Published in final edited form as:
J Am Stat Assoc. 2020 ; 115(529): 380–392. doi:10.1080/01621459.2018.1549050.

Matched Learning for Optimizing Individualized Treatment Strategies Using Electronic Health Records

Peng Wu [PhD candidate],
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032; (pw2394@cumc.columbia.edu)

Donglin Zeng [Professor],
Department of Biostatistics, University of North Carolina at Chapel Hill. (dzeng@email.unc.edu)

Yuanjia Wang[*] **[Professor]**
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032

Abstract

Current guidelines for treatment decision making largely rely on data from randomized controlled trials (RCTs) studying average treatment effects. They may be inadequate to make individualized treatment decisions in real-world settings. Large-scale electronic health records (EHR) provide opportunities to fulfill the goals of personalized medicine and learn individualized treatment rules (ITRs) depending on patient-specific characteristics from real-world patient data. In this work, we tackle challenges with EHRs and propose a machine learning approach based on matching (M-learning) to estimate optimal ITRs from EHRs. This new learning method performs matching instead of inverse probability weighting as commonly used in many existing methods for estimating ITRs to more accurately assess individuals' treatment responses to alternative treatments and alleviate confounding. Matching-based value functions are proposed to compare matched pairs under a unified framework, where various types of outcomes for measuring treatment response (including continuous, ordinal, and discrete outcomes) can easily be accommodated. We establish the Fisher consistency and convergence rate of M-learning. Through extensive simulation studies, we show that M-learning outperforms existing methods when propensity scores are misspecified or when unmeasured confounders are present in certain scenarios. Lastly, we apply M-learning to estimate optimal personalized second-line treatments for type 2 diabetes patients to achieve better glycemic control or reduce major complications using EHRs from New York Presbyterian Hospital.

Keywords

Personalized medicine; Individualized treatment rules; Matching; Observational studies; Machine learning

[*] yw2016@cumc.columbia.edu.

Supplementary Materials
The online supplementary materials contain additional simulations studies and EHR data analyses results.

1 Introduction

Personalized medicine calls for a paradigm shift from the universal strategy that assigns the same treatment to all patients affected by a disorder to selecting treatment strategies that optimize individual patient’s health outcomes according to individual characteristics (Collins & Varmus, 2015). Improvements in technologies for collecting personal data, accompanied with developments of machine learning and statistical methods to analyze these data, hold promise to enable healthcare providers to prescribe the right therapy to the right patient at the right time (Collins & Varmus, 2015; Chakraborty & Moodie, 2013). By treating each patient with the optimal individualized treatment, patients can potentially gain enhanced clinical benefits, experience less side effects, and be more adherent to treatments (Chakraborty & Moodie, 2013).

Machine learning approaches provide valuable tools to estimate individualized treatment rules (ITRs) and dynamic treatment rules (DTRs) due to their powerful computing capabilities. Previously proposed machine learning approaches include Q-learning (Watkins & Dayan, 1992; Qian & Murphy, 2011), outcome weighted learning (O-learning) (Zhao et al., 2012), boosting-based treatment selection (Kang et al., 2014), augmented O-learning (Liu et al., 2018, AOL), and subgroup identification methods (Fu et al., 2016). Most of these existing methods focus on analyzing randomized clinical trial (RCT) data. However, the ITRs estimated from RCTs may be inadequate to assist individualized treatment decision making in real-world settings due to stringent inclusion/exclusion criteria of RCTs, a lack of generalizability, and a lack of evidence for long-term outcomes.

Large-scale electronic health records (EHRs) provide new opportunities to learn ITRs using real-world patient data. In recent years, access to clinical data warehouses and databases continues to grow and an increasing trend of using EHRs for scientific research is observed (Weiskopf & Weng, 2013; Hripcsak & Albers, 2013; Hripcsak et al., 2016). As exclusive evidence generated from clinical trials is inadequate due to a lack of external validity, EHRs can serve as an important complement to evidence-based research for personalized medicine. For instance, a broad range of real-world medication use patterns not captured by RCTs were observed in EHRs (Hripcsak et al., 2016). Furthermore, as compared to RCTs, using EHRs to learn ITRs has benefits such as containing information on a large population over relatively longer time frames that reflects patients’ care management and disease course in more realistic settings.

However, EHRs are not collected for research purposes and conducting research with EHRs encounters great challenges. Critical issues including confounding bias and selection bias have been discussed (Hripcsak & Albers, 2013; Haneuse, 2016). In the context of estimating ITRs, common practice to adjust for confounding is inverse probability weighting (IPW) of propensity scores. The IPW approach requires a sophisticated model to estimate propensity scores with high accuracy. Machine learning methods are thus proposed to predict propensity scores (Lee et al., 2010, 2011; Austin & Stuart, 2015), but they may result in extreme weights with high variability. In addition, the IPW approaches may not adequately balance covariate distributions between treatment groups, especially when the distribution of propensity scores has less overlap between treatment arms (Crump et al., 2009).

On the other hand, matching has been successfully used to estimate population average treatment effects, including ratio matching (Smith, 1997), nearest neighbor matching (Dehejia & Wahba, 1999), and full matching (Stuart, 2010; Hansen, 2004). However, to the best of our knowledge, there is no method to leverage advantages of matching to estimate personalized treatment rules and apply to observational data such as EHRs. In this paper, we propose a machine learning approach, namely, Matched Learning (M-learning), to estimate ITRs through matching treated and untreated subjects with an application to EHRs. M-learning is a general framework that includes O-learning and AOL as special cases. M-learning introduces matching-based value function to match individual treatment responses under alternative treatments and alleviate confounding. Under a unified framework, an appropriate matching function can be used to compare outcomes for matched pairs to accommodate different types of data for measuring treatment response (continuous, discrete, or ordinal). The efficiency of M-learning can be improved by a de-noise procedure and doubly robust matching. The implementation is based on a matched-pairs weighted support vector machine. We establish the Fisher consistency and convergence rate of M-learning and conduct extensive simulation studies. We show that M-learning outperforms existing methods when propensity scores are misspecified and in certain scenarios when unmeasured confounders are present. Lastly, we tackle challenges of EHRs (e.g., confounding by indication, confounding bias, selection bias) and apply M-learning to estimate the optimal second-line treatments for type 2 diabetes (T2D) patients to achieve better glycemic control or reduce major complications using EHRs from New York Presbyterian Hospital.

2 Methodology

2.1 Individualized Treatment Rules (ITRs)

Let H_i denote the pre-treatment covariates and let A_i denote the binary treatment assignment taking values from $\{-1, 1\}$. Let R_i denote the clinical outcome post treatment (reward), and assume a larger R_i is more desirable (e.g., symptom reduction). An ITR is a decision rule, (H_i), that maps the domain of H_i to the treatment choices in $\{-1, 1\}$. The value function associated with $\mathscr{D}$ used to evaluate an ITR is defined as the expected post-treatment outcome by following $\mathscr{D}$ to assign treatments, that is, $V(\mathscr{D}) = E^{\mathscr{D}}(R_i)$.

For RCTs, the assumption that the potential outcomes are independent of treatment assignment given covariates is satisfied, and the treatment assignment probability, denoted by $\pi(a, h) = \Pr(A_i = a | H_i = h)$, is known by design. O-learning proceeds by re-expressing the value function as, $V(\mathscr{D}) = E\left[\frac{I(A_i = \mathscr{D}(H_i))R_i}{\pi(A_i, H_i)}\right]$, and then aims to maximize the empirical value function defined as

$$V_n(\mathscr{D}) = \frac{1}{n}\sum_{i=1}^{n}\frac{I(A_i = \mathscr{D}(H_i))R_i}{\pi(A_i, H_i)}. \qquad (1)$$

In an observational study, however, treatment propensities $\pi(A_i, H_i)$ are unknown and need to be estimated from data. Using the objective function (1) and IPW-based methods in observational studies suffer from instability and increased variance especially when weights are highly variable. In addition, IPW-based methods do not directly control the balance of covariate distributions between treatment groups.

2.2 Matched Learning (M-learning)

When comparing different treatment responses, matching methods can be designed to ensure balanced distribution at subgroup level and provide more flexible tools to control the matching quality of important confounders in subgroups or even on individual subjects. For example, covariates selection, distance metric and measure of covariates balance can be combined to optimize matching (Sekhon & Grieve, 2012) and identify matching subjects to guarantee numerical stability, especially when some subgroup of patients rarely receive one particular treatment. Denote the matched set for subject i as $\mathscr{M}_i$, which consists of subjects with opposite treatments but similar covariates as subject i, where similarity is defined under a suitable distance metric. That is, we let

$$\mathscr{M}_i = \left\{j: A_j = -A_i, d\left(H_j, H_i\right) \le \delta_i\right\},$$

where $d(\cdot,\cdot)$ is a metric defined in the covariate space and δ_i is a pre-specified positive threshold to determine the size of the matched set which may vary across subjects. For example, if we choose $\mathscr{M}_i$ to be the nearest neighbor, then δ_i is the minimal distance between subject i and any other subject with the opposite treatment. In some applications, subjects with empty matching sets may be excluded. In this paper, we use nearest neighbor in the matching step of M-learning in the simulations and application, and study its theoretical properties.

