

Supporting Information for “A Meta-Learner Framework to Estimate Individualized Treatment Effects for Survival Outcomes”

by Na Bo, Yue Wei, Lang Zeng, Chaeryon Kang, and Ying Ding

1 Appendix S1: Additional figures and tables from simulations

1.1 Unbalanced design

Web Figure 1 shows the box plots of biases and binned RMSEs and Web Table 1 shows the prediction accuracy of 100 simulations runs with a sample size of 4000 in the training dataset for the unbalanced design ($P(Z = 1) = 0.05$) at the median failure time where data were simulated from a Weibull distribution with nonzero treatment effect.

1.2 Dependent design

Web Figure 2 shows the box plots of biases and binned RMSEs at the median failure time to compare the performance of ITE estimates under the dependent design (the probability of treatment assignment depends on covariates X_1 and X_5). Web Table 2 presents the corresponding prediction accuracy.

1.3 Sensitivity analysis

Web Figures 3, 4, and 5 show the box plots of biases and binned RMSEs to compare the performance of ITE under balanced, dependent and unbalanced design at the 3rd quantile of failure times, respectively.

2 Appendix S2: Additional figures and tables for real data analysis

2.1 AREDS data analysis

Web Table 3 shows the summary of baseline characteristics of the AREDS data. Web Figure 6 plots the histograms of the estimated propensity scores for each treatment group in AREDS to check the unconfoundedness assumption. Web Figure 7 shows the mean treatment effect in RT and RC groups at year three for each method under three splits of data.

2.2 AREDS2 validation data analysis

Web Figure 8 plots the Kaplan-Meier curves for the RT (“recommended for treatment”) and RC (“recommended for control”) cohorts in AREDS2 (where the treatment recommendation rule is based on the D-X result on AREDS data) when ITE is estimated at year five. It shows that the overall 5-year progression-free probability is significantly higher in the RT group than in the RC group in this AREDS2 dataset (log-rank test $p = 0.011$).

3 Appendix S3: Convergence rate of ITE estimators

In this section, we developed convergence rate for the ITE estimators under the T- and X-learner with survival data. The conclusion is described in Theorem 1.

Applying Theorem 1 from Kennedy (14) to our survival outcome case, we can derive the minimax convergence rate of the ITE estimators for T-learner and X-learner, respectively.

Theorem 1. *Assume that the true survival function for the control group $S_0(t, \mathbf{x})$ is α_0 -smooth and the true survival function for the treatment group $S_1(t, \mathbf{x})$ is α_1 -smooth. Let m and n denote sample size used for estimating $S_0(t, \mathbf{x})$ and $S_1(t, \mathbf{x})$, respectively, and let $d = \dim(\mathbf{x})$. Then the minimax rate of T-learner and X-learner are both*

$$\mathbb{E}(\hat{\theta}(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 \lesssim m^{\frac{-2\alpha_0}{2\alpha_0+d}} + n^{\frac{-2\alpha_1}{2\alpha_1+d}}.$$

The definition of the function’s smoothness and detailed proofs of the convergence rate are stated below.

We define s -smooth functions in the Hölder class using definitions presented in (14).

Definition [s -smooth functions (14)]: Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ denote the set of functions in the Hölder class, which are k -times continuously differentiable and have bounded partial derivatives. Let $0 < p \leq 1$ and $s = k + p$. f is s -smooth if for all $x, y \in \mathbb{R}^d$,

$$\left| \frac{\partial^k f}{\partial x_1^{m_1} \dots \partial x_d^{m_d}}(x) - \frac{\partial^k f}{\partial x_1^{m_1} \dots \partial x_d^{m_d}}(y) \right| \lesssim \|x - y\|^p$$

where $m = (m_1, \dots, m_d)$ such that $\sum_{j=1}^d m_j = k$. $A \lesssim B$ denotes $A \leq cB$ for some constant $c > 0$. $\|\cdot\|$ denotes the Euclidean norm.

If a function f is s -smooth, then $\hat{f}(x)$ achieves the minimax rate denoted as

$$\mathbb{E} \left[\hat{f}(x) - f(x) \right]^2 \lesssim n^{\frac{-2s}{2s+d}},$$

where n is the sample size for the samples used for estimating $\hat{f}(x)$.

