Rejoinder

Guanhua Chen[a], Donglin Zeng[b], and Michael R. Kosorok[c]

[a]Department of Biostatistics, Vanderbilt University, Nashville, TN, USA; [b]Department of Biostatistics, University of North Carolina, Chapel, Hill, NC, USA; [c]Department of Biostatistics, and Professor, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC, USA

1. Introduction

We are very grateful for all of the discussants and their comments which have been constructive and idea-provoking. We are pleased that our work has been recognized as an important contribution to statistical methodology in personalized medicine research, and we are grateful for the opportunity to respond briefly to the comments of the discussants. We recognize that there are numerous important insights and questions raised by the discussants, and we regret that we are unable to adequately address each and everyone one of them. However, we have endeavored in what follows to highlight and provide meaningful responses to most of the main points raised. In Section 2, we respond to each of the discussion comments separately, and then we provide a few concluding thoughts in Section 3.

2. Response to Discussant Comments

Our responses to each of the discussant comments are ordered alphabetically by lead author.

Comments from Drs. Cai and Tian. Drs. Cai and Tian suggest a nonparametric approach to estimate the value associated with a given dose rule $f(X)$, wherein they first estimate the conditional mean of R given $A = f(X)$ and $f(X)$ through a bivariate kernel-smoothed regression estimator then search for the optimal $f(X)$ maximizing this conditional mean. Alternatively, in our approach, we directly estimate the value function using $\hat{\mathcal{V}}_\phi(f)$, which was defined as

$$n^{-1}\sum_{i=1}^{n}\frac{R_i}{2\phi p(A_i|X_i)}\min\left\{\phi^{-1}|A_i - f(X_i)|, 1\right\}$$

so entailed only one level of approximation. Comparatively, Cai and Tian's method should provide a more accurate approximation to the value function due to a finer approximation at each subgroup level, where the subgroup is defined by treatment assignment and rule assignment; however, using bivariate kernel smoothing potentially increases the variability compared to a single kernel approximation. Numerically, when the dose rule of interest is quite far from the dose assignment, the denominator of $\hat{\mu}_a$ in their method is likely to be close to zero, leading to potentially unstable computation in the value estimation.

The two-stage regression method, as suggested by Drs. Cai and Tian, provides a straightforward way to estimate the IDR value function. However, such a method requires (1) that the user has a strong belief in the given regression model to truly explain the relationship between (A, X) and R, and (2) that the solution to the optimal treatment based on this regression model is computationally feasible. As indicated in our article, and in the numerical studies presented therein, because of (2), a quadratic function is often assumed for the regression model. Note that by Taylor expansion, LASSO_{M2} is very close to the true nonlinear model in Scenario 2, and hence performs well. However, a misspecified model, such as LASSO_{M1}, leads to the inferior performance in Scenario 2. On the other hand, it is not uncommon for one to wish to tailor the dose based on high-dimensional training data such as genomic data. The following example, given in Table 1, shows that LASSO_{M2} is less resilient in high-dimension than O-learning. Specifically, under a similar setting to Scenario 2 but with a higher dimension ($d = 100$), a modified O-learning method (using weighted reinforcement learning trees (Zhu, Zeng, and Kosorok 2015)) performs better than LASSO_{M2}.

Drs. Cai and Tian also presented an interesting idea to learn an ordered scoring system when the dose has K fixed levels, a discrete approximation to the continuous case. They first estimated a continuous score function $\hat{S}(X)$ via minimizing some surrogate L_2-loss function; then obtained a discrete dose rule by searching for cut-off points of $\hat{S}(X)$ which maximizes the empirical value function as much as possible. We note that this two-stage estimation approach can be reformulated into a single maximization problem:

$$\max\sum_{k=1}^{K}\sum_{i=1}^{n}R_i I(A_i = a_k)I(c_{k-1} < f(X_i) \le c_k)$$

subject to $c_0 = -\infty < c_1 < c_2 < \cdots < c_{K-1} < c_K = \infty$, which is equivalent to

$$\max\sum_{k=1}^{K}\sum_{i=1}^{n}R_i\left\{I(A_i = a_k) - I(A_i = a_{k-1})\right\}I(f(X_i) > c_{k-1})$$

subject to $c_1 < c_2 < \cdots < c_{K-1}$. It is easy to show that this optimization is equivalent to

$$\max\sum_{k=1}^{K}\sum_{i=1}^{n}R_i I(Z_{ik}(f(X_i) - c_k) > 0) \text{ subject to } c_0 = -\infty < c_1 < c_2 < \cdots < c_{K-1} < c_K = \infty,$$

