Selecting Optimal Decisions via Distributionally Robust Nearest-Neighbor Regression

Ruidi Chen[1] and Ioannis Ch. Paschalidis[1,2,3]

[1] Division of Systems Engineering, Boston University

[2] Department of Electrical & Computer Engineering, Boston University

[3] Department of Biomedical Engineering, Boston University

Abstract

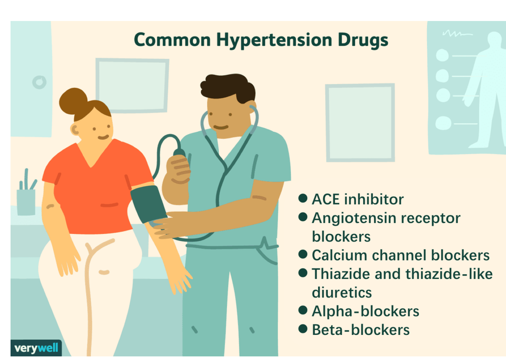
We develop a prediction-based prescriptive model for optimal decision making that

1. predicts the outcome under each action using a robust nonlinear model;
2. prescribes actions based on their predicted outcomes.

The *predictive* model combines *Distributionally Robust Linear Regression* (DRLR) with the *K-Nearest Neighbors (K-NN)* regression, which produces predictions that are robust to data perturbations and captures the nonlinearity embedded in the data. The *prescriptive* model selects each action with a probability inversely proportional to its exponentiated predicted outcome. We show theoretical guarantees on the out-of-sample performance of the predictive model, and prove the optimality of the randomized prescriptive policy in terms of the expected true future outcome. We demonstrate the proposed methodology on a diabetes and a hypertension dataset, showing that our prescribed treatment leads to a larger reduction in HbA_{1c} and systolic blood pressure compared to a series of alternatives.

Problem Description

- **Problem**: Given a set of drugs $[M] \triangleq \{1,\dots,M\}$, choose the one that yields the lowest future HbA_{1c}/systolic blood pressure y, with the aid of patient data $\mathbf{x}$ that is predictive of y.
- **Idea**: *Predict* the outcome y_m for each drug $m \in [M]$ using a robust nonlinear framework, and *prescribe* the actions based on their predictions.
- **Applications**: Prescribe optimal treatments for patients with diabetes or hypertension.



*Publicly available Internet images.

Robust Nonlinear Prescription

- Assumption: For any $m \in [M]$, $y_m = \mathbf{x}_m{}'\beta^*_m + h_m(\mathbf{x}_m) + \varepsilon_m$.
- Method:
  - For each $m \in [M]$, derive a robust estimate of β^*_m, denoted by $\hat{\beta}_m$, using *Wasserstein Distributionally Robust Optimization (DRO)*[1].
  - Given a new sample $\mathbf{x}$, find its K_m nearest neighbors, whose responses are denoted by $y_{m(i)}$, $i=[K_m]$, in each action group m using the metric:

$$\|\mathbf{x}-\mathbf{x}_{mi}\|_{\hat{\mathbf{W}}_m} = \sqrt{(\mathbf{x}-\mathbf{x}_{mi})'\hat{\mathbf{W}}_m(\mathbf{x}-\mathbf{x}_{mi})},$$

    where $\hat{\mathbf{W}}_m = \text{diag}((\hat{\beta}_{m1})^2,\dots,(\hat{\beta}_{mp})^2)$.
  - Prediction:

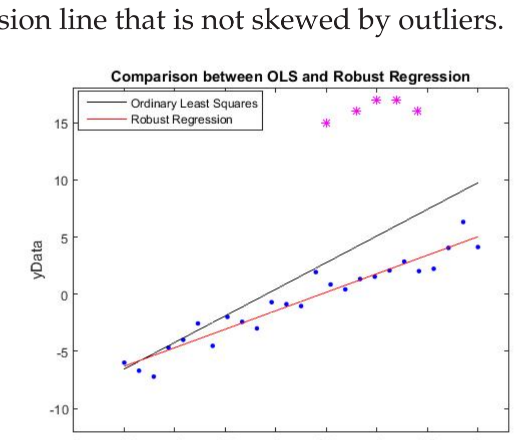
$$\hat{y}_m(\mathbf{x}) = \frac{1}{K_m}\sum_{i=1}^{K_m} y_{m(i)}.$$

  - Prescription: select action m with probability

$$e^{-\xi\hat{y}_m(\mathbf{x})} \Big/ \sum_{j=1}^{M} e^{-\xi\hat{y}_j(\mathbf{x})}.$$

Robust Regression

- Goal:
  - Estimate the regression line that is not skewed by outliers.



