.

Step 5: Take $\hat{l}^{AIPW}_{j,K}(\mathbf{H}_j)$ as the class label, and $\hat{C}^{AIPW}_{j,1}(\mathbf{H}_j)$ and $\hat{C}^{AIPW}_{j,2}(\mathbf{H}_j)$ as the weights to solve problems (2) and (3) using existing classification techniques.

Step 6: If $j > 1$, set $j = j - 1$ and repeat steps 1 to 6. If $j = 1$, stop.

When the outcome is cumulative (e.g., the sum of longitudinally observed values or a single continuous final outcome), we modify the pseudo-outcomes to reduce accumulated bias from the conditional mean models, following Huang et al. (2015). For stage j, $T - 1 \geq j \geq 1$, instead of using only the model-based values under optimal future treatments, i.e., $\mu_{j+1,g^{opt}_{j+1}}(\mathbf{H}_{j+1})$, we use the actual observed outcomes plus the expected future loss due to sub-optimal treatments.

Author Manuscript

Specifically, the modified pseudo-outcome is

$$PO_j^{'} = PO_{j+1}^{'} + \mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1}),$$

where a_{j+1} is the treatment that a patient actually received at stage $j+1$, and $\mu_{j+1,g_{j+1}^{opt}}(\mathbf{H}_{j+1}) - \mu_{j+1,a_{j+1}}(\mathbf{H}_{j+1})$ is the expected loss due to sub-optimal treatments at stage $j+1$ for a given patient, which is zero if $g_{j+1}^{opt}(\mathbf{H}_{j+1}) = a_{j+1}$ and positive otherwise. Again we set $PO'_T = Y$. This modification leads to more robustness against model misspecification and is less likely to accumulate bias from stage to stage during backward induction (Huang et al., 2015).

3. Simulation Studies

Author Manuscript

We conduct simulation studies to evaluate the performance of our proposed method in two aspects. First, we need to evaluate whether $\hat{g}^{opt}$ estimated through weighted classification with adaptive contrasts is close enough to the truth in numerical studies. Second, we aim to show the robustness of our methods with different levels of model misspecification. To achieve this, we purposely set all regression models μ to be misspecified, as is the case for most real data applications, and let the propensity model π be either correctly (e.g., randomized trials) or incorrectly (e.g., most observational studies) specified. We consider a single-stage scenario as in Section 2.2 and a multi-stage scenario as in Section 2.3, each with 500 replications. For both scenarios, we generate five baseline covariates $X_1, \dots, X_5$ according to $N(0,1)$, and set the expected counterfactual outcome under the optimal treatment regime, i.e., $E\{Y^*(\mathbf{g}^{opt})\}$, to be 8. We use CART to minimize the weighted misclassification error, which is implemented by the R package *rpart*.

3.1 *Scenario 1:* $T = 1$ *and* $K = 5$

In Scenario 1, we consider a single stage with five treatment options and sample size of 1000. We generate treatment A from $Multinomial(\pi_0/\pi_s, \pi_1/\pi_s, \pi_2/\pi_s, \pi_3/\pi_s, \pi_4/\pi_s)$, with $\pi_0 = 1$, $\pi_1 = \exp(0.5 - 0.5X_1)$, $\pi_2 = \exp(0.5X_1 + 0.2)$, $\pi_3 = \exp(0.5X_5 + 0.1)$, $\pi_4 = \exp(0.5X_5 - 0.1)$, and $\pi_s = \sum_{m=0}^{4} \pi_m$. We set A to take values in $\{0, \ldots, 4\}$ and generate outcomes as

$$Y = \exp[2.06 + 0.2X_3 - |X_1 + X_2|\varphi\{A, g^{opt}(\mathbf{H})\}] + \epsilon,$$

with $\varphi\{A, g^{opt}(\mathbf{H})\}$ taking the form of $\varphi^{(1)} = 3I\{A \neq g^{opt}(\mathbf{H})\}$ or $\varphi^{(2)} = \{A - g^{opt}(\mathbf{H})\}^2$, $g^{opt}(\mathbf{H}) = I(X_1 > -1)\{1 + I(X_2 > -0.4) + I(X_2 > 0.4) + I(X_2 > 1)\}$ and $\epsilon \sim N(0, 1)$.

The function $\varphi\{A, g^{opt}(\mathbf{H})\}$ indicates the penalty if a patient does not receive the optimal treatment. Given $\varphi^{(1)}$, misclassification to any of the four sub-optimal treatments leads to the same expected loss in the outcome for a given patient, which means that all $K - 1$ contrasts in (1) are actually the same for that patient. In this case, (2) and (3) are both identical to g^{opt} and we expect them to have good performances. With $\varphi^{(2)}$, we consider a more common situation where the differences among treatments vary, and misclassification to a treatment closer to the optimal one leads to a smaller expected loss in the outcome. In this case, the $K - 1$ contrasts are not all the same and therefore, (2) and (3) are not identical to g^{opt}. Simulation studies under $\varphi^{(1)}$ and $\varphi^{(2)}$ investigate the performance of ACWL and see how close (2) and (3) are tending to g^{opt}. Under each form of $\varphi\{A, g^{opt}(\mathbf{H})\}$, we further assess the robustness of our method. By using linear regression, we have a misspecified conditional mean model. For the propensity score, we consider both a correctly specified model $log(\pi_d/\pi_0) = \beta_{0d} + \beta_{1d}X_1 + \beta_{2d}X_5, d = 1, \ldots, 4$, and an incorrectly specified one $log(\pi_d/\pi_0) = \beta_{0d}$.