CONTACT Michael R. Kosorok kosorok@unc.edu W. R. Kenan, Jr. Distinguished Professor, Department of Biostatistics, and Professor, Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC .9957

Table 1. Average value from 200 replicates for Scenario 2 with dimension $d = 100$.

n	LASSO$_{M2}$	O-learning
200	2.97 (0.25)	**3.08 (0.17)**
800	4.54 (0.24)	**4.81 (0.20)**

where $Z_{ik} = \{I(A_i = a_k) - I(A_i = a_{k-1})\}$. Hence, we may consider this optimization as a weighted classification problem with linear constraints. The latter restricts certain parameters to a cone. Thus, when using the hinge-loss as a surrogate loss, we can show that the dual problem remains a quadratic programming problem with linear constraints, which can be solved readily using existing software packages.

Comments from Drs. Fan and Yuan. Drs. Fan and Yuan suggest using a more smooth kernel approximation for the empirical value function rather than the triangle kernel we use in our article (which is a little different than the Mexican hat kernel). Their small simulation study indicates potential gains in terms of the value estimation. We certainly welcome this effort to further improve the value approximation; however, we caution that a price may need to be paid for computation when using a smoother kernel approximation. For example, using $H_4(z)$, the optimization problem with a linear dose rule becomes equivalent to optimizing a truncated polynomial of the sixth order. Thus, we will no longer enjoy the flexibility of the DC algorithm and quadratic programming implementation at each iteration of the algorithm as we were able to do in our article.

Regardless of which kernel function is used in the value approximation, one challenging persistent issue is the choice of the bandwidths. As indicated in our article, there is no gold standard for choosing the bandwidth empirically. What was suggested in our article was to choose a bandwidth for which further reduction of the bandwidth was unlikely to lead to a higher value empirically. The rationale behind this was that a smaller bandwidth should lead to a more accurate optimal treatment rule due to less bias in the approximation to the value function. It would be interesting to further explore the challenging issues of finding data-adaptive bandwidths and how the choice of the bandwidth depends on different choices of kernels.

Comments from Drs. Luedtke and van der Laan. Drs. Luedtke and van der Laan present a number of interesting results, only a few of which can we address here. To begin with, they show that faster convergence rates, even nearly n^{-1}, can be obtained using the plug-in method, that is, the optimal dose rule is first estimated using Q-learning and then this result is plugged into the value estimator (this approach is also sometimes called the "indirect" approach). Smoothness assumptions were required to obtain these results, including differentiability of the Q-learning functions and the value functions as well as a certain concavity property of the value function. As also discussed, one advantage of our proposed method is its robustness against possible model misspecification of the Q-learning model, but at the price of a potentially slower convergence rate. These comparisons between our method and the plug-in method are illuminated in the simulation studies presented in our article.

Robustness and efficiency are ever-present conflicting tradeoffs in statistical learning. Cross-validation has been suggested as a way of resolving this conflict for the super learner developed by Drs. Luedtke and van der Laan. However, as discussed in our article, this cross-validation approach may not be applicable in the continuous dose finding setting, because the probability of observing $A = f(X)$ for a given dose rule f is zero. Thus, some caution needs to be taken when using the super-learner method in this context, but this approach is certainly worth investigating further. One alternative technique which could also be explored is an analogue to the one-step Newton–Raphson method in semiparametric inference: one could start with our method to obtain an initial dose rule, which is expected to maintain robustness to model misspecification and would thus be consistent (but with a slow convergence rate); one could then apply the Q-learning method via a one-step Newton–Raphson iteration to solve the score equation from the Q-learning model. Under certain smoothness conditions, we expect that the latter could lead to a faster convergence rate, since, in the semiparametric context, this approach can result in an $n^{-1/2}$ convergence rate for parameter estimation even if the initial estimator has a convergence rate only slightly faster rate than $n^{-1/4}$.

Comments from Dr. Moodie. Dr. Moodie presents very interesting and insightful thoughts on how practical treatment allocations in observational studies, such as dosage levels in the Warfarin study, might carry useful information on optimal treatment strategies. Her simulations and analyses reveal a surprising finding that G-estimation without using outcome information can sometimes lead to nearly optimal treatment strategies.