  - The samples $(\mathbf{x}_i, y_i)$, $i=1,\dots,N$, may be contaminated with outliers.
- Method: Inducing robustness by hedging against a set of uncertain parameters.

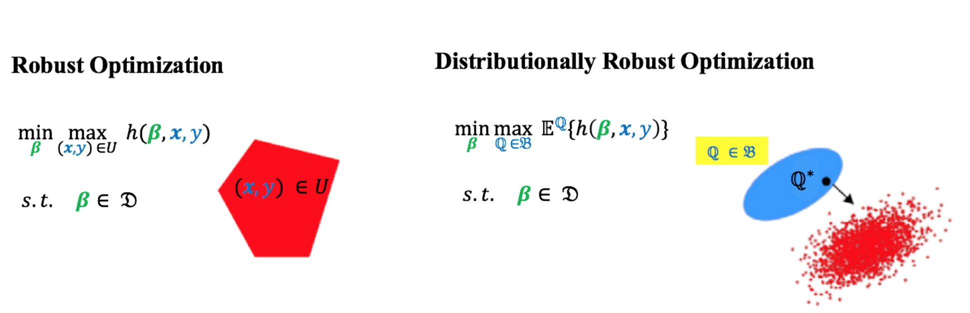


Fig. Comparisons of different optimization schemes.

Wasserstein Distributionally Robust Optimization

- The Wasserstein DRO problem:

$$\inf_{\beta\in\mathscr{D}} \sup_{\mathbb{Q}\in\mathscr{B}} E^{\mathbb{Q}}[|y-\mathbf{x}'\beta|].$$

- Notation:
  - β: the regression coefficient to be estimated; $\mathbb{Q}$: the probability distribution of $(\mathbf{x}, y)$.
  - $\mathscr{B}$: the Wasserstein ball of distributions centered at the empirical distribution $\hat{\mathbb{P}}_N$: $\mathscr{B} = \{\mathbb{Q}\in\mathscr{M}(\mathscr{Z}) : W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \le \varepsilon\}$, where the Wasserstein distance is defined through,

$$W_1(\mathbb{Q},\ \hat{\mathbb{P}}_N) \triangleq \min_{\Pi\in\mathscr{P}(\mathscr{Z}\times\mathscr{Z})}\left\{\int_{\mathscr{Z}\times\mathscr{Z}} \|(\mathbf{x}_1,y_1)-(\mathbf{x}_2,y_2)\|\ \Pi\big(d(\mathbf{x}_1,y_1),d(\mathbf{x}_2,y_2)\big)\right\},$$

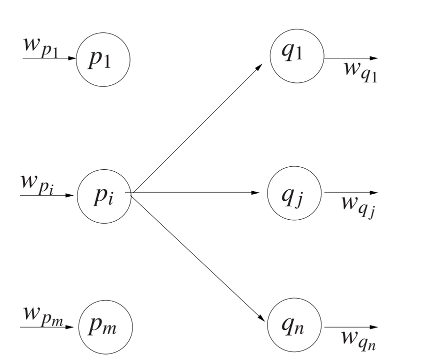
with Π the joint distribution of $(\mathbf{x}_1,y_1)$ and $(\mathbf{x}_2,y_2)$, with marginals $\mathbb{Q}$ and $\hat{\mathbb{P}}_N$.

Why the Wasserstein Metric?

- Other options: Kullback-Leibler distance, f-divergences.
- Wasserstein metric incorporates a notion of cost:

$$W_1(\mathbb{Q}_1,\mathbb{Q}_2) = \inf_{\Pi\in\mathscr{P}(\mathscr{Z}\times\mathscr{Z})}\left\{\int_{\mathscr{Z}\times\mathscr{Z}} s(\mathbf{z}_1,\mathbf{z}_2)\ \Pi(d\mathbf{z}_1,d\mathbf{z}_2)\right\}.$$