We apply the proposed ACWL algorithm to each simulated dataset and denote the methods using the two adaptive contrasts as ACWL-C_1 and ACWL-C_2, respectively. For comparison, we use the regression-based conditional mean models directly to infer the optimal DTRs and we denote this method as RG. We also use the contrasts and orders estimated from the conditional mean models to apply weighted classification (2) and (3), and denote these two methods as RG-C_1 and RG-C_2. Furthermore, we apply the OWL method by Zhao et al. (2012) with CART.

[Table 1 about here.]

Table 1 summarizes the performances of all methods considered in Scenario 1, in terms of the

Author Manuscript

percentage of subjects correctly classified to their optimal treatments, denoted as $opt\%$, and the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime, denoted as $\hat{E}\{Y^*(\hat{g}^{opt})\}$. $opt\%$ shows how likely the estimated optimal regime is to assign a new patient to his or her real optimal treatment and $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows how much the entire population of interest will benefit from following $\hat{g}^{opt}$. The regression-based methods RG, RG-C_1 and RG-C_2 have relatively poor performances since the conditional mean model is misspecified. They classify $55 \sim 59\%$ patients to their optimal treatments, resulting in a $\hat{E}\{Y^*(\hat{g}^{opt})\}$ much smaller than the true value of 8. OWL has relatively good performance only when the propensity score model is correctly specified, as expected, and it is the least efficient among all methods considered with large empirical standard deviations (SDs) for both $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$. Our proposed method classifies over 84% patients to their optimal treatments in all cases and achieves $\hat{E}\{Y^*(\hat{g}^{opt})\}$ close to 8. ACWL is highly robust against model misspecification with only slight decrease in performance from using a correctly specified propensity score model to using an incorrectly specified one. Under $\varphi^{(1)}$ when all $K-1$ contrasts are the same, both (2) and (3) are equal to g^{opt} and thus yield satisfactory $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$. From $\varphi^{(1)}$ to $\varphi^{(2)}$, the regression-based methods show improved $\hat{E}\{Y^*(\hat{g}^{opt})\}$ despite similar $opt\%$, indicating higher sensitivity to subjects with larger contrasts given varying expected losses due to sub-optimal treatments. Although $K-1$ contrasts are not all the same under $\varphi^{(2)}$, ACWL-C_1 and ACWL-C_2 show very slight deterioration in $opt\%$ and $\hat{E}\{Y^*(\hat{g}^{opt})\}$, compared to the results under $\varphi^{(1)}$, and are still much better than the other methods. These results confirm the feasibility of estimating g^{opt} via ACWL with adaptive AIPW contrasts.

Author Manuscript

3.2 *Scenario 2: $T = 2$ and $K_1 = K_2 = 3$*

In this section, we generate data under a two-stage DTR with three treatment options at each stage. We consider the outcome of interest as the sum of the rewards from each stage, i.e., $Y = R_1 + R_2$, and set φ to be the form as $\varphi^{(2)}$ in Scenario 1. We evaluate the performance of our proposed method given a misspecified conditional mean model through linear regression, while allowing the propensity score models to be either correctly or incorrectly specified. Furthermore, since we apply CART for classification, we consider both a tree-type underlying optimal DTR and a non-tree-type one. We consider sample sizes of 500 and 1000.

Treatment variables are set to take values in $\{0, 1, 2\}$ at each stage. For stage 1, we generate A_1 from $Multinomial(\pi_{10}, \pi_{11}, \pi_{12})$, with $\pi_{10} = 1/\{1 + \exp(0.5 - 0.5X_3) + \exp(0.5X_4)\}$, $\pi_{11} = \exp(0.5 - 0.5X_3)/\{1 + \exp(0.5 - 0.5X_3) + \exp(0.5X_4)\}$ and $\pi_{12} = 1 - \pi_{10} - \pi_{11}$. We generate stage 1 reward as

$$R_1 = \exp[1.5 - |1.5X_1 + 2|\{A_1 - g_1^{opt}(\mathbf{H}_1)\}^2] + \epsilon_1,$$

with tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -1)\{I(X_2 > -0.5) + I(X_2 > 0.5)\}$ or non-tree-type $g_1^{opt}(\mathbf{H}_1) = I(X_1 > -0.5)\{1 + I(X_1 - X_2 > 0)\}$, and $\epsilon_1 \sim N(0, 1)$.