It is well known that treatment allocations in observational studies are not random and, most likely, clinicians and doctors prescribe the treatments which they think are "most beneficial" to patients. However, such "most beneficial" treatments may or may not actually be the most beneficial. First, treatment prescription correctness is limited by the knowledge or experience of clinicians and doctors; for example, what is perceived as the best could be based on a limited list of drugs or doses which they have used in the past, or a limited number of patients they have seen recently, or an incomplete understanding of the underlying drug–disease mechanisms. Indeed, in Dr. Moodie's simulation studies, for the scenarios where the proportion of the patients actually treated optimally is not high, naive estimation of optimal treatment strategies by ignoring outcome information can lead to very biased results. Second, in observational studies, it is impossible to know what other significant factors influence which drugs patients actually take. They can be based on patient preference, cost of insurance claims, or even the influence of other patients or social media. Therefore, perception of the "most beneficial" treatment actually received in practice can be misleading.

On the other hand, in situations where the mechanism of treating diseases is largely understood, such as seems to be the case with the Warfarin study, one can believe that the actual treatment allocation is close to optimal. Such a belief can be incorporated into the analysis of estimating optimal treatment regimes, including in our dose finding method. A conceptual solution to this could be the following: we first start from the standard propensity score approach to control the observational nature of treatments and then apply either G-estimation or O-learning to learn the optimal treatment rule, say $\widehat{f}(X)$. We then refine the treatment propensity scores through some regularization so that the updated treatment propensity score is high if the actual treatment allocation is close to $\widehat{f}(X)$.

For example, suppose that we fit a logistic regression model to estimate the propensity of binary A given feature variables X. We can then consider the following regularized estimation:

$$\sum_{i=1}^{n}\left\{A_i\beta^T X_i - \log(1+e^{\beta^T X_i})\right\} + \lambda\sum_{i=1}^{n}(\beta^T X_i - \widehat{f}(X_i))^2.$$

Note that the second term in the above forces the updated propensity score estimates to be not far from the preliminary estimate of the optimal treatment strategy when λ is large. In other words, if we have a strong belief that the actual treatment allocation is close to optimal, we can use a large λ to incorporate this belief to refine the estimation of the propensity scores, which reduces the bias of the final optimal treatment estimation. Alternatively, instead of the two-step approach, we may consider directly including this belief in the value optimization. For example, we could replace the objective function (where we assume for this discussion that A_i is dichotomous) in O-learning by

$$\sum_{i=1}^{n} R_i I(A_i f(X_i) < 0)/\widehat{\pi}_i + \lambda\sum_{i=1}^{n} I(A_i f(X_i) \geq 0),$$

where $\widehat{\pi}_i$ is the estimated propensity score, and the second term forces the derived rule to be close to the actual treatment rule.

Finally, we appreciate the comment that we should be aware of the dangers of trying to evaluate or compare competing methods using observed treatment as a proxy for optimal treatment. The ideal way to do this would be through conducting a future, large-scale, and confirmatory study, while simultaneously seeking to understand the true underlying biological mechanisms.

Comments from Dr. Ogburn. Dr. Ogburn's discussion addresses practically important issues of generalizability and surrogate outcomes in the context of dynamic treatment regimes. She provides a complete review of recent work using S-admissibility in causal DAGs for evaluating generalizability and makes several useful suggestions for extending the definition of consistent surrogates to more complicated settings, including to continuous treatment dosing. Regarding this generalizability, we agree that when we really want to generalize the derived optimal treatment rule from population 1 to population 2, sampling selection should be independent of the potential outcomes in the two populations, and treatment assignments in the final group should be defined using all the effect modifiers. This can be achieved by including $P(S = 1|X, W)$ as part of the effect modifier, even though there may not exist such an effect modification involving the sample selection score.

Note that the covariate density discrepancy between the training (1) and target (2) population is also called the "data shift." The data drift problem can be alleviated by incorporating sampling weights (estimated from nonparametric methods such as kernel mean matching, (Gretton et al. 2013)) when building the model using the training data. The idea is that the weighted covariate density of the training data will mimic that of the target population. An important problem pointed out by Dr. Ogburn, which is relevant to this, is to identify a set of covariates that suffice to generalize the optimal treatment effects, which can be the ones predicting sample selection, treatments, and potential outcomes. Certain penalization methods could be used for this purpose.