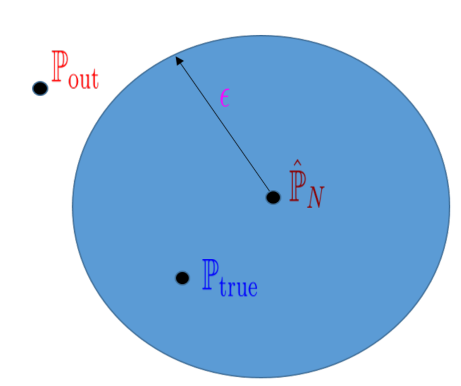
- Allow support out of the observed samples.
- Wasserstein a.k.a. optimal mass transport, earth mover's distance. Discrete case: transportation problem

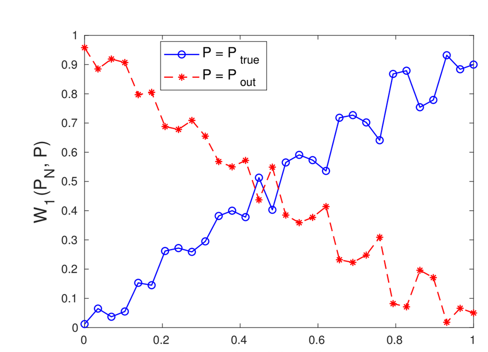


$$\begin{aligned} W_1(\mathbb{P},\mathbb{Q}) = \min_{\pi} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} \pi(i,j)s(i,j) \\ \text{s.t.} \quad & \sum_{i=1}^{m} \pi(i,j) = w_{q_j}, \quad \forall j, \\ & \sum_{j=1}^{n} \pi(i,j) = w_{p_i}, \quad \forall i, \\ & \pi(i,j) \ge 0, \quad \forall i,j. \end{aligned}$$

Robustness to Outliers

- Suppose we generate training data from a mixture of $\mathbb{P}_{\text{true}}$ (w.p. $1-q$) and $\mathbb{P}_{\text{out}}$ (w.p. q).
- Then, for $q<0.5$, $W_1(\mathbb{P}_{\text{true}},\ \hat{\mathbb{P}}_N) < W_1(\mathbb{P}_{\text{out}},\ \hat{\mathbb{P}}_N)$. Can exclude outliers!





Robustness of the Wasserstein Set

Theorem 1 *Suppose we are given two probability distributions $\mathbb{P}_{true}$ and $\mathbb{P}_{out}$, and the mixture distribution $\mathbb{P}_{mix}$ is a convex combination of the two: $\mathbb{P}_{mix} = q\mathbb{P}_{out} + (1-q)\mathbb{P}_{true}$. Then,*

$$\frac{W_1(\mathbb{P}_{out},\mathbb{P}_{mix})}{W_1(\mathbb{P}_{true},\mathbb{P}_{mix})} = \frac{1-q}{q}.$$

- When $q<0.5$, and $W_1(\mathbb{P}_{\text{true}},\mathbb{P}_{\text{mix}}) < W_1(\mathbb{P}_{\text{out}},\mathbb{P}_{\text{mix}}) \implies$ the set $\mathscr{B}$ will include the true distribution $\mathbb{P}_{\text{true}}$ and exclude the outlying one $\mathbb{P}_{\text{out}}$.

Tractable Relaxation of Wasserstein DRO

- The Wasserstein DRO problem could be relaxed to ($\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ used in the Wasserstein metric)

$$\inf_{\beta\in\mathscr{D}} \varepsilon\|(-\beta,1)\|_* + \frac{1}{N}\sum_{i=1}^{N}|y_i - \mathbf{x}_i'\beta|.$$

- Incorporates a class of models, e.g., regularized LAD, GLASSO with ℓ_1-loss.
  - Connects sparsity with robustness.
  - New interpretation for the regularization coefficient ε.
  - The regularizer controls the amount of ambiguity in the data.

Estimation Bias of the Wasserstein DRO

Theorem 2 *Under mild conditions, when the sample size $N_m \ge n_m$, with probability at least δ_m,*

$$\|\beta^*_m - \hat{\beta}_m\|_2 \le \tau_m.$$

- The parameters n_m, δ_m, τ_m are related to the Gaussian width of the unit ball in $\|\cdot\|_\infty$, the sub-Gaussian norm of $(\mathbf{x}_m, y_m)$, the eigenvalues of the covariance matrix of $(\mathbf{x}_m, y_m)$, and the geometric structure of the true regression coefficient β^*_m.