M-learning is developed to maximize a matching-based value function defined in (2). The motivation of M-learning is that when two subjects are matched in confounders or propensity scores of treatments but are observed to receive opposite treatments, the subject with a larger clinical outcome should be more likely to have received the optimal treatment among two options. Based on this rationale, one expects that if $j \in \mathscr{M}_i$ and $R_j \ge R_i$, then the optimal ITR for subject i should more likely to be A_j, and vice versa. Furthermore, the likelihood is expected to be greater if the difference between R_j and R_i is larger. Specifically, for any given ITR $\mathscr{D}$, define the matching-based value function as

$$\begin{aligned} V_n(\mathscr{D}; g) = n^{-1}\sum_{i=1}^{n} |\mathscr{M}_i|^{-1} \sum_{j \in \mathscr{M}_i} \Big\{ I\left(R_j \ge R_i, \mathscr{D}(H_i) = -A_i\right) \\ + I\left(R_j \le R_i, \mathscr{D}(H_i) = A_i\right)\Big\} g\left(\left|R_j - R_i\right|\right), \end{aligned} \quad (2)$$

where $|\mathscr{M}_i|$ is the size of $\mathscr{M}_i$ and $g(\cdot)$ is a monotonically increasing function specified by users to weight different pairs of subjects. Typical choices of $g(\cdot)$ can be $g(x) = 1$ or $g(x) = x$. Furthermore, let $\mathscr{D}(H) = \mathrm{sign}(f(H))$ for some ITR decision function f, then the matching-based value function (2) is equivalent to

$$V_n(f;g) = n^{-1}\sum_{i=1}^{n}|\mathscr{M}_i|^{-1}\sum_{j\in\mathscr{M}_i} I\big(f(H_i)A_i\mathrm{sign}\big(R_j - R_i\big) \le 0\big)g\big(|R_j - R_i|\big).$$

M-learning maximizes $V_n(f; g)$, or equivalently, minimizes

$$n^{-1}\sum_{i=1}^{n}|\mathscr{M}_i|^{-1}\sum_{j\in\mathscr{M}_i} I\big(f(H_i)A_i\mathrm{sign}\big(R_j - R_i\big) \ge 0\big)g\big(|R_j - R_i|\big), \tag{3}$$

in order to identify the optimal ITR.

The objective function (3) can be further expanded by allowing $\mathscr{M}_i = i$ (match subject i with himself/herself). If in addition we replace R_j in (3) by zero (when $R_j > 0$ for all subjects) or the smallest observed outcome when negative outcomes are present and choose $g(x) = x$, M-learning reduces to the original O-learning in Zhao et al. (2012). Similarly, if we replace R_j by subject i's predicted outcome estimated from a parametric model including only the main effects of H_i, M-learning reduces to the single-stage AOL in Liu et al. (2018). Thus, O-learning and single-stage AOL are special cases of M-learning, where they compare the observed outcome R_i with a constant or the predicted outcome given H_i averaged across treatments. In contrast, M-learning compares observed individual outcomes from two subjects in the matched set, where the treatment assignment is approximately “random” given H_i but the received treatments are opposite. Thus, M-learning is more informative in taking account of information on patient’s outcome at the individual level (R_i and R_j), instead of comparing a patient’s outcome with the predicted outcome averaged over treatments (as done in O-learning or AOL).

Minimizing the matching-based value function (3) is not feasible due to the discontinuity of the indicator function. Similar to O-learning, we replace the zero-one loss by other surrogate loss functions. In particular, when using the hinge-loss, the objective function to be optimized is the loss function for the weighted support vector machine (SVM) with matched pairs:

$$\begin{aligned} V_{n,\phi}(f;g) = {} & n^{-1}\sum_{i=1}^{n}|\mathscr{M}_i|^{-1}\sum_{j\in\mathscr{M}_i}\phi\big(-f(H_i)A_i\mathrm{sign}\big(R_j - R_i\big)\big)g\big(|R_j - R_i|\big) \\ & + \lambda_n\|f\|_{\mathscr{H}_K}, \end{aligned} \tag{4}$$

where $\phi(x) = (1 - x)_+$, λ_n is a tuning parameter and $\mathscr{H}_K$ is a reproducing kernel Hilbert space (RKHS) with kernel function $K(\cdot,\cdot)$. The solution to M-learning is obtained by minimizing $V_{n,\phi}(f;g)$. In terms of implementation, the dual problem of (4) is a quadratic problem which can be solved by any off-the-shelf quadratic programming packages.

Taking linear ITR decision rules as an example, we describe solution to the quadratic programming problem using Lagrange multipliers. Assume f in $V_{n,\phi}(f;g)$ is linear and $f(h) = \langle\beta, h\rangle + \beta_0$ where $\langle\,\cdot\,,\ \cdot\,\rangle$ denotes the inner product operator and $\|f\|_{\mathscr{H}_K}$ represents

$\|f\|^2$ in Euclidean space. It is computationally convenient to re-write (4) in an equivalent form as

$$\min \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\sum_{j \in \mathscr{M}_i} |\mathscr{M}_i|^{-1} g(|R_j - R_i|)\xi_{ij},$$

subject to: $A_i \text{sign}(R_i - R_j)(\langle \beta, H_i\rangle + \beta_0) \geq (1 - \xi_{ij}), \xi_{ij} \geq 0, \forall i$ and $j \in \mathscr{M}_i$, where ξ_{ij} is a slack variable that represents misclassification error for the jth subject in the matched set of the ith subject, C is a cost parameter, and $|\mathscr{M}_i|^{-1} g(|R_j - R_i|)$ is the individual-specific weight in a weighted SVM framework.

The Lagrange primal function follows as

$$\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{n}\sum_{j \in \mathscr{M}_i} |\mathscr{M}_i|^{-1} g(|R_j - R_i|)\xi_{ij}$$
$$-\sum_{i=1}^{n}\sum_{j \in \mathscr{M}_i} \alpha_{ij}\left\{A_i \text{sign}(R_i - R_j)\left(H_i^T\beta + \beta_0\right) - (1 - \xi_{ij})\right\} - \sum_{i=1}^{n}\sum_{j \in \mathscr{M}_i} \mu_{ij}\xi_{ij},$$

where we minimize with respect to β, β_0 and ξ_{ij}. By taking the respective derivatives and setting them to zero to obtain,

$$\begin{cases} \beta = \sum_{i=1}^{n} \sum_{j \in \mathscr{M}_i} \alpha_{ij} A_i \text{sign}(R_i - R_j) H_i, \\ 0 = \sum_{i=1}^{n} \sum_{j \in \mathscr{M}_i} \alpha_{ij} A_i \text{sign}(R_i - R_j), \\ \alpha_{ij} = C|\mathscr{M}_i|^{-1} g(|R_j - R_i|) - \mu_{ij}, \forall i \text{ and } j \in \mathscr{M}_i. \end{cases}$$

By substituting above equations into Lagrangian dual function, we obtain $\max \sum_{i=1}^{n} \sum_{j \in \mathscr{M}_i} \alpha_{ij} - \frac{1}{2}\sum_{i=1}^{n} \sum_{i'=1}^{n} \sum_{j \in \mathscr{M}_i} \sum_{j' \in \mathscr{M}_{i'}} \alpha_{ij}\alpha_{i'j'} A_i A_{i'}$ sign $(R_i - R_j)$ sign $(R_{i'} - R_{j'})\langle H_i, H_{i'}\rangle$ subject to $0 \leq \alpha_{ij} \leq C|\mathscr{M}_i|^{-1} g(|R_j - R_i|)$ and $\sum_{i=1}^{n} \sum_{j \in \mathscr{M}_i} \alpha_{ij} A_i \text{sign}(R_i - R_j) = 0$. In addition, subject to Karush-Kuhn-Tucker conditions for $\forall i$ and $j \in \mathscr{M}_i$ (Zhao et al., 2012):

$$\begin{cases} \alpha_{ij}\left[A_i \text{sign}(R_i - R_j)\left(H_i^T\beta + \beta_0\right) - (1 - \xi_{ij})\right] = 0, \\ \mu_{ij}\xi_{ij} = 0, \\ A_i \text{sign}(R_i - R_j)\left(H_i^T\beta + \beta_0\right) - (1 - \xi_{ij}) \geq 0, \end{cases}$$

the solution to the primal and dual problem is optimal. It is straightforward to extend the algorithm to other kernels (e.g., Gaussian kernel) and obtain a nonparametric ITR based on kernel function $K(\cdot,\cdot)$ in the RKHS.

2.3 Improved M-Learning

To improve the performance of M-learning, we use a de-noise procedure first reported in Liu et al. (2018). We replace R_i by a surrogate residualized outcome $\tilde{R}_i = R_i - s(H_i)$ in $V_n(\mathcal{D}; g)$ for any measurable function of H_i, denoted as $s(H_i)$. These residualized outcomes remove the main effects of covariates, which improves efficiency of identifying tailoring variables exhibiting quantitative or qualitative interaction with treatment. The residuals can be obtained through a regression model and the value function to be maximized becomes

$$\begin{aligned} V_n(\mathcal{D}; g) &= n^{-1} \sum_{i=1}^{n} |\mathcal{M}_i|^{-1} \sum_{j \in \mathcal{M}_i} \left\{ I\left(\tilde{R}_j \ge \tilde{R}_i, \mathcal{D}(H_i) = -A_i\right) \right. \\ &\quad \left. + I\left(\tilde{R}_j \le \tilde{R}_i, \mathcal{D}(H_i) = A_i\right) \right\} g\left(\left|\tilde{R}_j - \tilde{R}_i\right|\right). \end{aligned}$$

As shown in Liu et al. (2018), by removing the main effects of covariates, more stable weights are used in the weighted SVM to boost efficiency in estimating ITRs.