For stage 2, we have treatment $A_2 \sim Multinomial(\pi_{20}, \pi_{21}, \pi_{22})$, with $\pi_{20} = 1/\{1+\exp(0.2R_1 - 1) + \exp(0.5X_4)\}$, $\pi_{21} = \exp(0.2R_1 - 1)/\{1 + \exp(0.2R_1 - 1) + \exp(0.5X_4)\}$ and $\pi_{22} = 1 - \pi_{20} - \pi_{21}$. We generate stage 2 reward as

$$R_2 = \exp[1.26 - |1.5X_3 - 2|\{A_2 - g_2^{opt}(\mathbf{H}_2)\}^2] + \epsilon_2,$$

with tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > -1)\{I(R_1 > 0.5) + I(R_1 > 3)\}$ or non-tree-type $g_2^{opt}(\mathbf{H}_2) = I(X_3 > 0) + I(X_3 + R_1 > 2.5)$, and $\epsilon_2 \sim N(0, 1)$.

We apply the proposed ACWL algorithm with the modified pseudo-outcome to each simulated dataset. For comparison, we use the regression-based conditional mean models directly to infer the optimal DTR, which is Q-learning. We also apply the backward OWL (BOWL) method by Zhao et al. (2015) with CART. As BOWL does not involve outcome regression models, only subjects whose observed treatments are optimal at stage 2 can be used for identifying the optimal regime at stage 1, resulting in a significantly reduced sample size. Therefore, we also consider BOWL combined with Q-learning, denoted as BOWL-Q. Basically, at stage 1, we use the conditional mean

Author Manuscript

model from Q-learning to predict the pseudo-outcome for patients whose observed treatments are not optimal at stage 2 and then apply OWL using all subjects to identify the optimal regime.

[Table 2 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

Results for Scenario 2 are shown in Table 2. The regression-based conditional mean models explain about 34% of the total variance at stage 2 and 20% of the total variance at stage 1. Q-learning is relatively stable with different sample sizes while all classification-based methods show clear improvement with an increased sample size. The two OWL methods are the least efficient with large empirical SDs. BOWL-Q has higher $opt\%$ and $\hat{E}\{Y^*(\hat{\mathrm{g}}^{opt})\}$ but also larger variability than BOWL, implying a bias and variance trade-off by incorporating misspecified but informative regression models. Similarly as in Scenario 1, ACWL has the best performance among all methods considered with average $opt\%$ over 80% and $\hat{E}\{Y^*(\hat{\mathrm{g}}^{opt})\}$ closest to 8 in all cases. ACWL is also very robust against misspecification of the propensity score model while BOWL and BOWL-Q have significant deterioration in performance with a misspecified propensity score model. From a tree-type underlying optimal DTR to a non-tree-type one, all CART-based methods show worse performance. For our proposed method, $opt\%$ decreases by approximately 10% and $\hat{E}\{Y^*(\hat{\mathrm{g}}^{opt})\}$ drops $0.3 \sim 0.6$, yet still much better than all the other methods. Figures 1 and 2 further shows how the methods perform in predicting the optimal treatments for new subjects with correctly specified propensity score models, sample size of 1000 and the underlying optimal DTR being tree-type and non-tree-type, respectively. We only present the results from ACWL-C_2 given the similarity between ACWL-C_1 and ACWL-C_2. In Figure 1, ACWL leads to clear differentiation of the three regions, which matches the true underlying DTR, while in Figure 2, there are more misclassified cases near the borders, likely due to the use of CART for the non-tree-type underlying DTR. In both figures, ACWL shows superior performances compared to Q-learning and BOWL.

Author Manuscript

Notably, in both single-stage and multi-stage scenarios, ACWL is robust and efficient compared to the other methods, even with misspecified models for both outcome and propensity score. This may be due to the following reasons. First, the treatment effect ordering and adaptive contrasts are constructed using the doubly robust AIPW estimator. Second, we utilize the flexible weighted classification, instead of using the orders and contrasts directly, to estimate the optimal DTR, which further improves robustness. Comparing ACWL-C_1 and ACWL-C_2, we do not have a clear conclusion on which one is better. We suggest implementing both and choosing the optimal DTR by taking the common part or by incorporating background knowledge. Additional simulation studies can be found in the Web-based Supplementary Materials (S-Tables 1-3), which lead to a similar conclusion. ACWL becomes less efficient with more treatment options or more stages but still performs much better than the other competing methods.

4. Application to the Esophageal Cancer Example

As a further illustration, we apply ACWL to the esophageal cancer data collected by MD Anderson Cancer Center from 1998 to 2012. At baseline, we have $n = 1170$ patients with about 90% at overall cancer stage II or III (Byrd et al., 2010). The general disease management strategy is chemotherapy or chemoradiation therapy (CRT) followed by surgery (Lloyd and Chang, 2014).

[Figure 3 about here.]