Convergence rate for T-learner. Under the T-learner framework, $\hat{\theta}(t, \mathbf{x}) = \hat{S}_1(t, \mathbf{x}) - \hat{S}_0(t, \mathbf{x})$. Assume that survival probability for control arm $S_0(t, \mathbf{x})$ is α_0 -smooth and survival probability for treatment arm $S_1(t, \mathbf{x})$ is α_1 -smooth. We use m and n to denote samples size for control group and treatment group used for estimating $\hat{S}_0(t, \mathbf{x})$ and $\hat{S}_1(t, \mathbf{x})$, respectively. Thus, the minimax rate of T-learner is

$$\begin{aligned}\mathbb{E}(\hat{\theta}(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 &\leq 2\{\mathbb{E}(\hat{S}_0(t, \mathbf{x}) - S_0(t, \mathbf{x}))^2 + \mathbb{E}(\hat{S}_1(t, \mathbf{x}) - S_1(t, \mathbf{x}))^2\} \\ &\lesssim m^{\frac{-2\alpha_0}{2\alpha_0+d}} + n^{\frac{-2\alpha_1}{2\alpha_1+d}}\end{aligned}$$

by using $(a + b)^2 \leq 2(a^2 + b^2)$ in the first inequality.

Convergence rate for X-learner. Denote $W = \{W_0, W_1\}$ as the datasets for the first stage and second stage estimation with an equal sample size of N , respectively. We use m_0 and n_0 to denote the sample size for the control group and treatment group in dataset W_0 , and m_1 and n_1 to denote the sample size for the control group and treatment group in dataset W_1 . Following the Theorem 1 in Kennedy (14), we estimate nuisance parameters $\hat{S}_1(t, \mathbf{x})$, $\hat{S}_0(t, \mathbf{x})$ and propensity score $g(\mathbf{x})$ in the first stage using dataset W_0 . In second stage, we regress pseudo outcomes on covariates for treatment group and control group respectively using dataset W_1 , denoted as $\hat{\theta}_1(t, \mathbf{x}) = \hat{\mathbb{E}}_n(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 1)$ and $\hat{\theta}_0(t, \mathbf{x}) = \hat{\mathbb{E}}_n(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0)$, where the pseudo outcome is constructed as $\tilde{D}(t) = \hat{S}_1(t, \mathbf{x}) - \hat{S}_0(t, \mathbf{x})$.

The ITE estimator from the X-learner is constructed as $\hat{\theta}(t, \mathbf{x}) = g(\mathbf{x})\hat{\theta}_0(t, \mathbf{x}) + \{1 - g(\mathbf{x})\}\hat{\theta}_1(t, \mathbf{x})$. We apply Theorem 1 in (14) to $\hat{\theta}_0(t, \mathbf{x})$ and $\hat{\theta}_1(t, \mathbf{x})$, respectively. Thus, the minimax rate of X-learner is

$$\mathbb{E}(\hat{\theta}(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 \lesssim m_0^{\frac{-2\alpha_0}{2\alpha_0+d}} + n_0^{\frac{-2\alpha_1}{2\alpha_1+d}}.$$

The proof is provided below. We focus on the derivation of the minimax rate for $\hat{\theta}_0(t, \mathbf{x})$. The minimax rate for $\hat{\theta}_1(t, \mathbf{x})$ can be derived by symmetry.

Define $\tilde{\theta}_0(t, \mathbf{x})$ as the oracle regression of $\tilde{D}(t)$ on $\mathbf{X}$ assuming we have access to the nuisance parameters. Applying (14)'s Theorem 1, the minimax rate of $\hat{\theta}_0(t, \mathbf{x})$ is:

$$\mathbb{E}(\hat{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 \lesssim \mathbb{E}(\tilde{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 + \mathbb{E}[\{\mathbb{E}(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0, W_0, t) - \theta(t, \mathbf{x})\}^2].$$

The oracle regression in the second stage is

$$\tilde{\theta}_0(t, \mathbf{x}) = \hat{\mathbb{E}}_n(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0) = \hat{\mathbb{E}}_n(S_1(t, \mathbf{x}) - S_0(t, \mathbf{x}) \mid \mathbf{X} = \mathbf{x}, Z = 0).$$

Then, $\mathbb{E}(\tilde{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x})) = 0$.