Furthermore, prognostic scores can be incorporated into M-learning under the framework of doubly robust matching estimator (DRME) proposed in Antonelli et al. (2018). The DRME uses both propensity scores and prognostic scores to construct a matching set $\mathcal{M}(i, \theta)$, where $\theta = (\theta_1, \theta_2)^T$ denotes parameters for the propensity score and prognostic score models:

$$\pi(H) = P(A = 1|H) = u_1\left(H^T \theta_1\right), m(H) = E(R|A = -1, H) = u_2\left(H^T \theta_2\right). \qquad (5)$$

Antonelli et al. (2018) showed that only one of the two models in (5) is required to be correctly specified to ensure consistency of DRME, which achieves double robustness. Applying DRME to M-learning, both propensity scores and prognostic scores will be included in the matching step to create informative matched pairs. The doubly robust M-learning is consistent even if one of the propensity score model or prognostic model is misspecified, and it will be more efficient than regular M-learning if both models are correctly specified. Note that M-learning can be applied to RCT data where only prognostic scores need to be included in the matching step to improve efficiency.

3 Theoretical Properties

In this section, we establish the theoretical properties including Fisher consistency, different choices of $g(x)$ and convergence rate of of M-learning.

3.1 Fisher Consistency

Theorem 3.1 *Under regularity assumptions including* $\max_{i=1}^{n} \delta_i \to 0$, *and that the density of* H *and* $E[R|H, A = 1]$ *is continuously differentiable in the support of H, it holds that*

$$V_n(f, g) \to_{a.s} V(f, g),$$

where

$$V(f;g) = E\left\{\widetilde{E}\left[I\left(f(H)A\text{sign}\left(\widetilde{R} - R\right) \le 0\right)g\left(\left|\widetilde{R} - R\right|\right)\middle|\widetilde{A} = -A, \widetilde{H} = H\right]\right\},$$

$\widetilde{E}$ is the expectation with respect to $(\widetilde{R}, \widetilde{H}, \widetilde{A})$, an independent copy of (R, H, A). In addition, define

$$\Delta_g(r,h) = E\left[\frac{g(|R-r|)}{|R-r|}(R-r)\middle|A = 1, H = h\right] - E\left[\frac{g(|R-r|)}{|R-r|}(R-r)\middle|A = -1, H = h\right],$$

then for any h in the support of H,

$$sign\left(f^*(h)\right) = sign\int_r \Delta_g(r,h)dF(r|H = h),$$

where $F(r|H = h)$ is the distribution of $R = r$ given $H = h$ and f^* is the optimal function minimizing $V(f, g)$.

The proof of Theorem 3.1 is given in the Appendix. Here we make a few remarks.

Remark 1. When $g(x) = x$ and $r = 0$, i.e. $\Delta_g(r, h) = E(R|A = 1, H = h) - E(R|A = -1, H = h)$, Theorem 3.1 implies that the optimal treatment rule obtained from M-learning is the same as the optimal rule from O-learning, and thus M-learning is Fisher consistent for the usual optimal ITR.

Remark 2. When $g(x) = 1$, we obtain

$$\begin{aligned} \Delta_g(r,h) &= E[\text{sign}(R-r)|A = 1, H = h] - E[\text{sign}(R-r)|A = -1, H = h] \\ &= 2[P(R > r|A = 1, H = h) - P(R > r|A = -1, H = h)]. \end{aligned}$$

Remark 2 suggests that for subjects with $H = h$, the optimal rule chooses the treatment with a higher probability of having a greater outcome than the average outcome across treatments. Such choice of $g(x)$ ensures robustness against outliers of R. When R is an ordinal or binary random variable, this choice is especially suitable. For example, consider an ordinal outcome with three levels, then the optimal rule $f^*(h)$ has a desirable property

$$\text{sign}\left(f^*(h)\right) = \text{sign}[\text{AUC}_{13}(h) - \text{AUC}_{23}(h)], \quad (6)$$

where $\text{AUC}_{jk}(h)$ is the conditional AUC for comparing $R = j$ with $R = k$ for subjects with $H = h$. More generally, the function $\Delta_g(r, h)$ is similar to creating comparisons based on a reference level r of the outcome. Therefore, for a particular target value r (e.g., the value under a universal "one-size-fits-all" treatment assignment, or a clinically meaningful level for an ordinal outcome), one can construct $g(x)$ so that the weights concentrate on the difference from the reference value r.

Remark 3. Lastly, when applied to observational studies, the condition of no unmeasured confounders ensures that the optimal rule estimates the treatment with a higher potential outcome, since $\Delta_g(r, h) = E\left(R^{(1)} | H = h\right) - E\left(R^{(-1)} | H = h\right)$, where $R^{(k)}$ denotes the potential outcome under treatment k.

3.2 Convergence Rate of M-Learning

In this section, we establish the convergence rate of the risk bound for the estimated decision rule. We consider the nearest neighborhood matching, $\mathscr{H}_K$ is the RKHS based on a Gaussian kernel function with bandwidth σ_n, and assume R and H are bounded. Furthermore, we need the following assumptions:

(A.1) The density of $H = h$ with respect to the dominating measure and $E(R|A = a, H = h)$ are continuously differentiable in H's support for $a = -1$ and 1. Moreover, the density of H is bounded from below on the support of H, denoted by $\mathscr{X}_H$.

(A.2) The probability measure has a geometric noise exponent $\alpha > 0$ as in Definition of Steinwart and Scovel (2007). That is, if let τ_H be the distance from any H to the decision boundary $\{h : f^*(h) = 0\}$, it holds

$$E\left[\left|f^*(H)\right|\exp\left\{-\tau_H^2/t\right\}\right] \leq ct^{\alpha d/2}, \quad t > 0.$$

(A.3) There exists $\gamma > 0$ and $r_0 > 0$ such that $|\mathscr{X}_H \cap B(h, r)| \geq \gamma |B(h, r)|$ for any $h \in \mathscr{X}_H$ and $0 < r < r_0$, where $B(h, r)$ is a ball centered at h with radius r, and $|A|$ denotes the volume of set A in $\mathscr{X}_H$.

Condition (A.1) is necessary to ensure the conistency of approximation in the nearest-neighbor based matching. Condition (A.2) is commonly assumed for SVMs and a similar condition has been considered for classification problem (c.f., Steinwart and Scovel (2007)) and establishing the learning rate for ITRs (Zhao et al., 2012). When the decision rule is completely separable, the exponent α can be as large as possible. The third condition (A.3) is used to obtain the convergence for the nearest-neighbor estimator (Devroye et al., 2013)

Theorem 3.2 *Under the above assumptions and letting* $\sigma_n = \lambda_n^{1/p(1+\alpha)}$, *it holds*

$$V\left(f^*; g\right) - V\left(\hat{f}; g\right) = O_p\left[\frac{1}{\sqrt{n}\lambda_n^{\beta_1}} + \frac{1}{\lambda_n^{\beta_2}}\left\{\left(\frac{m_n}{n}\right)^{1/p} + \sqrt{\frac{\log n}{m_n}}\right\} + \lambda_n^{\alpha/(1+\alpha)}\right],$$

where $\beta_1 = p/4 + (1/2 - p/8)d/[(1 + \alpha)]$, $\beta_2 = 1/2p(1 + \alpha) + 1/2$, *and* m_n *is the size of the nearest neighbor.*

The proof of Theorem 3.2 is given in the Appendix. Note that the convergence rate will depend on the dimension, the geometric noise exponent α and the choice of tuning parameter σ_n. Moreover, we observe that when $\lambda_n = n^{-\theta}$ with a constant θ and the size of nearest-neighbor equals to $n^{2/(p+2)}$, the polynomial convergence rate can be attained.

4 Simulation Studies

We conducted extensive simulation studies to compare M-learning with Q-learning and single-stage AOL as improved O-learning (Liu et al., 2018). Data were simulated under an observational study design where treatment assignment depends on pre-treatment variables H. Simulation settings and analyses we considered include: (1) No unmeasured confounder and the propensity score model given H is correctly specified in the analyses; (2) No unmeasured confounder but the propensity score model is misspecified; and (3) Unmeasured confounders are present and some components of H are not observed and not included in the analyses.

In these simulations, one-to-one matching with replacement was used and features were matched using shortest Euclidean distance function (one nearest-neighbor). The tuning parameters for AOL and M-learning (including choice of kernel as linear or Gaussian, inverse radius, and cost C) were selected by three-fold cross validation. The value function corresponding to the estimated optimal rule was computed on a large independent testing set with a sample size of 10, 000 using empirical average. Q-learning was fit with a linear model including feature variables and their interaction with treatment as covariates. We varied sample size of training data from 100 to 1, 000 and repeated the simulations 100 times.

We first considered continuous responses in two settings:

$$S_1 : R = 2H_3 - H_4 + A(H_1 - H_2) + 6\mathrm{sign}(H_1) + N(0, 1)$$

and

$$S_2 : R = 1 + 2H_1 + H_2 + 0.5H_3 + A\left(H_2 + H_1^2 - 1\right) + 6\mathrm{sign}(H_1) + N(0, 1) .$$

Uncorrelated feature variables H_k with standard normal distributions were simulated. Since heterogeneity and clustering effects are observed in the real-world patient population (e.g., Figure A3.2 of NYPH EHRs in Supplementary Materials), we considered the distribution of reward outcomes to be clustered in strata depending on the first feature variable H_1. The true optimal treatment decision boundary is linear in setting S_1, and nonlinear in setting S_2. The true optimal value is 1.20 in S_1 and 2.29 in S_2. In the continuous response scenario, $g(x) = x$ was used for M-learning. In setting S_1 and S_2, M-learning and doubly robust M-learning by stratifying on prognostic scores (referred to as “M-learning Stratified” in Figure 1 and 2) were considered. For the latter, prognostic scores were obtained using random forest. Prognostic factors used in the matching step were created by dichotomizing the prognostic scores based on the median split.