For continuous dose selection, we believe that a consistent surrogate should be able to preserve ranking of the true outcome. In other words, $E[G|A = a, X] > E[G|A = a', X]$ if and only if $E[R|A = a, X] > E[R|A = a', X]$. Equivalently, there is a monotonic relationship between $E[R|A = a, X]$ and $E[G|A = a, X]$ for each group with the same covariate value X. For the Warfarin example, since the target is to have the INR remain within the range of 2–3, some transformed INR value may be used as a surrogate to maintain the above monotonic relationship. For example, we could define G to be the distance of the INR from the interval [2, 3].

Comments from Dr. Qian. The idea Dr. Qian proposes of using truncated L_2-loss instead of L_1-loss is interesting. It certainly leads to an advantage in computation due to its greater differentiability. As illustrated in each iteration of the DC algorithm in her comment, the L_2-loss there essentially leads to an update of the parameters which is similar in spirit to ridge regression, whereas the L_1-loss we use corresponds to a weighted least-absolute deviation estimator. Thus, we expect that the usual comparisons between ridge regression and LAD estimation could potentially be applicable here, and thus each method may have different trade-offs in terms of robustness and statistical efficiency.

The doubly robust loss proposed by Dr. Qian is an interesting and useful extension of our approach to the situation, where $p(a|X)$ must be learned from observational data. Dr. Qian has also outlined a nice framework for using doubly robust loss to learn the optimal dose rule. The main purpose of using the doubly robust loss is to ensure no bias in the value function; however, based on our experience, the trade-off is that the additional estimation of $p(a|X)$ and $Q(X, A)$, especially when both are misspecified, which is a possibility in practice, may result in large variability of the value estimator. Therefore, an interesting question is whether the doubly robust loss could be modified to reduce the mean squared error of the value estimator, taking into account both bias and variance.

Dr. Qian has also brought our attention to the practical situation, where dose assignments are sequential, either depending on the patient's on-going response or responses from previous patients. In this case, we could still potentially use our method provided we allowed $p(a|X)$ to depend on each subject as well as on timing of the administration of dose a, where X could include the patient's ongoing response and previous patient responses in the study. However, modeling $p(a|X)$ may be potentially difficult due to the dynamic nature of X for newly enrolled patients. Some additional model assumptions may also be needed; for example, we could assume that $p(a|X)$ only depends on the responses in a given time period in the past or on some summary quantities of previous patient outcomes.

Comments from Dr. Rosenblum. We agree that when the optimal rule is not unique for some subgroup of patients, one should always apply the dose with the least toxicity or side effects. Furthermore, when the clinical benefit of the derived treatment rule is less than a clinically meaningful threshold, the least toxic dose (usually the zero dose) should be used. Dr. Rosenblum also suggests to penalize the constant term in a linear dose rule so as

to favor the zero dose when there is no obvious clinical benefit. He also suggests to incorporate costs due to side effects, adverse events, and/or wasted healthcare resources as part of the outcome. Both ideas are interesting approaches to reducing impact of the nonuniqueness of the optimal rule. Alternatively, we could consider the following modification to our method: we first apply the dose regression model to estimate the dose–outcome relationship by including feature variables as effect modifiers; we then obtain the estimated optimal dose benefit relative to the zero dose, which we denote by $\hat{\delta}(X)$. Note that $\hat{\delta}(X)$ may be allowed to sometimes differ from the truth, but we require $\hat{\delta}(X)$ to be close to zero when the truth is actually zero (this can be achieved, e.g., by fitting a sparse linear model). Finally, we apply the proposed method but replace the penalty $\|f(X)\|$ by $\|f(X)/\hat{\delta}(X)\|$. It can be easily shown that this inverse benefit-weighted penalty leads to the same optimal dose rule as before whenever $\hat{\delta}(X)$ is strictly different from zero; however, when $\hat{\delta}(X) = 0$, this new penalty forces $f(X) = 0$, resulting in a zero dose for the subgroups who do not benefit from any nonzero dose level.

Inference for the assessment of the value and estimated rules has been a challenging and open problem in optimal treatment regime estimation, and more broadly, in machine learning-based methods. The main challenges include the complexity of treatment rules (which typically involve nonparametric estimation) and the boundary proximity problem. The latter particularly refers to the situation where some treatments are not distinguishable for a subgroup of patients. For binary treatments, it has been shown that standard inference approaches, including resampling, may not yield appropriate inference in this context. A number of attempts have recently been proposed to address this inference challenge, including penalized methods (Song et al. 2015) , m out of n resampling (Chakraborty, Laber, and Zhao 2013), adaptive but very conservative methods (Laber et al. 2014), and more recently data splitting in Luedtke and Van Der Laan (2016). However, it is unclear how these methods work in practical situations with low signal-to-noise ratio and high-dimensional feature variables. One thought is to consider treatment equivalence classes wherein subjects for whom the treatments are indistinguishable are grouped together and then this structure is used in formulating the inference. Interestingly, since treatment misallocation near the boundary does not really contribute to the overall expected value, empirical evidence suggests that incorrect inference due to this boundary issue may not affect the inference of the value function significantly.