Define $R_0 = \mathbb{E}(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0, W_0, t) - \theta(t, \mathbf{x})$, then we have

$$\mathbb{E}(R_0^2) = \mathbb{E}[\{\mathbb{E}(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0, W_0, t) - \theta(t, \mathbf{x})\}^2].$$

Working out the algebra, we further have

$$\begin{aligned}R_0 &= \mathbb{E}(\tilde{D}(t) \mid \mathbf{X} = \mathbf{x}, Z = 0, W_0, t) - \theta(t, \mathbf{x}) \\ &= \mathbb{E}(\hat{S}_1(t, \mathbf{x}) - \hat{S}_0(t, \mathbf{x}) \mid \mathbf{X} = \mathbf{x}, Z = 0, W_0, t) - \theta(t, \mathbf{x}) \\ &= \{\hat{S}_1(t, \mathbf{x}) - S_1(t, \mathbf{x})\} + \{S_0(t, \mathbf{x}) - \hat{S}_0(t, \mathbf{x})\}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}(R_0^2) &\leq \mathbb{E}\{2(\hat{S}_1(t, \mathbf{x}) - S_1(t, \mathbf{x}))^2\} + \mathbb{E}\{2(S_0(t, \mathbf{x}) - \hat{S}_0(t, \mathbf{x}))^2\} \\ &\lesssim n_0^{\frac{-2\alpha_1}{2\alpha_1+d}} + m_0^{\frac{-2\alpha_0}{2\alpha_0+d}}.\end{aligned}$$

Therefore, the minimax rate of $\hat{\theta}_0(t, \mathbf{x})$ is:

$$\mathbb{E}(\hat{\theta}_0(t, \mathbf{x}) - \theta_0(t, \mathbf{x}))^2 \lesssim n_0^{-\frac{2\alpha_1}{2\alpha_1+d}} + m_0^{-\frac{2\alpha_0}{2\alpha_0+d}}.$$

By symmetry, the minimax rate of $\hat{\theta}_1(t, \mathbf{x})$ is:

$$\mathbb{E}(\hat{\theta}_1(t, \mathbf{x}) - \theta_1(t, \mathbf{x}))^2 \lesssim m_0^{-\frac{2\alpha_0}{2\alpha_0+d}} + n_0^{-\frac{2\alpha_1}{2\alpha_1+d}}.$$

Thus, the minimax rate of $\hat{\theta}(t, \mathbf{x})$ is:

$$\begin{aligned} & \mathbb{E}(\hat{\theta}(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 \\ &= \mathbb{E}[g(\mathbf{x})\hat{\theta}_0(t, \mathbf{x}) + \{1 - g(\mathbf{x})\}\hat{\theta}_1(t, \mathbf{x}) - \theta(t, \mathbf{x})]^2, \\ &= \mathbb{E}[g(\mathbf{x})(\hat{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x})) + \{1 - g(\mathbf{x})\}(\hat{\theta}_1(t, \mathbf{x}) - \theta(t, \mathbf{x}))]^2 \\ &\leq 2\mathbb{E}[g^2(\mathbf{x})(\hat{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 + \{1 - g(\mathbf{x})\}^2(\hat{\theta}_1(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2] \\ &= 2g^2(\mathbf{x})\mathbb{E}(\hat{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 + 2\{1 - g(\mathbf{x})\}^2\mathbb{E}(\hat{\theta}_1(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 \\ &\leq 2\{\mathbb{E}(\hat{\theta}_0(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2 + \mathbb{E}(\hat{\theta}_1(t, \mathbf{x}) - \theta(t, \mathbf{x}))^2\} \\ &\lesssim m_0^{-\frac{2\alpha_0}{2\alpha_0+d}} + n_0^{-\frac{2\alpha_1}{2\alpha_1+d}}. \end{aligned}$$

The second inequality holds because $0 < g(\mathbf{x}) < 1$.