In the first set of simulations, distribution of A depends on H and no unmeasured confounder is present. Clinical response outcomes were simulated under setting S_1 and S_2, and the true propensity model was specified as $P(A = 1|H) = \mathrm{expit}(1 + 2H_1 + H_2)$. In this case, H_1 and H_2 are observed confounders. The propensity scores were estimated through a logistic regression model with treatment as binary outcome and features H_1, H_2 as linear predictors. On average, 64% of subjects received an active treatment and 36% received a control treatment. Simulation results are presented in the top panel of Figure 1. For setting S_1, Q-learning has the best performance since the linear function is the true optimal treatment separation boundary. Doubly robust M-learning performs similarly as Q-learning with larger sample size. It is clear that doubly robust M-learning improves efficiency. For S_2 with a nonlinear boundary, both M-learning and doubly robust M-learning achieve a higher empirical value than AOL and Q-learning. In this case Q-learning and AOL lose efficiency because they do not capture the information in prognostic scores, even though the propensity scores were consistently estimated.

In the second set of simulations, the true propensity score model was specified as $P(A = 1|H) = \text{expit}(1 + \exp(H_2))$. The propensity scores were estimated through a logistic regression model with linear predictors, and thus the model was misspecified. On average, 88% of subjects received one treatment and 12% received the other. Simulation results are presented in the bottom panel of Figure 1. In both setting S_1 and S_2, the results suggest that M-learning is more robust to misspecified propensity model compared to Q-learning and O-learning. The best performance is achieved by the doubly robust M-learning, where the estimated value function is very close to the true optimal value with a large sample size. Matching using prognostic scores in doubly robust M-learning has protected against deteriorated performance when the propensity score model is misspecified.

In the third set of simulations, we considered presence of unmeasured confounders.

The clinical outcomes were simulated as

$$S_3: R = 2H_3 - H_4 + A(H_1 - H_2 + X) + 6\text{sign}(H_1) + N(0, 1)$$

and

$$S_4: R = 1 + 2H_1 + H_2 + 0.5H_3 + A\left(H_2 + H_1^2 + X - 1\right) + 6\text{sign}(H_1) + N(0, 1)$$

where $P(A = 1|H, X) = \text{expit}(1 + R^{(-1)} - R^{(1)} + 2X + H_1)$ and X is an unmeasured confounder (not included in any analysis in any method) and $R^{(-1)}$, $R^{(1)}$ are potential outcomes under each treatment.

After introducing unmeasured confounding, the true optimal value function is 1.37 in S_3 and 2.61 in S_4. From Figure 2, we see that in S_3 with a linear decision boundary, Q-learning performs the best. Doubly robust M-learning has a higher mean value than M-learning. Matching-based methods have an advantage over AOL. Specifically, the value function of ITR estimated by AOL has a large variability, especially when the sample size is small. In S_4 with nonlinear decision boundary, two M-learnings much outperform AOL and Q-learning. In this case, the unmeasured confounder has a greater impact on AOL and Q-learning than M-learning.

We also examine M-learning with ordinal outcomes and report results in Supplementary Materials A1. For linear decision boundary, since ordinal outcomes were generated by discretizing a continuous outcome, M-learning does not give an advantage over Q-learning and AOL. For nonlinear boundary, M-learning using matching function $g(x) = 1$ and $g(x) = x$ both achieves a higher value than Q-learning and AOL.

5 Application to EHRs to Learn Optimal Treatment Sequence for T2D patients

We apply various methods to a large clinical data warehouse (CDW) at New York Presbyterian Hospital (NYPH). NYPH CDW is one of the earliest pioneer CDWs in the United States developed 25 years ago, long before the wide adoption of EHRs and informatics methods. The database encompasses about 4.5 million patients in the New York City population, making it a useful data source for research and supports new research initiatives including eMERGE (Gottesman et al., 2013) and precision medicine initiative. The details of the informatics technology of NYPH CDW is described in Section A2 of Supplementary Materials.

Our research goal is to optimize treatment sequence for T2D patients based on their person-specific characteristics. Current treatment guideline recommends metformin (MET) as the first line treatment for T2D patients (Diabetes Control and Complications Trial Research Group, 1993). Literature reveals barriers of timely insulin initiation in clinical practice when patients do not achieve adequate glycemic control by using metformin alone, and the optimal sequence of treatments for insulin therapy versus second-line oral hypoglycemic agents (OHA) largely remains unknown (American Diabetes Association, 2014). In this work, we aim to estimate the optimal second-line treatment for T2D patients who received MET as the first-line treatment using real-world EHRs. Targeting the second-line treatments (metformin + insulin versus metformin + SFU, where SFU refers to oral agent sulfonylureas that includes glyburide and glipizide) partially reduces confounding by indication, where treatment uncertainty is present in real-world practice.

We excluded subjects with extreme baseline HbA1c values (greater than 10%), and used a new-user cohort design (Ray, 2003). Such design is often used in other studies of EHRs to properly capture time-varying confounding and early treatment responses. Specifically, the study design is illustrated in Figure 3. Subjects who started a second-line treatment (new users) are anchored at the treatment initiation (index date), and information before and after index date will be analyzed. Subjects were included in the analyses if they had MET as the first-line treatment, had insulin or SFU as the second-line treatment, and had at least one observation post index date. The median baseline period was around one year and the median follow up time post second-line treatment was about 18 months.

In Section A2 of Supplementary Materials, we describe details of patient records extraction and feature extraction. We constructed patterns of laboratory measurements to handle challenges in the analyses of EHRs (e.g., confounding bias and selection bias). Extracted features encompass information from five domains (Figure A3.1 of Supplementary Materials): demographics, medication prescription, ICD diagnosis codes, laboratory test values, and lab test measurement patterns. Propensity scores were estimated using two distinct logistic regression models for lab measurement pattern features and demographics covariates. The matching step in M-learning was performed using extracted features from lab test values, ICD counts, and two propensity scores. In addition, to improve efficiency and perform doubly robust matching, we also included a prognostic score estimated from a linear regression model in the matching step. Mahalanobis distance was the matching similarity measure and one nearest-neighbor was used to select matched pairs. To address selection bias in missing post-treatment outcomes, we used the IPW method and constructed a logistic regression model predicting whether a subject had any post treatment lab measure to compute the weights. To handle incompleteness in features, imputation with chained equations was used (Buuren & Groothuis-Oudshoorn, 2011).

Our final EHR data for learning optimal ITR consists of 740 patients, among whom 292 (39%) received insulin as the second-line treatment while 448 (61%) received SFU. The outcome is the HbA1c level (%) at 6 month post second-line treatment initiation estimated from a linear mixed effect model with subject-specific random intercepts and random slopes. Feature variables for learning optimal ITR include initial lab test values (HbA1c, glucose, HDL, LDL, BMI) and rate of change of measurements before index date, demographic variables, the cluster membership estimated from a subset of features (Online Supplementary Materials, Section A2, Figure A3.2), counts of other non-glycemic medications and counts of positive ICD diagnosis codes. Two-fold cross validation was used to estimate the value function of fitted ITRs.

We divided our cohort to two groups according to the initial HbA1c level (high baseline HbA1c group: >= 8.5 and low baseline HbA1c group: < 8.5) and analyzed the groups separately to further reduce patient heterogeneity. We compared the cross-validated value function of doubly robust M-learning to non-personalized universal rules, Q-learning, and AOL. In the rest of this section, we refer doubly robust M-learning as M-learning and AOL as O-learning for simplicity. The results are displayed in Figure 4 and Table 1. In the low baseline group, there were 380 patients in total (240 received SFU, 140 received insulin). For universal rules, the IPW-adjusted mean HbA1c level is 7.99 for those treated by SFU and is 8.05 for insulin. M-learning achieves the best glycemic control among all methods (lowest post-treatment HbA1c at 6 month) with a median and mean of 7.85 that is much lower than both universal rules. Q-learning does not provide much improvement compared to universal rules and its estimated post-treatment HbA1c is slightly smaller than assigning SFU to all. In the high baseline group, there were 152 patients who received insulin and 208 received SFU. The universal rules for HbA1c level in SFU group is 8.90 and in insulin group is 9.21. O-learning and M-learning have very similar performance and both reduce the average post-treatment HbA1c level to 8.57, again much lower than universal rules.

By examining M-learning in all patients using a linear kernel in the low baseline group, we identified several features that are most informative in determining the optimal treatment: pre-treatment rate of change of BMI, initial value of glucose and LDL at the index date, co-medication count, patient cluster membership and race. These feature variables can be considered by healthcare practitioner when recommending second-line treatment for T2D patients. There were 263 (69%) of the 380 patients predicted to have “MET + SFU” as the optimal choice and 117 (31%) with “MET + Insulin” as the optimal choice. Of the 240 patients who were prescribed SFU as the second-line treatment, majority of times (66%) medication was also the predicted optimal treatment in terms of lowering HbA1c level. In contrast, among the 140 patients who were prescribed insulin, only 36 (26%) were optimal. In the high baseline group, the important features we identified are initial value of HDL, age, sex and patient cluster membership. 294 (82%) of the 360 patients were recommended to “MET + SFU”. Of the 208 patients who were prescribed SFU, 168 (81%) also had as the predicted optimal treatment. Among the 152 patients who received insulin treatment, only 26 (17%) were optimal. These results seem to suggest that some patients who received insulin as the second-line treatment might be better treated with SFU.