MSE of Wasserstein DRO Informed K-NN

- The bias-variance decomposition implies (η_m is the standard deviation of the noise ε_m):

$$\begin{aligned} \text{MSE}\big(\hat{y}_m(\mathbf{x})\big|\mathbf{x},\mathbf{x}_{mi}, i=[N_m]\big) &\triangleq \mathbb{E}\Big[\big(\hat{y}_m(\mathbf{x})-y_m(\mathbf{x})\big)^2\Big|\mathbf{x},\mathbf{x}_{mi},i=[N_m]\Big] \\ &= \Big(\frac{1}{K_m}\sum_{i=1}^{K_m}\big((\mathbf{x}-\mathbf{x}_{m(i)})'\beta^*_m + h_m(\mathbf{x}) - h_m(\mathbf{x}_{m(i)})\big)\Big)^2 + \frac{\eta_m^2}{K_m} + \eta_m^2. \end{aligned}$$

- For MSE to be small:
  - $\|\beta^*_m - \hat{\beta}_m\|_2$ is small;
  - $\|\mathbf{x}-\mathbf{x}_{m(i)}\|_{\hat{\mathbf{W}}_m}$ is small for $i=[K_m]$; and
  - $h_m(\mathbf{x}) - h_m(\mathbf{x}_{m(i)})$ is small for $i=[K_m]$.

Distance to the K Nearest Neighbors

Theorem 3 *Suppose we are given N_m i.i.d. samples $(\mathbf{x}_{mi}, y_{mi})$, $i\in[N_m]$, drawn from some unknown probability distribution with finite fourth moment. Every $\mathbf{x}_{mi}$ has independent, centered coordinates:*

$$\mathbb{E}(\mathbf{x}_{mi}) = \mathbf{0}, \qquad cov(\mathbf{x}_{mi}) = diag\big(\sigma^2_{m1},\dots,\sigma^2_{mp}\big), \forall i\in[N_m].$$

For a fixed predictor $\mathbf{x}$, and any given positive definite diagonal matrix $\mathbf{W}\in\mathbb{R}^{p\times p}$ with diagonal elements w_j, $j\in[p]$, and $|w_j|\le\bar{B}^2$, suppose:

$$|(x_{mij}-x_j)^2 - (\sigma^2_{mj}+x_j{}^2)| \le T_m,\ a.s.,\ \forall i\in[N_m],\ j\in[p],$$

where x_{mij}, x_j are the j-th components of $\mathbf{x}_{mi}$ and $\mathbf{x}$, respectively. Under the condition that $\bar{w}^2_m > \bar{B}^2\sum_{j=1}^{p}(\sigma^2_{mj}+x_j^2)$, with probability at least $1 - I_{1-p_{m0}}(N_m-K_m+1,K_m)$,

$$\|\mathbf{x}-\mathbf{x}_{m(i)}\|_{\mathbf{W}} \le \bar{w}_m,\ i\in[K_m].$$

Predictive Performance

Theorem 4 *Given a fixed predictor $\mathbf{x}=(x_1,\dots,x_p)$, and some scalar $\bar{w}_m$, assuming*

1. *$h_m(\cdot)$ is Lipschitz continuous with a Lipschitz constant L_m.*
2. *$\bar{w}^2_m > \bar{B}^2_m\sum_{j=1}^{p}(\sigma^2_{mj}+x_j^2)$.*
3. *$|(x_{mij}-x_j)^2 - (\sigma^2_{mj}+x_j^2)| \le T_m$, $\forall i,j$.*
4. *The coordinates of any feasible solution to Wasserstein DRO have absolute values greater than or equal to some positive number b_m (dense estimators).*

When $N_m \ge n_m$, with probability at least $\delta_m - I_{1-p_{m0}}(N_m-K_m+1,K_m)$ w.r.t. the measure of samples,

$$\mathbb{E}\Big[(\hat{y}_m(\mathbf{x})-y_m(\mathbf{x}))^2\Big|\mathbf{x},\mathbf{x}_{mi},i=[N_m]\Big] \le \left(\frac{\bar{w}_m\tau_m}{b_m} + \sqrt{p}\bar{w}_m + \frac{L_m\bar{w}_m}{\bar{B}_m}\right)^2 + \frac{\eta_m^2}{K_m} + \eta_m^2.$$