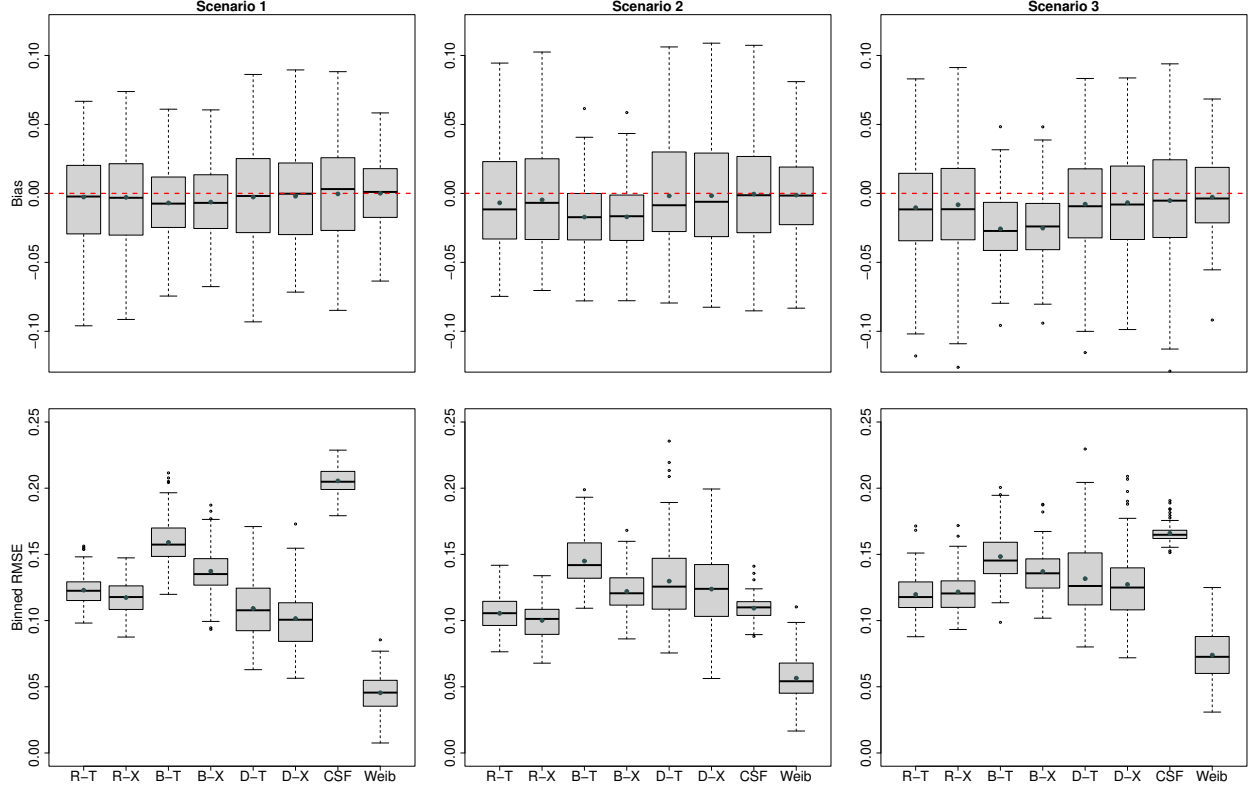
References

- [1] Age-Related Eye Disease Study Research Group: The age-related eye disease study (AREDS): design implications. AREDS report no. 1. Controlled Clinical Trials **20**(6), 573–600 (1999)
- [2] Belitser, Svetlana V., Edwin P. Martens, Wiebe R. Pestman, Rolf H.H. Groenwold, Anthonius de Boer, and Olaf H. Klungel: Measuring Balance and Model Selection in Propensity Score Methods. Archives of Ophthalmology **20**(11), 1115–29 (2011)
- [3] Cascella, R., Strafella, C., Caputo, V., Errichiello, V., Zampatti, S., Milano, F., Potenza, S., Mauriello, S., Novelli, G., Ricci, F., Cusumano, A., Giardina, E.: Towards the application of precision medicine in age-related macular degeneration. Progress in Retinal and Eye Research **63**, 132–146 (2018)
- [4] Chew, E.Y., Clemons, T., SanGiovanni, J.P., Danis, R., Domalpally, A., McBee, W., Sperduto, R., Ferris, F.L.: The age-related eye disease study 2 (AREDS2): study design and baseline characteristics (AREDS2 report number 1). Ophthalmology **119**(11), 2282–2289 (2012). URL <http://www.sciencedirect.com/science/article/pii/S0161642012004538>
- [5] Chew, E.Y., Klein, M.L., Clemons, T.E., Agrón, E., Abecasis, G.R.: Genetic testing in persons with age-related macular degeneration and the use of the AREDS supplements: to test or not to test? Ophthalmology **122**(1), 212–215 (2015)