However, Bianchi and Del Prato (2011) suggested that tight glycemic control need to be studied carefully in different group of T2D patients to determine the balance of its negative and positive effect and treatment personalization should be recommended considering multiple factors such as risk of complications (e.g. cardiovascular events). Given a low rate of insulin predicted to be optimal among patients who were treated with insulin, we explored whether insulin could be prescribed based on other considerations such as risk of complications in addition to achieving glycemic control. We estimated the optimal ITR that reduces major complications of T2D measured by three ICD diagnosis counts including essential hypertension, hyperlipidemia and hypercholesterolemia as ordinal outcomes (0, 1, 2, 3). M-learning was implemented with $g(x) = x$. The results are displayed in Figure 5 and Table 2. In the low baseline group, O-learning is moderately better than M-learning with an average count of 0.72. Based on M-learning, SFU was predicted to be optimal for 274 (72%) patients. Among patients who indeed received SFU, 175 (73%) were predicted to be optimal with regarding to reducing complications while 41 (29%) of the patients who received insulin were predicted to be optimal. In the high baseline group, M-learning performs the best with an average value of 0.84. Further investigation shows that insulin was predicted to be the optimal choice for 234 (65%) patients. In this group, among 152 patients who indeed received insulin, 106 (70%) were predicted to be optimal with regard to reducing complications, while only 80 (38%) of the patients who received SFU were predicted to be optimal.

In conclusion, the optimal ITRs outperform universal rules in all groups for both outcomes. M-learning performs better than Q-learning in all cases and better than O-learning in most cases. In addition, the proportion of patients treated by insulin and with insulin predicted to be optimal is higher when considering reducing complications as the outcome as compared to controlling for HbA1c (from 17% to 70% in the high baseline group). This result suggests that the rationale to prescribe insulin might be also based on concerns of complications especially when the baseline HbA1c is high (greater than 8.5%).

6 Discussion

We have proposed a machine learning approach based on matching, M-learning, to estimate the optimal ITR from observational data. We show that M-learning is a general approach that includes O-learning and some of its derivatives as special case and it satisfies Fisher consistency. A general matching function is proposed to analyze continuous or discrete outcomes where in some cases the objective function maximizes a certain function of AUC. The choice of $g(\cdot)$ function provides a flexible tool to weight outcome measures: $g(x) = 1$ gives the most robust estimation which only concerns with the ranking of outcomes; while other robust choices can prevent sensitivity to outliers of R_i's. Moreover, multivariate outcomes can be incorporated in the M-learning framework by creating suitable g function. The matching function $g(x)$ can be selected from a pool of non-decreasing functions to estimate the optimal ITRs in a data-driven way, which may lead to a better post treatment response.

M-learning has a few advantages over O-learning or other IPW-based methods. It does not rely on the validity of propensity score models and no inverse weighting is involved. Thus instability can be avoided when there are extremely small weights. The choice of $\mathscr{M}_i$ in M-learning is flexible and can include a large suite of matching tools including nearest neighbor, metrics defined on a dimension-reduced space determined by propensity scores or prognostic scores, yielding double robustness. For example, methods based on greedy matching or optimal matching algorithm are available to be implemented in M-learning. Different calipers can also be specified for individual subject and hence allow more "personalization". This strategy will introduce more flexibility but at the price of some computational complexity.

The choice of matching variables is important in M-learning. The performance of M-learning may be affected by the presence of high-dimensional features in the matching step. We suggest a dimension reduction approach to match on a lower dimensional space consisting of propensity score, prognostic score, and/or cluster membership of patients. We also included some key covariates as part of the matching criteria. A more general practical guide during the matching step is: first, choose major confounders according to domain knowledge or preliminary studies to achieve covariates balance; second, construct several propensity scores to reduce the dimensionality of the space of matching covariates; and third, include prognostic scores in order to improve robustness and efficiency. In the EHR analysis here, we considered this general guideline and constructed domain-wise propensity scores as well as prognostic scores, and matching was performed based on these scores. Other variable selection techniques can be considered, for example, to estimate propensity and prognostic scores by penalized regression.

Single-stage M-learning can be generalized to multi-stage setting by changing the value function $V(\mathscr{D})$ to a corresponding matching-based value function involving multiple stages and applying the backward learning methods (Liu et al., 2018). In each stage, M-learning will have the flexibility to choose different matching function and matched features. Furthermore, an extension to handle efficacy and safety outcomes (e.g., glycemic control and risk of complications) simultaneously when learning ITR is desirable. Here we only considered choosing between two treatment options. M-learning is ready to be generalized to more than two treatments by, for example, adopting one-versus-one or one-versus-all strategies for multicategory learning (Allwein et al., 2001). Lastly, our analyses were restricted to EHRs from those who had at least one second-line T2D treatment documented at a single academic medical center. It would be of interest to examine the performance of our methods on other EHR databases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research is support by U.S. NIH grants NS073671, GM124104, MH117458.

Appendix

Proof of Theorem 3.1. After some algebra, we can show that the value function is equal to

$$\begin{aligned} E\Bigg[& I(f(H) > 0)\Bigg\{ \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{+}|\widetilde{A} = 1, \widetilde{H} = H] \\ & + \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{-}|\widetilde{A} = -1, \widetilde{H} = H]\Bigg\}\Bigg] \\ + E\Bigg[& I(f(H) \le 0)\Bigg\{ \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{+}|\widetilde{A} = -1, \widetilde{H} = H] \\ & + \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{-}|\widetilde{A} = 1, \widetilde{H} = H]\Bigg\}\Bigg]. \end{aligned}$$

Hence, the optimal decision function, denoted by $f^{*}(H)$, should have the same sign as

$$\begin{aligned} & E\Bigg[\Bigg\{ \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{+}|\widetilde{A} = 1, \widetilde{H} = H] \\ & \quad + \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{-}|\widetilde{A} = -1, \widetilde{H} = H]\Bigg\}|H\Bigg] \\ & - E\Bigg[\Bigg\{ \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{+}|\widetilde{A} = -1, \widetilde{H} = H] \\ & \quad + \widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)_{-}|\widetilde{A} = 1, \widetilde{H} = H]\Bigg\}|H\Bigg] \\ & = E\Bigg[\widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)|\widetilde{A} = 1, \widetilde{H} = H]|H\Bigg] \\ & \quad - E\Bigg[\widetilde{E}[\frac{g(|\widetilde{R} - R|)}{|\widetilde{R} - R|}(\widetilde{R} - R)|\widetilde{A} = -1, \widetilde{H} = H]|H\Bigg]. \end{aligned}$$

In other words, if we define

$$\Delta_{g}(r, h) = E\left[\frac{g(|R - r|)}{|R - r|}(R - r)|A = 1, H = h\right] - E\left[\frac{g(|R - r|)}{|R - r|}(R - r)|A = -1, H = h\right],$$

then for any h in the support of H,

$$\operatorname{sign}\left(f^{*}(h)\right) = \operatorname{sign}\int_{r} \Delta_{g}(r, h) dF(r|H = h),$$

where $F(r|H = h)$ is the distribution of $R = r$ given $H = h$. □

Proof of Theorem 3.2. For convenience of notation, we use $\|\cdot\|_{n}$ to denote the norm in the RKHS and omit g in the definition of the loss function, i.e., denote $L_{n}(f; g)$ as $L_{n}(f)$. We use

c to denote a constant that is independent of n in the following proof. The M-learning algorithm estimates the decision function f as $\widehat{f}$ that minimizes (4), which can be rewritten as

$$L_n(f) \equiv \mathbf{P}_n\left[\frac{\int I\left(d\left(\widetilde{H}, H\right) < \delta_n\right)\phi\left(-f(H)A\text{sign}\left(\widetilde{R} - R\right)\right)g\left(\left|\widetilde{R} - R\right|\right)d\widetilde{\mathbf{P}}_n}{\int I\left(d\left(\widetilde{H}, H\right) < \delta_n\right)d\widetilde{\mathbf{P}}_n}\right] + \lambda_n\|f\|_n^2$$

Here, $\widetilde{\mathbf{P}}_n$ and $\widetilde{\mathbf{P}}$ to be used later refer to the measures with respect to an independent copy of random variables, $(\widetilde{R}, \widetilde{A}, \widetilde{H})$. We further define

$$Q_n(R, A, H; f) = \frac{\int I\left(d\left(\widetilde{H}, H\right) < \delta_n\right)\phi\left(-f(H)A\text{sign}\left(\widetilde{R} - R\right)\right)g\left(\left|\widetilde{R} - R\right|\right)d\widetilde{\mathbf{P}}_n}{\int I\left(d\left(\widetilde{H}, H\right) < \delta_n\right)d\widetilde{\mathbf{P}}_n}$$

and

$$Q(R, A, H; f) = \widetilde{E}\left[\phi\left(-f(H)A\text{sign}\left(\widetilde{R} - R\right)\right)g\left(\left|\widetilde{R} - R\right|\right)\middle|\widetilde{H} = H\right] .$$

Clearly, $L_n(f) = \mathbf{P}_n Q_n(R, A \,.\, H; f) + \lambda_n\|f\|_n^2$.