Prescriptive Performance

Theorem 5 *Given any $\mathbf{x}\in\mathbb{R}^p$, denote its predicted and true future outcome under action m by $\hat{y}_m(\mathbf{x})$ and $y_m(\mathbf{x})$, respectively. For any $k\in[M]$, the expected true outcome under the randomized prescriptive policy satisfies:*

$$\begin{aligned} \sum_{m=1}^{M}\frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}y_m(\mathbf{x}) &\le y_k(\mathbf{x}) + \Big(\hat{y}_k(\mathbf{x}) - \frac{1}{M}\sum_{m=1}^{M}\hat{y}_m(\mathbf{x})\Big) \\ &\quad + \xi\Big(\frac{1}{M}\sum_{m=1}^{M}\hat{y}_m^2(\mathbf{x}) + \sum_{m=1}^{M}\frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}y_m^2(\mathbf{x})\Big) + \frac{\log M}{\xi}. \end{aligned}$$

Activate the Randomized Strategy

- In consideration of the health care costs and treatment transients, we do not want to switch patients' treatments too frequently.
- Threshold $T(\mathbf{x})$ to activate the randomized strategy:

$$m_{\text{f}}(\mathbf{x}) = \begin{cases} m, \text{ w.p. } \frac{e^{-\xi\hat{y}_m(\mathbf{x})}}{\sum_{j=1}^{M}e^{-\xi\hat{y}_j(\mathbf{x})}}, & \text{if } \sum_k \frac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) \le x_{\text{co}} - T(\mathbf{x}), \\ m_{\text{c}}(\mathbf{x}), & \text{otherwise.} \end{cases}$$

- Find the largest $T(\mathbf{x})$ such that the probability of the expected improvement being less than $T(\mathbf{x})$ is small.

  Theorem 6 *Assume that the distribution of the predicted outcome $\hat{y}_m(\mathbf{x})$ conditional on $\mathbf{x}$, is sub-Gaussian, and its ψ_2-norm is equal to $\sqrt{2}C_m(\mathbf{x})$, for any $m\in[M]$ and any $\mathbf{x}$. Given a small $0<\bar{\varepsilon}<1$, in order to satisfy*

$$\mathbb{P}\left(\sum_k \frac{e^{-\xi\hat{y}_k(\mathbf{x})}}{\sum_j e^{-\xi\hat{y}_j(\mathbf{x})}}\hat{y}_k(\mathbf{x}) > x_{co} - T(\mathbf{x})\right) \le \bar{\varepsilon},$$

  it suffices to set a threshold

$$T(\mathbf{x}) = \max\Big(0,\ \min_m\Big(x_{co} - \mu_{\hat{y}_m}(\mathbf{x}) - \sqrt{-2C_m^2(\mathbf{x})\log(\bar{\varepsilon}/M)}\Big)\Big),$$

  where $\mu_{\hat{y}_m}(\mathbf{x}) = \mathbb{E}[\hat{y}_m(\mathbf{x})|\mathbf{x}]$.

- As $\xi\to\infty$, the randomized policy becomes deterministic.

$$m_{\text{f}}(\mathbf{x}) = \begin{cases} \arg\min_m \hat{y}_m(\mathbf{x}), & \text{if } \min_m \hat{y}_m(\mathbf{x}) \le x_{\text{co}} - T(\mathbf{x}), \\ m_{\text{c}}(\mathbf{x}), & \text{otherwise.} \end{cases}$$

  A slight modification to the threshold level $T(\mathbf{x})$ is given below:

$$T(\mathbf{x}) = \max\Big(0,\ \min_m\Big(x_{\text{co}} - \mu_{\hat{y}_m}(\mathbf{x}) - \sqrt{-2C_m^2(\mathbf{x})\log\bar{\varepsilon}}\Big)\Big).$$

Estimate $\mu_{\hat{y}_m}(\mathbf{x})$ and $C_m^2(\mathbf{x})$

Algorithm 1 Estimating the conditional mean and standard deviation of the predicted outcome.

Input: a feature vector $\mathbf{x}$; a_m: the number of subsamples used to compute $\hat{\beta}_m$, $a_m < N_m$; d_m: the number of repetitions.
for $i=1,\dots,d_m$ **do**
  Randomly pick a_m samples from group m, and use them to estimate a robust regression coefficient $\hat{\beta}_{m_i}$ through solving Wasserstein DRO.
  The future outcome for $\mathbf{x}$ under action m is predicted as $\hat{y}_{m_i}(\mathbf{x}) = \mathbf{x}'\hat{\beta}_{m_i}$.
end for
Output: Estimate the conditional mean of $\hat{y}_m(\mathbf{x})$ as:

$$\mu_{\hat{y}_m}(\mathbf{x}) = \frac{1}{d_m}\sum_{i=1}^{d_m}\hat{y}_{m_i}(\mathbf{x}),$$

and the conditional standard deviation as:

$$C_m(\mathbf{x}) = \sqrt{\frac{1}{d_m-1}\sum_{i=1}^{d_m}\Big(\hat{y}_{m_i}(\mathbf{x}) - \mu_{\hat{y}_m}(\mathbf{x})\Big)^2}.$$