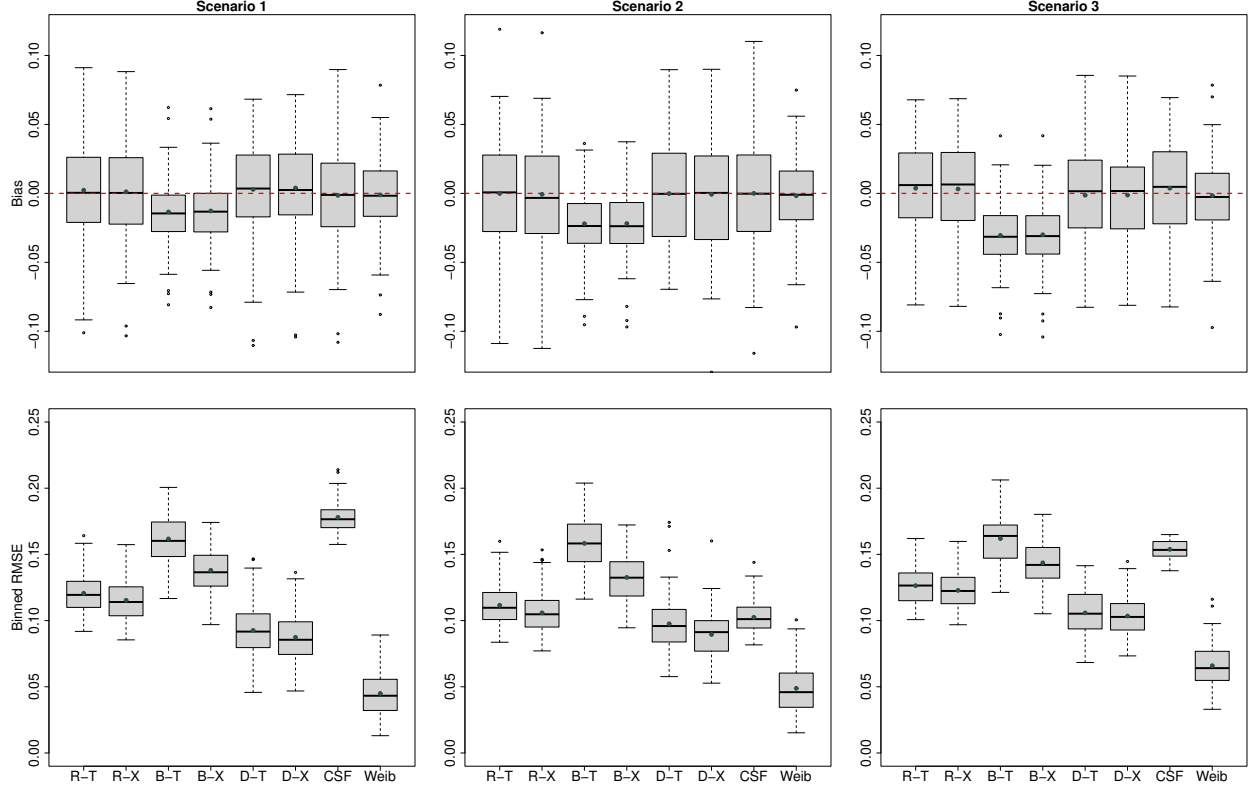
- [6] Cui, Y., Kosorok, M.R., Wager, S., Zhu, R.: Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv:2001.09887*. (2020)
- [7] Curth, A., van der Schaar, M.: Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In: A. Banerjee, K. Fukumizu (eds.) *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research*, vol. 130, pp. 1810–1818. PMLR (2021)
- [8] Foster, J.C., Taylor, J.M., Ruberg, S.J.: Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867–2880 (2011)
- [9] Greenwell, B., Boehmke, B., Cunningham, J., Developers, G.: *gbm: Generalized Boosted Regression Models* (2020). URL <https://CRAN.R-project.org/package=gbm>. R package version 2.1.8
- [10] Henderson, N.C., Louis, T.A., Rosner, G.L., Varadhan, R.: Individualized treatment effects with censored data via fully nonparametric bayesian accelerated failure time models. *Biostatistics* **21**(1), 50–68 (2020)
- [11] Hill, J.L.: Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217–240 (2011)
- [12] Ishwaran, H., Kogalur, U.: Random survival forests for r. *R News* **7**(2), 25–31 (2007). URL <https://CRAN.R-project.org/doc/Rnews/>
- [13] Johansson, F.D., Shalit, U., Kallus, N., Sontag, D.: Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research* **23**(166), 1–50 (2022). URL <http://jmlr.org/papers/v23/19-511.html>
- [14] Kennedy, E.H.: Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv:2004.14497* (2020)
- [15] Klein, M.L., Francis, P.J., Rosner, B., Reynolds, R., Hamon, S.C., Schultz, D.W., Ott, J., Seddon, J.M.: *CFH* and *LOC387715/ARMS2* genotypes and treatment with antioxidants and zinc for age-related macular degeneration. *Ophthalmology* **115**(6), 1019–1025 (2008)
- [16] Koch, B., Sainburg, T., Geraldo, P., Jiang, S., Sun, Y., Foster, J.G.: Deep learning of potential outcomes. *arXiv:2110.04442* (2021)
- [17] Kunzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B.: Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **116**(10), 4156–4165 (2018)
- [18] Kursu, M.B., Rudnicki, W.R.: Feature selection with the boruta package. *Journal of Statistical Software* **36**, 1–13 (2010)