Let $L_\phi(f) = E[Q(R, A, H; f)]$. From the general property of the weighted hinge-loss as shown in Theorem 3.2 of Zhao et al. (2012), we have

$$V(f^*; g) - V\left(\widehat{f}; g\right) \le c\left\{L_\phi\left(\widehat{f}\right) - L_\phi(f^*)\right\}.$$

Therefore, it is sufficient to obtain a bound for the right-hand side. First, since $L_{\phi n}\left(\widehat{f}\right) \le L_{\phi n}(0)$, we obtain $\lambda_n\left\|\widehat{f}\right\|_n^2 \le 1$. Let f_{0n} be the minimizer of $L_\phi(f) + \lambda_n\|f\|_n^2$ over $f \in \mathscr{H}_K$ Therefore,

$$\begin{aligned}
& L_\phi\left(\widehat{f}\right) - L_\phi(f^*) \\
\le\ & E\left[Q\left(R, A, H; \widehat{f}\right)\right] - E[Q(R, A, H; f_{0n})] + E[Q(R, A, H; f_{0n})] - V(f^*) \\
\le\ & -(P_n - P)\left[Q\left(R, A, H; \widehat{f}\right) - Q(R, A, H; f_{0n})\right] \\
& + P_n\left[Q\left(R, A, H; \widehat{f}\right)\right] - P_n[Q(R, A, H; f_{0n})] \\
& + E[Q(R, A, H; f_{0n})] - V(f^*) \\
\le\ & \sup_{f: \|f\|_n \le \lambda_n^{-1/2}} |(P_n - P)Q(R, A, H; f)| \\
& + P_n\left[Q\left(R, A, H; \widehat{f}\right) - Q_n\left(R, A, H; \widehat{f}\right)\right] - P_n[Q(R, A, H; f_{0n}) - Q_n(R, A, H; f_{0n})] \\
& + L_n\left(\widehat{f}\right) - \lambda_n\left\|\widehat{f}\right\|_n^2 - L_n(f_{0n}) \\
& + E[Q(R, A, H; f_{0n})] + \lambda_n\|f_{0n}\|_n^2 - V(f^*)
\end{aligned}$$

$$\le \sup_{f: \|f\|_n \le \lambda_n^{-1/2}} |(P_n - P)Q(R, A, H; f)| \qquad \text{(I)}$$

$$+\sup_{R,A,H}\left|Q\left(R,A,H;\hat{f}\right)-Q_n\left(R,A,H;\hat{f}\right)\right| \tag{II}$$

$$+\sup_{R,A,H}\left|Q(R,A,H;f_{0n})-Q_n(R,A,H;f_{0n})\right| \tag{III}$$

$$+E[Q(R,A,H;f_{0n})]+\lambda_n\|f_{0n}\|_n^2-V(f^*) \tag{IV}$$

We refer the terms in the right-hand side as (I), (II), (III) and (IV) in turn.

For term (I), we compute the bracket covering number of some finite balls in $\mathscr{H}_K$. First, from Theorem 3.1 in Steinwart and Scovel (2007), the entropy number for the unit ball in $\mathscr{H}_K$, denoted by $\mathscr{O}_n$, satisfies

$$\log\mathscr{N}\left(\epsilon,\mathscr{O}_n,\|\cdot\|_\infty\right)\le c\sigma_n^{-(1-p/4)d}\epsilon^{-p}$$

for a constant c depending on p and d, so it yields

$$\log\mathscr{N}_{[]}\left(\epsilon,\mathscr{O}_n,\|\cdot\|L^2(P)\right)\le c\sigma_n^{-(1-p/4)d}\epsilon^{-p}.$$

Thus, we obtain

$$\log\mathscr{N}_{[]}\left(\epsilon,\left\{f:f\in\mathscr{H}_{\sigma_n},\|f\|_n\le\lambda_n^{-1/2}\right)\right\},\|\cdot\|L^2(P)\right)\le c\sigma_n^{-(1-p/4)d}\epsilon^{-p}(1/\lambda_n)^{p/2}.$$

Note that $Q(R, A, H; f)$ is Lipschitz continuous with respect to f in the sense that

$$|Q(R,A,H;f_1)-Q(R,A,H;f_2)|\le c|f_1(H)-f_2(H)|,$$

where c is a constant bounding $g\left(\left|R-\tilde{R}\right|\right)$. Therefore, we obtain

$$\log\mathscr{N}_{[]}\left(\epsilon,\left\{Q(R,A,H;f):\|f\|_n\le\lambda_n^{-1/2}\right\},\|\cdot\|L^2(P)\right)\le c\sigma_n^{-(1-p/4)d}\epsilon^{-p}/\lambda_n^{p/2}.$$

According to Theorem 2.14.2 in Van Der Vaart and Wellner (1996), we obtain that term (I) is bounded by

$$\begin{aligned}&O_p(1)\left\{n^{-1/2}\int_0^c\sqrt{1+\log\mathscr{N}_{[]}\left(\epsilon\left\{Q(R,A,H;f):\|f\|_n\le\lambda_n^{-1/2}\right\},\|\cdot\|L^2(P)\right)d\epsilon}\right\}\\&=O_p(1)n^{-1/2}\sigma_n^{-(1/2-p/8)d}/\lambda_n^{p/4}.\end{aligned}$$

For term (II), since $\|\hat{f}\|_n\le\lambda_n^{-1/2}$, Theorem 4.48 in Steinwart and Christmann (2008), implies that $\hat{f}$ is differentiable with derivative bounded by $c\sigma_n^{-1}\|\hat{f}\|_n=c\sigma_n^{-1/2}\lambda_n^{-1/2}$. Using the uniform convergence rate result for nearest-neighbor estimators (Devroye et al., 2013; Jiang, 2017) and assumptions (A.1)–(A.3), we obtain that term (II) is bounded by

$$O_p(1)(\sigma_n\lambda_n)^{-1/2}\left[\left(\frac{m_n}{n}\right)^{1/p} + \sqrt{\frac{p\log n}{m_n}}\right].$$

The same bound holds for term (III). Finally, the last term is the approximation error as defined in Steinwart and Christmann (2008) but with a different definition of the loss function as $Q(R, A, H; f)$. We can follow exactly the same argument in Theorem 2.7 of Steinwart and Christmann (2008) to obtain its upper bound as $c(\sigma_n^{-p}\lambda_n + \sigma_n^{\alpha p})$ for any positive α.

In conclusion, we have shown

$$\begin{aligned} &L_\phi(\hat{f}) - L_\phi(f^*) \\ &\le O_p(1)\left[\frac{n^{-1/2}}{\sigma_n^{(1/2 - p/8)d}\lambda_n^{p/4}} + (\sigma_n\lambda_n)^{-1/2}\left\{\left(\frac{m_n}{n}\right)^{1/p} + \sqrt{\frac{\log n}{m_n}}\right\} + \sigma_n^{-p}\lambda_n + \sigma_n^{\alpha p}\right]. \end{aligned}$$

By choosing $\sigma_n = \lambda_n^{1/p(1+\alpha)}$, we obtain the result in Theorem 2.

As a remark, the tail probability, $P(|V(f^*; g) - V(\hat{f}; g)| \ge t)$ where $t > 0$, can also be obtained under similar arguments. Theorem 3.2 provides a stochastic bound for term (I) in the Appendix. One can obtain the bound of the tail probability for this term using the tail bound for empirical processes (Chapter 2.14, Van Der Vaart and Wellner (1996)). Then the tail probability, $P(|V(f^*; g) - V(\hat{f}; g)| \ge t)$, will follow. □

References

Allwein EL, Schapire RE, & Singer Y (2001). Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research, 1, 113–141.

American Diabetes Association. (2014). Standards of medical care in diabetes—2014. Diabetes Care, 37 (Supplement 1), S14–S80. [PubMed: 24357209]

Antonelli J, Cefalu M, Palmer N, & Agniel D (2018). Doubly robust matching estimators for high dimensional confounding adjustment. Biometrics

Austin PC, & Stuart EA (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in Medicine, 34 (28), 3661–3679. [PubMed: 26238958]

Bianchi C, & Del Prato S (2011). Metabolic memory and individual treatment aims in type 2 diabetes – outcome-lessons learned from large clinical trials. The Review of Diabetic Studies, 8 (3), 432–440. [PubMed: 22262079]

Buuren S, & Groothuis-Oudshoorn K (2011). MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45 (3).

Chakraborty B, & Moodie EE (2013). Statistical methods for dynamic treatment regimes New York: Springer-Verlag.

Collins FS, & Varmus H (2015). A new initiative on precision medicine. New England Journal of Medicine, 372 (9), 793–795. [PubMed: 25635347]

Crump RK, Hotz VJ, Imbens GW, & Mitnik OA (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96 (1), 187–199.

Dehejia RH, & Wahba S (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association, 94 (448), 1053–1062.

Devroye L, Gÿorfi L, & Lugosi G (2013). A probabilistic theory of pattern recognition (Vol. 31) New York: Springer-Verlag.

Diabetes Control and Complications Trial Research Group. (1993). The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. New England Journal of Medicine, 329 (14), 977–986. [PubMed: 8366922]

Fu H, Zhou J, & Faries DE (2016). Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. Statistics in Medicine, 35 (19), 3285–3302. [PubMed: 26892174]

Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, … et al. (2013). The electronic medical records and genomics (eMERGE) network: past, present, and future. Genetics in Medicine, 15 (10), 761–771. [PubMed: 23743551]

Haneuse S (2016). Distinguishing selection bias and confounding bias in comparative effectiveness research. Medical Care, 54 (4), e23–e29. [PubMed: 24309675]

Hansen BB (2004). Full matching in an observational study of coaching for the SAT. Journal of the American Statistical Association, 99 (467), 609–618.

Hripcsak G, & Albers DJ (2013). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20 (1), 117–121. [PubMed: 22955496]

Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, … et al. (2016). Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences, 113 (27), 7329–7336.

Jiang H (2017). Rates of uniform consistency for k-NN regression. arXiv preprint arXiv:1707.06261.