Figure 3 shows the two-stage disease management before surgery in our observational data. At baseline, all patients had records of basic characteristics and disease status, including a total 11 covariates, denoted by $X_{1,1}, \ldots, X_{1,11}$. At treatment stage 1, about 40% of the patients received induction chemotherapy (ICT), denoted by A_1 with YES for treated and NO for untreated. Tumor response was measured right before treatment stage 2, denoted as X_2, which is an intermediate variable. X_2 takes values from 0 to 5 with 0 being progression and 5 being complete response, compared to baseline tumor measures. At stage 2, all patients received CRT with one of three radiation modalities: 3D conformal radiotherapy (3DCRT, 39% of the total patients), intensity-

Author Manuscript

modulated radiation therapy (IMRT, 45%) and proton therapy (PT, 16%). We use A_2 to denote the stage 2 treatment variable. After CRT, tumor response and development of new lesions were measured within three months (before surgery), denoted as $R_{3,1}$ (same scale as X_2) and $R_{3,2}$ (0 for development of new lesions and 1 for none), respectively. We focus on these two stages to estimate the optimal DTR to decide whether a patient should receive ICT at stage 1 and what radiation modality should be used for CRT at stage 2. We define a single outcome $Y = R_{3,1} + 2R_{3,2}$ to measure the effectiveness of the two-stage treatments, and side effects (e.g., nausea, anorexia and fatigue) are not included in the evaluation because most of them would go away shortly after CRT. Missing data is imputed using IVEware (Raghunathan et al., 2002).

We apply the ACWL algorithm to the data described as above. Specifically, the covariate and treatment history just prior to stage 2 treatment is $\mathbf{H}_2 = (X_{1,1}, \ldots, X_{1,11}, A_1, X_2)$ and the number of treatment options at stage 2 is $K_2 = 3$. We fit a linear regression model for $\mu_{2,A_2}(\mathbf{H}_2)$ as in (5) using Y as the outcome and all variables in $\mathbf{H}_2$ as predictors that interact with A_2. For the propensity score, we fit a multinomial logistic regression model including main effects of all variables in $\mathbf{H}_2$. We use CART with pruning for weighted classification. We repeat the same procedure for stage 1 except that we have $\mathbf{H}_1 = (X_{1,1}, \ldots, X_{1,11})$, $K_1 = 2$ and $PO_1^{'} = Y + \hat{\mu}_{2,\hat{g}_2^{opt}}(\mathbf{H}_2) - \hat{\mu}_{2,a_2}(\mathbf{H}_2)$.

We find very similar results using ACWL-C_1 and ACWL-C_2, and thus combine the results by using variables that both methods identify as important (CART variable importance ≥ 15). For stage 1, the most important variables are tumor length (mm, continuous) and overall clinical stage (I/II vs. III/IV). For stage 2, the most important variables are stage 1 treatment A_1, intermediate tumor response X_2 and baseline tumor differentiation (well/moderate vs. poor). The estimated optimal DTR $\hat{\mathbf{g}}^{opt} = c(\hat{g}_1^{opt}, \hat{g}_2^{opt})$ is

$$\hat{g}_1^{opt}(\mathbf{H}_1) = \begin{cases} \text{YES} & \text{if tumor length } \geq 36mm \text{ or stage = III/IV} \\ \text{NO} & \text{otherwise} \end{cases}$$

Author Manuscript

and

$$\hat{g}_2^{opt}(\mathbf{H}_2) = \begin{cases} \text{PT} & \text{if } A_1 = \text{NO and tumor differentiation = poor} \\ \text{IMRT} & \text{if } A_1 = \text{YES and intermediate tumor response } < 4 \\ \text{3DCRT} & \text{otherwise} \end{cases}$$

As suggested by the estimated optimal DTR, ICT is recommended at stage 1 for patients with larger tumor or worse clinical stage, which is consistent with clinical findings that the addition of ICT is appropriate for advanced disease with high risk for local or distant failure (Haddad et al., 2013). Some studies have shown that ICT is beneficial overall for both tumor control and prolonging survival (Jin et al., 2004) but there have not been randomized trials or studies focusing on subgroups of patients. At stage 2, our result suggests that patients who do not use ICT and have poor tumor differentiation should use PT in CRT, patients with ICT and minor or worse tumor response after ICT should use IMRT and all other patients should use 3DCRT. Currently, there has not been any large trial comparing the three radiation modalities. Some studies have shown that PT and IMRT are more efficient at targeting the tumors and less toxic than 3DCRT (Lloyd and Chang, 2014), which may explain why our result suggests PT or IMRT for patients with worse conditions.

5. Discussion

We have proposed a robust and efficient method ACWL to estimate the optimal DTR, which can effectively handle multiple treatment options at each stage. The adaptive contrasts we develop at each stage simplify the problem of optimization with multiple treatment comparisons to a dynamic weighted learning procedure, and our simulations studies show that this simplification leads to excellent numerical performances. Our method combines robust semiparametric regression estimators with flexible machine learning methods. With regression models at each stage, one can predict the future outcomes under optimal treatments for patients whose assigned treatments are not all optimal at future stages, thus improving efficiency if the regression models are well approximated. The doubly robust AIPW estimator and nonparametric classification method that

Author Manuscript

we utilize help improve the robustness of ACWL against model misspecification. Therefore, our proposed method is capable of dealing with observational data. Moreover, the dynamic ACWL algorithm can be easily implemented with existing regression and classification methods.