- [19] Liaw, A., Wiener, M.: Classification and regression by randomforest. *R News* **2**(3), 18–22 (2002). URL <https://CRAN.R-project.org/doc/Rnews/>
- [20] Rubin, D.B.: Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**(4), 688–701 (1974)
- [21] Seddon, J., Silver, R., Rosner, B.: Response to AREDS supplements according to genetic factors: survival analysis approach using the eye as the unit of analysis. *British Journal of Ophthalmology* **100**(12), 1731–1737 (2016)
- [22] Shalit, U., Johansson, F.D., Sontag, D.: Estimating individual treatment effect: generalization bounds and algorithms. In: D. Precup, Y.W. Teh (eds.) *Proceedings of the 34th International Conference on Machine Learning, *Proceedings of Machine Learning Research**, vol. 70, pp. 3076–3085. PMLR (2017). URL <https://proceedings.mlr.press/v70/shalit17a.html>
- [23] Shi, C., Blei, D., Veitch, V.: Adapting neural networks for the estimation of treatment effects. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds.) *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc. (2019). URL <https://proceedings.neurips.cc/paper/2019/file/8fb5f8be2aa9d6c64a04e3ab9f63feee-Paper.pdf>
- [24] Splawa-Neyman, J., Dabrowska, D., Speed, T.: On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science* **5**(4), 465–472 (1990)
- [25] Sun, T., Ding, Y.: Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics* **22**, 315–330 (2019). DOI <https://doi.org/10.1093/biostatistics/kxz032>
- [26] Sun, T., Wei, Y., Chen, W., Ding, Y.: Genome-wide association study-based deep learning for survival prediction. *Statistics in Medicine* **39**, 4605–4620 (2020)
- [27] Tabib, S., Larocque, D.: Non-parametric individual treatment effect estimation for survival data with random forests. *Bioinformatics* **36**(2), 629–636 (2020)
- [28] Wager, S., Athey, S.: Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242 (2018)
- [29] Wei, Y., Hsu, J.C., Chen, W., Chew, E.Y., Ding, Y.: Identification and inference for subgroups with differential treatment efficacy from randomized controlled trials with survival outcomes through multiple testing. *Statistics in Medicine* **40**(29), 6523–6540 (2021)
- [30] Yan, Q., Ding, Y., Liu, Y., Sun, T., Fritsche, L.G., Clemons, T., Ratnapriya, R., Klein, M.L., Cook, R.J., Liu, Y., Fan, R., Wei, L., Abecasis, G.R., Swaroop, A., Chew, E.Y., Group, A.R., Weeks, D.E., Chen, W.: Genome-wide analysis of disease progression in age-related macular degeneration. *Human Molecular Genetics* **27**(5), 929–940 (2018)