Kang C, Janes H, & Huang Y (2014). Combining biomarkers to optimize patient treatment recommendations. Biometrics, 70 (3), 695–707. [PubMed: 24889663]

Lee BK, Lessler J, & Stuart EA (2010). Improving propensity score weighting using machine learning. Statistics in Medicine, 29 (3), 337–346. [PubMed: 19960510]

Lee BK, Lessler J, & Stuart EA (2011). Weight trimming and propensity score weighting. PloS One, 6 (3), e18174. [PubMed: 21483818]

Liu Y, Wang Y, Kosorok MR, Zhao Y, & Zeng D (2018). Augmented outcome-weighted learning for estimating optimal dynamic treatment regimens. Statistics in Medicine, 37, 3776–3788. [PubMed: 29873099]

Qian M, & Murphy SA (2011). Performance guarantees for individualized treatment rules. The Annals of Statistics, 39 (2), 1180. [PubMed: 21666835]

Ray WA (2003). Evaluating medication effects outside of clinical trials: new-user designs. American Journal of Epidemiology, 158 (9), 915–920. [PubMed: 14585769]

Sekhon JS, & Grieve RD (2012). A matching method for improving covariate balance in cost-effectiveness analyses. Health Economics, 21 (6), 695–714. [PubMed: 21633989]

Smith HL (1997). Matching with multiple controls to estimate treatment effects in observational studies. Sociological Methodology, 27 (1), 325–353.

Steinwart I, & Christmann A (2008). Support vector machines New York: Springer-Verlag.

Steinwart I, & Scovel C (2007). Fast rates for support vector machines using gaussian kernels. The Annals of Statistics, 575–607.

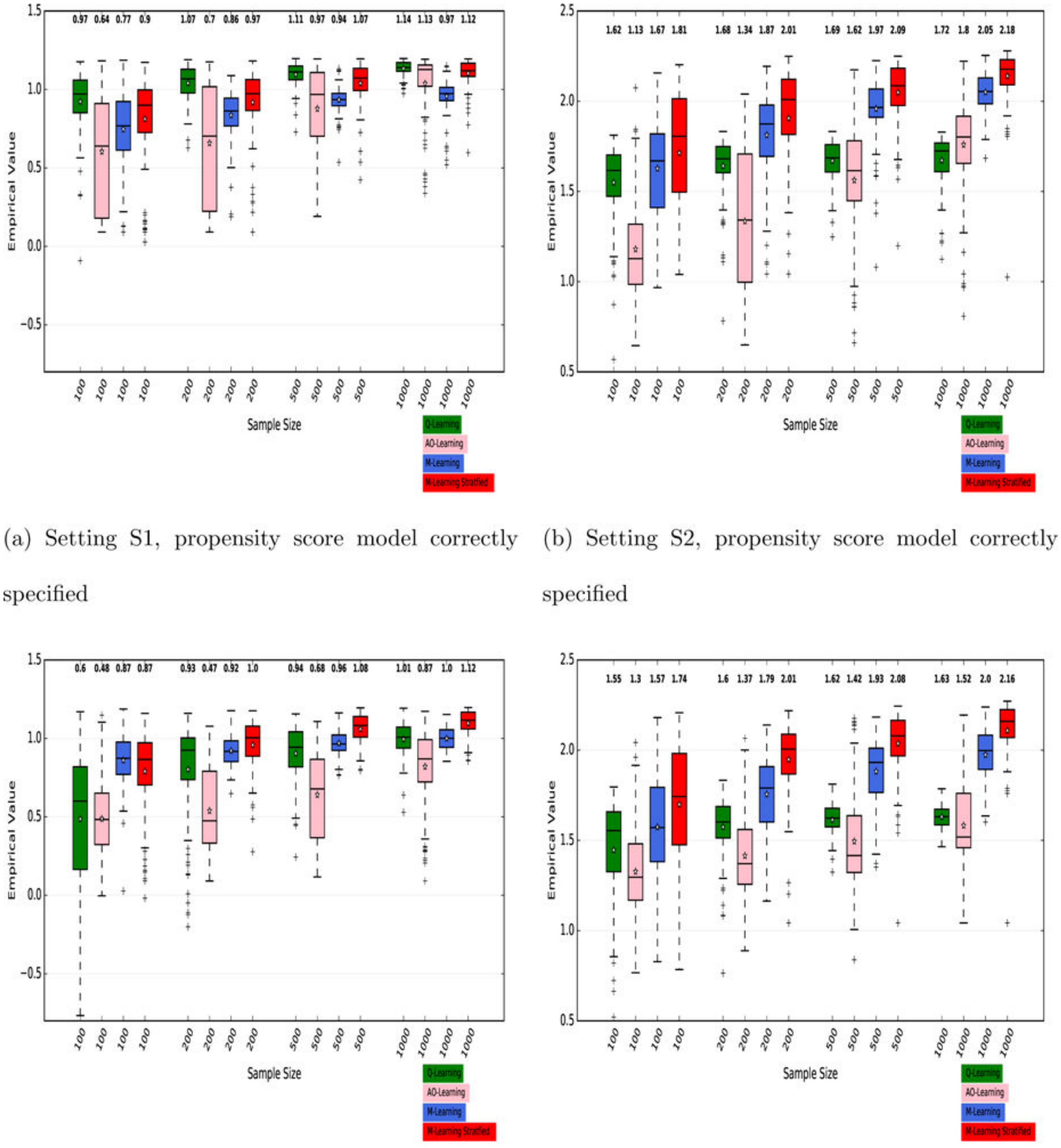
Stuart EA (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25 (1), 1–21. [PubMed: 20871802]

Van Der Vaart AW, & Wellner JA (1996). Weak convergence and empirical processes: With applications to statistics New York: Springer.

Watkins CJ, & Dayan P (1992). Q-learning. Machine learning, 8 (3–4), 279–292.

Weiskopf NG, & Weng C (2013). Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association, 20 (1), 144–151. [PubMed: 22733976]

Zhao Y, Zeng D, Rush AJ, & Kosorok MR (2012). Estimating individualized treatment rules using outcome weighted learning. Journal of the American Statistical Association, 107 (499), 1106–1118. [PubMed: 23630406]



(a) Setting S1, propensity score model correctly specified

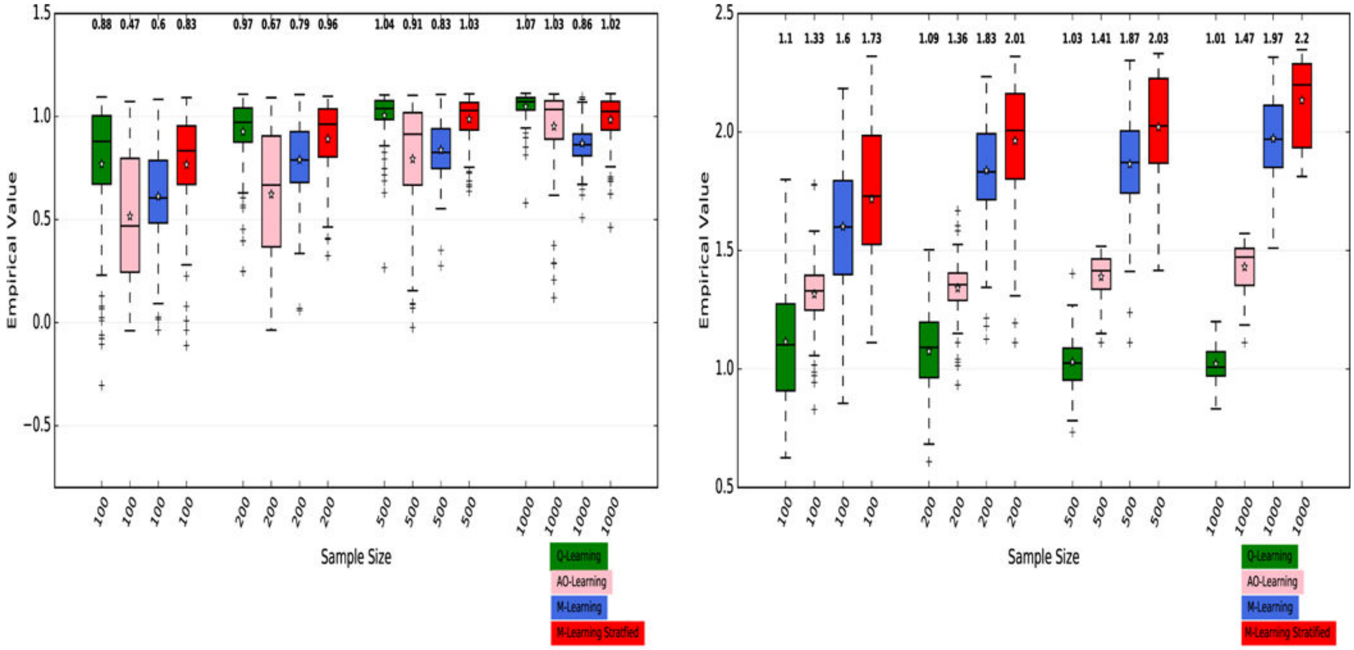
(b) Setting S2, propensity score model correctly specified

(c) Setting S1, propensity score model misspecified

(d) Setting S2, propensity score model misspecified

Figure 1:

Value comparison of four methods with propensity scores correctly specified (top panel) and misspecified (bottom panel). The numbers at the top of each subfigure are mean values.



(a) Setting S3: unmeasured confounders present

(b) Setting S4: unmeasured confounders present

Figure 2:

Value comparison of four methods in the presence of unmeasured confounders. The numbers at the top of each subfigure are mean values.