Several improvements and extensions can be explored in future studies. Generalizing the ACWL method to handle informatively censored data is clinically meaningful as many studies focus on prolonging patients' survival. Goldberg and Kosorok (2012) has developed a method within the Q-learning framework by using inverse-probability-of-censoring weighting (IPCW). With ACWL, one may combine the probability of treatment with the probability of censoring in the AIPW estimator. Due to the flexibility in the ACWL algorithm, many other machine learning methods can be considered, for both the classification part (e.g., SVM or other tree-based learning methods) and the backward induction part (e.g., A-learning). Moreover, with high dimensional data, one can incorporate variable selection at each stage for the conditional mean models. In addition, it may be of great practical interest to explore generalization of ACWL with continuous treatment options, such as radiation dose.

6. Supplementary Materials

Web Tables referenced in Section 3 for additional simulation studies and sample R codes for implementing the proposed method are available with this paper at the *Biometrics* website on Wiley Online Library.

Author Manuscript

References

Bather, J. (2000). *Decision Theory: An Introduction to Dynamic Programming and Sequential Decisions*. New York: Wiley.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Breiman, L., Freidman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Byrd, D. R., Compton, C. C., Fritz, A. G., Greene, F. L., and Trotti, A. (2010). *AJCC Cancer Staging Manual (Vol. 649).* Springer, New York.

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive moutofn bootstrap scheme. *Biometrics* **69,** 714–723.

Cortes, C. and Vapnik, V. (1995). Support-vector netowrks. *Machine Learning* **20,** 273–297.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of Statistics* **40,** 529–560.

Haddad, R., O'Neill, A., Rabinowits, G., Tishler, R., Khuri, F., Adkins, D., Sarlis, N amd Jochen Lorch, J., Beitler, J. J., Limaye, S., Riley, S., and Posner, M. (2013). Induction chemotherapy followed by concurrent chemoradiotherapy (sequential chemoradiotherapy) versus concurrent chemoradiotherapy alone in locally advanced head and neck cancer (PARADIGM): a randomised phase 3 trial. *The Lancet Oncology* **14,** 257–264.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *Journal of the American Statistical Association* **96,** 440–448.

Huang, X., Choi, S., Wang, L., and Thall, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Statistics in Medicine.* Advance online publication. doi:10.1002/sim.6558.

Jin, J., Liao, Z., Zhang, Z., Ajani, J., Swisher, S., Chang, J. Y., Jeter, M., Guerrero, T., Stevens, C. W., Vaporciyan, A., Putnam, J. J., Walsh, G., Smythe, R., Roth, J., Yao, J., Allen, P., Cox, J. D., and Komaki, R. (2004). Induction chemotherapy improved outcomes of patients with resectable esophageal cancer who received chemoradiotherapy followed by surgery. *International Journal of Radiation Oncology*Biology*Physics* **60,** 427–436.

Lloyd, S. and Chang, B. W. (2014). Current strategies in chemoradiation for esophageal cancer. *Journal of Gastrointestinal Oncology* **5,** 156–165.

Author Manuscript

Menard, S. (2002). *Applied Logistic Regression Analysis*. Thousand Oaks, CA: Sage, 2nd edition.

Moodie, E. E., Chakraborty, B., and Kramer, M. S. (2012). Q-learning for estimating optimal dynamic treatment rules from observational data. *Canadian Journal of Statistics* **40,** 629–645.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65,** 331–355.

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24,** 1455–1481.

Raghunathan, T. E., Solenberger, P., and Van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software User Guide*. Survey Methodology Program, University of Michigan, Ann Arbor, Michigan.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS* **113,** 159.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, pages 69–117. New York: Springer.

Robins, J. M. (2000). Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. New York: Springer.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. New York: Springer.

Robins, J. M. and Hernán, M. A. (2009). Estimation of the causal effects of time-varying

Author Manuscript

exposures. In *Longitudinal Data Analysis*, pages 553–599. Chapman and Hall/CRC Press: Boca Raton.

Robins, J. M., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Rotnitzky, A., Robins, J., and Scharfstein, D. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93,** 1321–1339.

Scharfstein, D., Rotnitzky, A., and Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94,** 1096–1120.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science* **29,** 640–661.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning: An Introduction.* MIT Press, Cambridge.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E., and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine* **26,** 4687–4707.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2,** 1–40.

Wagner, E. H., Austin, B. T., Davis, C., Hindmarsh, M., Schaefer, J., and Bonomi, A. (2001). Improving chronic illness care: translating evidence into action. *Health affairs* **20,** 64–78.

Wang, F., Wang, L., and Song, P. X. (2016). Fused lasso with the adaptation of parameter ordering (FLAPO) in merging multiple studies with repeated measurements. *Biometrics,* DOI:10.1111/biom.12496.

Wang, L., Rotnitzky, A., Lin, X., Millikan, R. E., and Thall, P. F. (2012). Evaluation of viable

Author Manuscript

dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association* **107,** 493–508.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine Learning* **8,** 279–292.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. B. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1,** 103–114.

Zhao, Y., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association* **110,** 583–598.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107,** 1106–1118.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2015). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association,* DOI:10.1080/01621459.2015.1093947.