Prescribe Optimal Treatments

- Goal: develop optimal prescriptions for patients with type-2 diabetes and hypertension using the Electronic Health Records (EHRs).
- Predictors: demographics, diagnoses, lab tests, and past admission records.
- Response: HbA_{1c}, and systolic blood pressure.
- The reduction in HbA_{1c}/systolic blood pressure, mean (std.):

	Diabetes		Hypertension	
	Deterministic	Randomized	Deterministic	Randomized
LASSO	-0.51 (0.16)	-0.51 (0.16)	-4.22 (0.20)	-4.22 (0.19)
CART	-0.45 (0.13)	-0.42 (0.14)	-4.48 (0.55)	-4.51 (0.49)
OLS+K-NN	-0.53 (0.13)	-0.53 (0.13)	-4.27 (0.32)	-4.29 (0.31)
DRO+K-NN	-0.56 (0.06)	-0.55 (0.08)	-6.58 (0.70)	-6.78 (0.73)
Current therapy	-0.22 (0.04)		-2.50 (0.16)	
Standard of care	-0.22 (0.03)		-2.37 (0.11)	

Refinement of the Policy

- K-NN is sensitive to the number of neighbors K_m.
- Propose a patient-specific number of neighbors K'_m, where the neighbors that are relatively far away from the patient in query are discarded.
- Denote by d_i^m the distance between the patient in query and her i-th closest neighbor in group m, and define $j^*_m = \arg\max_j\Big(d_j^m - \sum_{i=1}^{j-1}\frac{d_i^m}{j-1}\Big)$.

$$K'_m = \begin{cases} j^*_m - 1, & \text{if } \frac{d^m_{j^*_m} - \sum_{i=1}^{j^*_m-1}\frac{d_i^m}{j^*_m-1}}{\sum_{i=1}^{j^*_m-1}\frac{d_i^m}{j^*_m-1}} > \tilde{T}, \\ K_m, & \text{otherwise,} \end{cases}$$

  where $\tilde{T}$ is some threshold that can be tuned using cross-validation.
- Results on the diabetes and hypertension datasets:

	Diabetes		Hypertension	
	Deterministic	Randomized	Deterministic	Randomized
LASSO	-0.54 (0.19)	-0.54 (0.20)	-4.34 (0.28)	-4.33 (0.28)
CART	-0.62 (0.32)	-0.57 (0.27)	-4.46 (0.46)	-4.49 (0.50)
OLS+K-NN	-0.65 (0.25)	-0.64 (0.25)	-4.30 (0.35)	-4.30 (0.32)
DRO+K-NN	-0.68 (0.20)	-0.67 (0.23)	-7.42 (0.46)	-7.58 (0.51)
Current therapy	-0.23 (0.05)		-2.56 (0.14)	
Standard of care	-0.22 (0.03)		-2.37 (0.11)	

Conclusions

- All models outperform the current prescription and the standard of care.
- The Wasserstein DRO+K-NN model leads to the largest reduction in outcomes with a relatively stable performance.
- The best DRO+K-NN model leads to a 69% reduction in future systolic blood pressure compared to the 2nd best model.
- Using a patient-specific K'_m in general leads to a more significant reduction in outcomes.
- The randomized policy achieves a similar (slightly better) performance than the deterministic one.

References

[1] Chen, R., and Paschalidis, I.C. (2018). A robust learning algorithm for regression models using distributionally robust optimization under the Wasserstein metric, **Journal of Machine Learning Research**, 19, 1-48.

[2] Chen, R., and Paschalidis, I.C. (2018). Learning optimal personalized treatment rules using robust regression informed K-NN, **NIPS Machine Learning for Health (ML4H) workshop**, Montreal, Canada.

[3] Bertsimas, D., Kallus, N., Weinstein, A.M., and Zhuo, Y.D. (2017). Personalized diabetes management using electronic medical records, **Diabetes Care**.