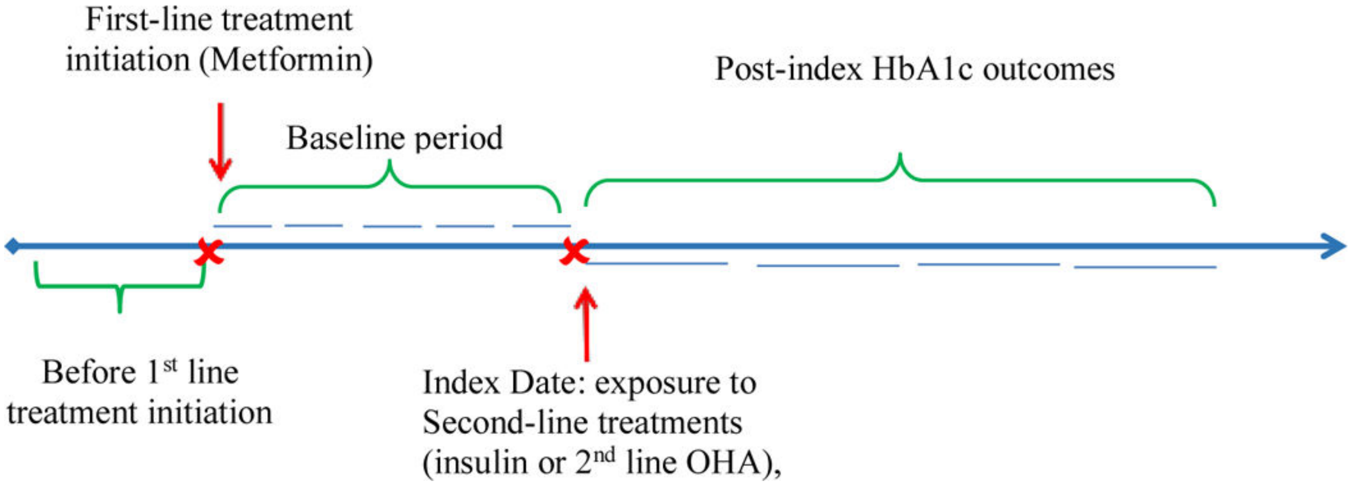
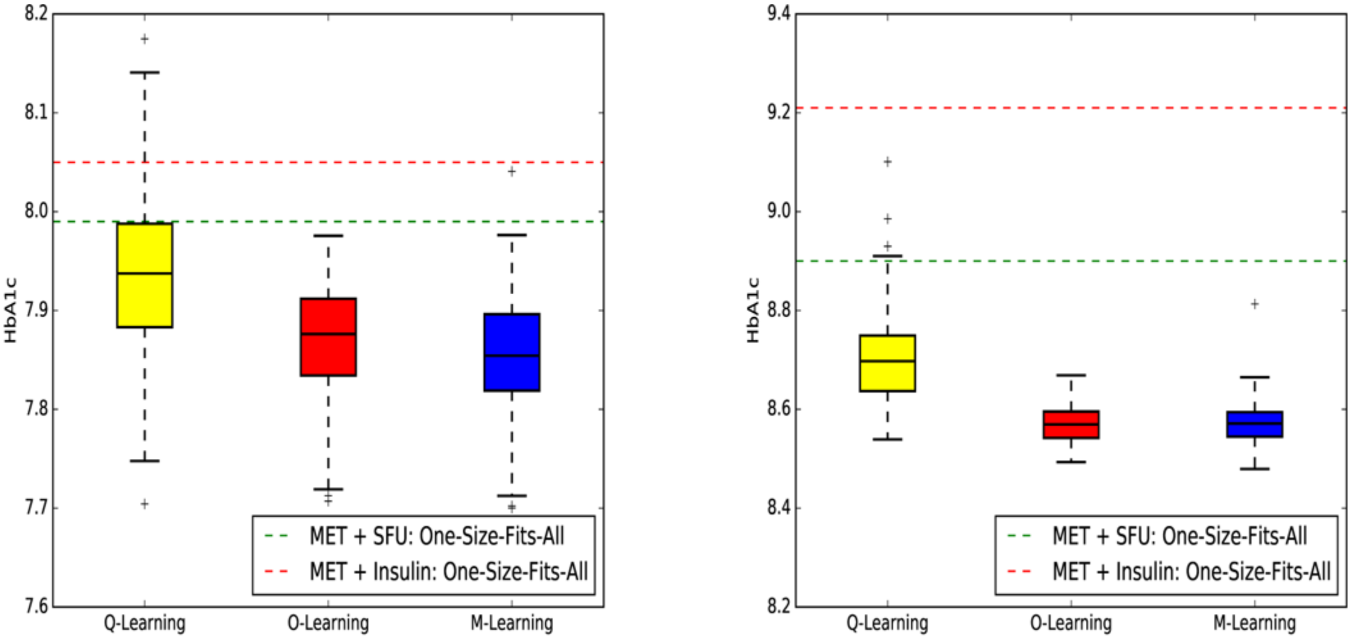

First-line treatment initiation (Metformin)
Post-index HbA1c outcomes
Baseline period
Before 1st line treatment initiation
Index Date: exposure to Second-line treatments (insulin or 2nd line OHA),

Figure 3:
T2D EHR Study Design

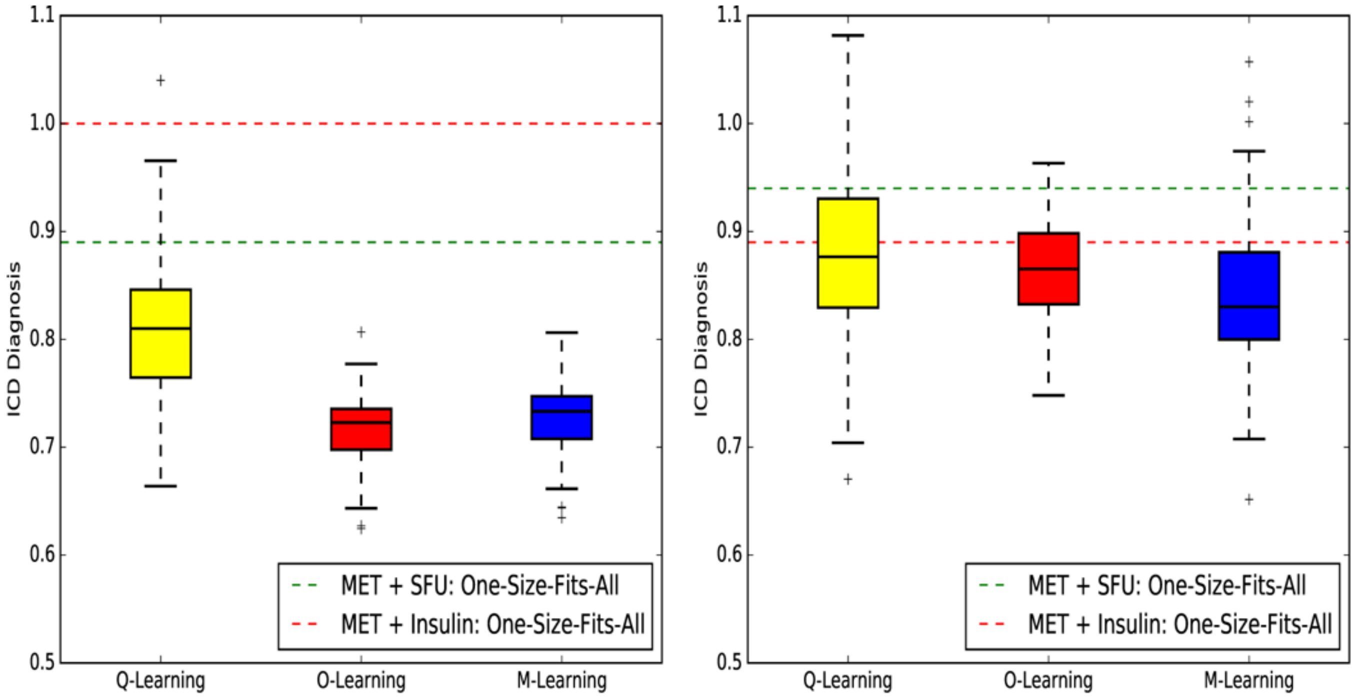


(a) Low Baseline Group

(b) High Baseline Group

Figure 4:

Empirical value function of HbA1c in EHR data with 100 2-fold cross-validations (a low value is desirable)



(a) Low Baseline Group

(b) High Baseline Group

Figure 5:

Empirical value function of ICD diagnosis count in EHR data with 100 2-fold cross-validations (a low value is desirable)

Table 1:

Cross-validated Empirical Value Function for HbA1c

High Baseline Group		
Universal rules: MET + SFU: 8.90, MET +0020Insulin: 9.21		
ITR Method	**Mean (Sth)**	**Median (Q1, Q3)**
Q-Learning	8.72 (0.124)	8.70 (8.64, 8.75)
O-learning	8.57 (0.038)	8.57 (8.54, 8.60)
M-Learning	8.57 (0.045)	8.57 (8.55, 8.59)
Low Baseline Group		
Universal rules: MET + SFU: 7.99, MET + Insulin: 8.05		
ITR Method	**Mean (Sth)**	**Median (Q1, Q3)**
Q-Learning	7.94 (0.083)	7.94 (7.88, 7.99)
O-learning	7.87 (0.061)	7.88 (7.83, 7.91)
M-Learning	7.85 (0.068)	7.85 (7.82, 7.90)

Table 2:

Cross-validated Empirical Value Function for the Number of Major Complications

High Baseline Group		
Universal rules: MET + SFU: 0.94, MET + Insulin: 0.89		
ITR Method	**Mean (Sth)**	**Median (Q1, Q3)**
Q-Learning	0.88 (0.078)	0.88 (0.83, 0.93)
O-Learning	0.86 (0.050)	0.87 (0.83, 0.90)
M-Learning	0.84 (0.068)	0.83 (0.80, 0.88)
Low Baseline Group		
Universal rules: MET + SFU: 0.89, MET + Insulin: 1.00		
ITR Method	**Mean (Sth)**	**Median (Q1, Q3)**
Q-Learning	0.81 (0.063)	0.81 (0.76, 0.85)
O-Learning	0.72 (0.033)	0.72 (0.70, 0.74)
M-Learning	0.73 (0.032)	0.73 (0.71, 0.75)