Author Manuscript

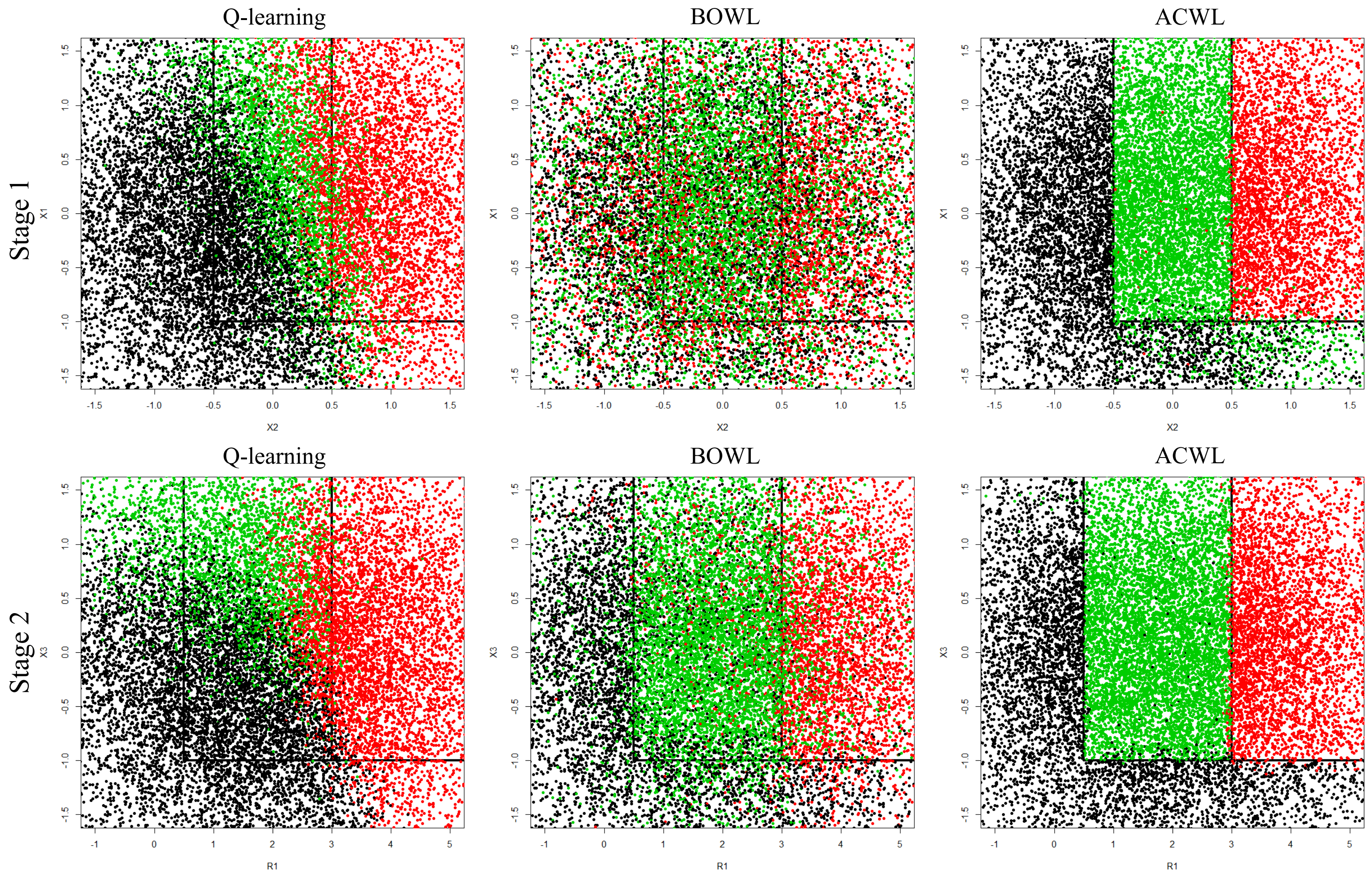

Figure 1: Predicted optimal treatments in simulation Scenario 2 with a tree-type underlying optimal DTR, correctly specified propensity score model and sample size of 1000. The true regions at stage 1 are red for $X_1 > -1$ and $X_2 > 0.5$, green for $X_1 > -1$ and $-0.5 < X_2 \leq 0.5$ and black elsewhere. The true regions at stage 2 are red for $R_1 > 3$ and $X_3 > -1$, green for $0.5 < R_1 \leq 3$ and $X_3 > -1$ and black elsewhere.