In our proposed dose finding approach, the objective function is nonconvex so it is possible that a local minimum is attained at the end of the DC algorithm. In addition to using different initial values, including a one-size-fit-all estimate and the estimated rule from standard Q-learning, we can also consider evaluating the estimated rule compared to one based on dichotomizing treatment at a given dose level. The latter could also be computed using the O-learning method in Zhao et al. (2012) which is guaranteed to achieve the optimum due to convex optimization. When the value obtained from the personalized dose rule based on our proposed approach is larger than a sufficient number of the dichotomized rules, we would have increased confidence that a global optimum is achieved.

The article by Kennedy et al. (2016) proposes an interesting idea to estimate nonparametric dose effects based on the construction of doubly robust pseudo-outcomes. We believe a similar doubly robust procedure could be developed for the personalized dose finding setting by replacing their objective function with the value function we propose in our article. Furthermore, we acknowledge that doubly robust estimation has been well developed for G-estimation procedures by Robins (2004) and Moodie, Richardson, and Stephens (2007). More recently, our group has proposed a doubly robust O-learning method to infer optimal dynamic treatment regimes in multi-stage treatment settings, where the idea is to use the history of those patients who received nonoptimal treatments in future stages as auxiliary information so data can be augmented to include their information, instead of only using patients who are treated optimally in future stages as done in Zhao et al. (2015). We note, however, that although doubly robust estimation leads to unbiased estimation of optimal treatment strategies in a broader range of models, it may not necessarily lead to improved value estimation in terms of value gain and reduced variability. Thus, robust procedures focusing on value estimation would be a welcomed addition to personalized medicine research.

3. Conclusion

We once again express appreciation for the opportunity to participate in this discussion. Clearly, many important questions regarding personalized medicine methodology remain. In fact, overall, more questions were raised in this discussion than answered. Nevertheless, it is clear that as a discipline we now have the technical capacity to perform important, meaningful, and reproducible research which can advance personalized medicine and successfully find individualized treatment rules. Nevertheless, there remain many open questions and many opportunities for improvement. The challenges that remain include solving interesting and practically important problems in optimal estimation, computational efficiency, model robustness, valid statistical inference, causal inference, and in other areas. Taken together, the recent progress and interconnected research activity in personalized medicine study design and analysis is converging to become a new subdiscipline in statistics which is stimulating developments in theory and practice and on the interface with other disciplines. We look forward to watching this process continue to unfold and especially look forward to the accompanying improvements in human health.

References

Chakraborty, B., Laber, E. B., and Zhao, Y. (2013), "Inference for Optimal Dynamic Treatment Regimes Using an Adaptive m-Out-of-n Bootstrap Scheme," *Biometrics*, 69, 714–723. [1546]

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2013), *Chapter8: Covariate Shift by Kernel Mean Matching in Dataset Shift in Machine Learning*, Cambridge, MA: MIT Press. [1545]

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2016), "Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment Effects," *Journal of the Royal Statistical Society*, Series B, in press. [1546]

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225–1272. [1546]

Luedtke, A. R., and Van Der Laan, M. J. (2016), "Statistical Inference for the Mean Outcome Under a Possibly Non-Unique Optimal Treatment Strategy," *The Annals of Statistics*, 44, 713–742. [1546]

Moodie, E. E. M., Richardson, T. S., and Stephens, D. A. (2007), "Demystifying Optimal Dynamic Treatment Regimes," *Biometrics*, 63, 447–455. [1546]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics Analysis of Correlated Data*, New York: Springer, pp. 189–326. [1546]

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015), "Penalized Q-Learning for Dynamic Treatment Regimens," *Statistica Sinica*, 25, 901–920. [1546]

Zhao, Y., Zeng, D., Rush, J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [1546]

Zhao, Y.-Q., 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. [1546]

Zhu, R., Zeng, D., and Kosorok, M. R. (2015), "Reinforcement Learning Trees," *Journal of the American Statistical Association*, 110, 1770–1784. [1543]