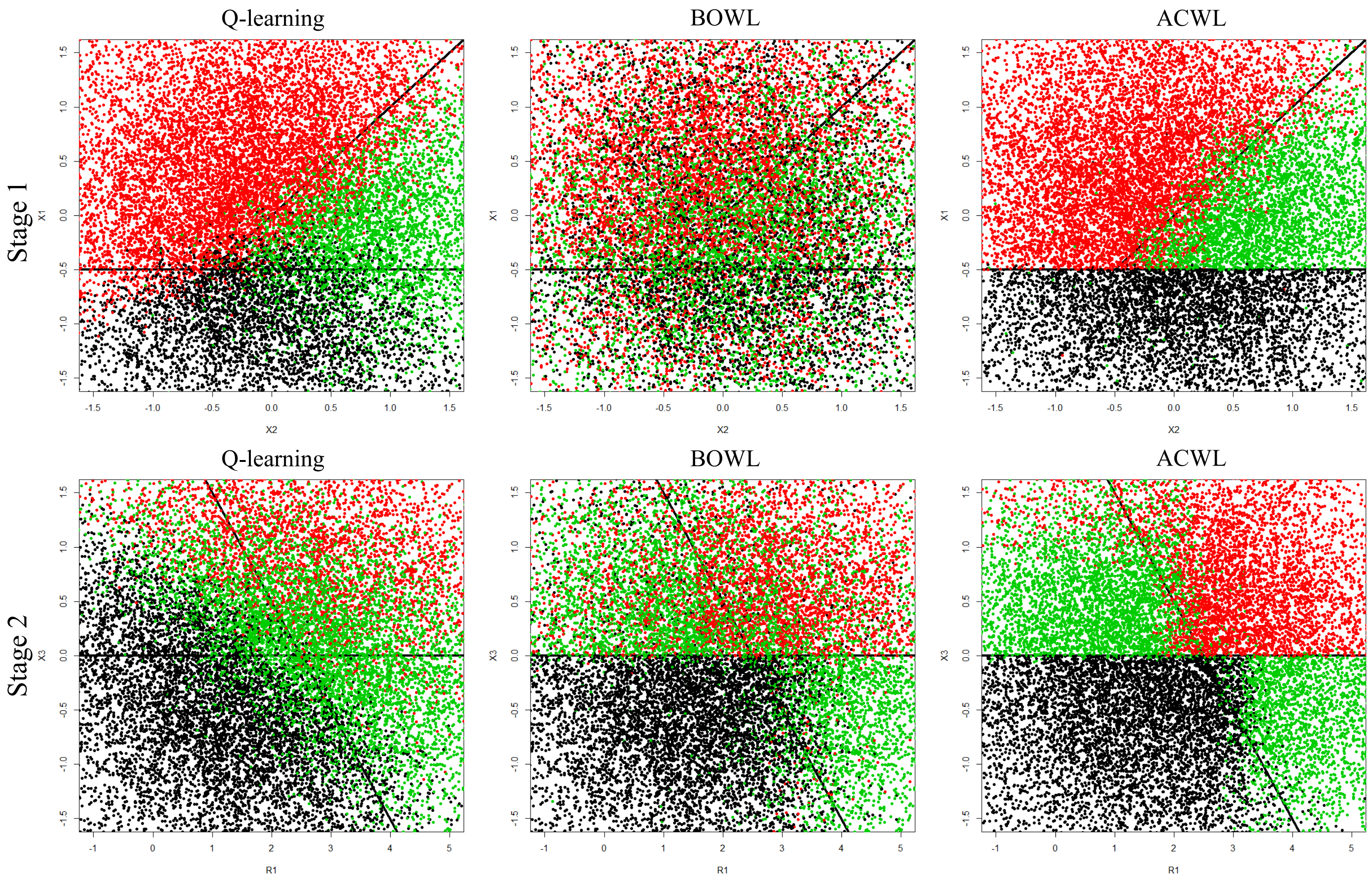

Figure 2: Predicted optimal treatments in simulation Scenario 2 with a non-tree-type underlying optimal DTR, correctly specified propensity score model and sample size of 1000. The true regions at stage 1 are red for $X_1 > 0$ and $X_1 > X_2$, black for $X_1 \leq 0$ and green elsewhere. The true regions at stage 2 are red for $X_3 > 0$ and $R_1 + X_3 > 2.5$, black for $X_3 \leq 0$ and $R_1 + X_3 \leq 2.5$, and green elsewhere.

Auth

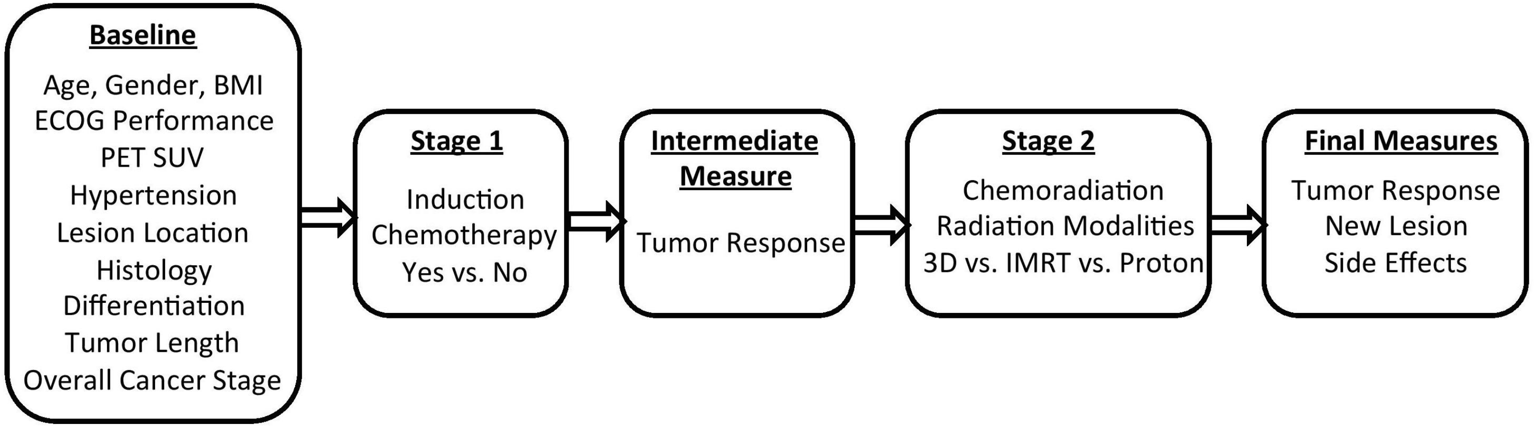

Figure 3: Two-stage disease management for esophageal cancer patients.

Table 1: Simulation results for Scenario 1 with a single stage and five treatment options (500 replications, $n = 1000$). π is the propensity score model. $\varphi^{(1)}$ and $\varphi^{(2)}$ indicate equal and varying penalties for misclassification. $opt\%$ shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{g}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal regime. $E\{Y^*(g^{opt})\} = 8$.

π	Method	$\varphi^{(1)}$		$\varphi^{(2)}$	
		$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{g}^{opt})\}$
-	RG	58.1 (2.6)	5.39 (0.17)	59.5 (3.8)	5.99 (0.25)
	RG-C_1	55.1 (4.1)	5.20 (0.29)	58.2 (6.0)	6.00 (0.37)
	RG-C_2	55.7 (3.8)	5.24 (0.29)	58.4 (5.7)	6.00 (0.34)
Correct	OWL	83.2 (9.2)	6.92 (0.60)	74.6 (11.6)	6.80 (0.56)
	ACWL-C_1	94.2 (3.5)	7.69 (0.21)	88.7 (5.5)	7.60 (0.22)
	ACWL-C_2	90.4 (6.1)	7.38 (0.40)	86.4 (8.4)	7.36 (0.38)
Incorrect	OWL	60.0 (13.8)	5.57 (0.89)	52.0 (11.0)	5.89 (0.65)
	ACWL-C_1	92.5 (4.1)	7.60 (0.23)	84.2 (6.7)	7.47 (0.24)
	ACWL-C_2	90.2 (6.0)	7.37 (0.38)	85.6 (8.2)	7.35 (0.36)

Author Manuscript

Table 2: Simulation results for Scenario 2 with two stages and three treatment options at each stage (500 replications). π is the propensity score model. $opt\%$ shows the empirical mean and standard deviation (SD) of the percentage of subjects correctly classified to their optimal treatments. $\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$ shows the empirical mean and SD of the expected counterfactual outcome obtained using the true outcome model and the estimated optimal DTR. $E\{Y^*(\mathbf{g}^{opt})\} = 8$.

π	Method	Tree-type DTR				Non-tree-type DTR			
		$n = 500$		$n = 1000$		$n = 500$		$n = 1000$	
		$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$	$opt\%$	$\hat{E}\{Y^*(\hat{\mathbf{g}}^{opt})\}$
-	Q-learning	51.2 (3.3)	5.83 (0.23)	53.3 (2.6)	5.97 (0.21)	55.5 (4.2)	6.07 (0.23)	58.0 (3.2)	6.24 (0.20)
Correct	BOWL	28.9 (6.1)	4.30 (0.44)	38.6 (7.9)	4.66 (0.52)	26.1 (5.8)	4.25 (0.40)	34.8 (7.6)	4.56 (0.50)
	BOWL-Q	38.1 (9.3)	5.04 (0.66)	63.2 (10.5)	6.41 (0.59)	31.7 (7.4)	4.69 (0.50)	49.1 (10.3)	5.49 (0.55)
	ACWL-C_1	85.1 (4.7)	7.29 (0.21)	93.3 (3.3)	7.57 (0.13)	74.1 (5.7)	6.68 (0.29)	83.3 (3.8)	7.10 (0.16)
	ACWL-C_2	85.4 (5.3)	7.31 (0.24)	93.7 (3.3)	7.60 (0.13)	77.8 (5.4)	6.83 (0.24)	86.6 (3.6)	7.25 (0.15)
Incorrect	BOWL	23.7 (5.9)	4.05 (0.42)	26.3 (6.5)	4.10 (0.42)	22.1 (4.9)	4.02 (0.34)	23.6 (5.6)	4.11 (0.38)
	BOWL-Q	26.4 (7.3)	4.31 (0.51)	30.3 (8.7)	4.43 (0.61)	24.7 (5.6)	4.30 (0.40)	26.4 (6.7)	4.41 (0.46)
	ACWL-C_1	84.1 (4.9)	7.27 (0.25)	91.8 (3.7)	7.42 (0.17)	72.3 (6.1)	6.60 (0.33)	81.1 (4.1)	7.03 (0.18)
	ACWL-C_2	83.8 (5.9)	7.25 (0.28)	91.8 (3.8)	7.43 (0.18)	76.9 (5.9)	6.65 (0.31)	82.9 (4.0)	7.09 (0.17)

Author Manuscript