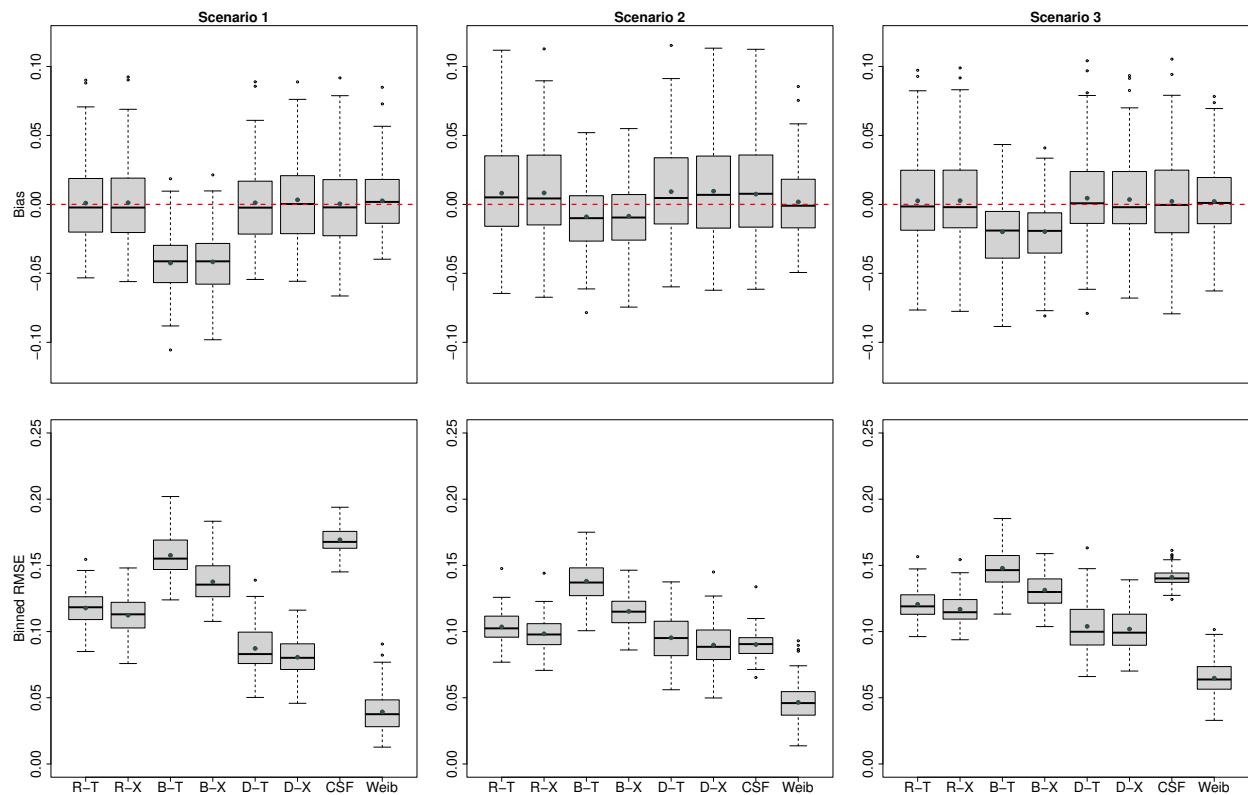
- [31] Yoon, J., Jordon, J., van der Schaar, M.: GANITE: Estimation of individualized treatment effects using generative adversarial nets. In: International Conference on Learning Representations (2018). URL <https://openreview.net/forum?id=ByKWUeWA->
- [32] Zhao, L., Tian, L., Cai, T., Claggett, B., Wei, L.J.: Effectively selecting a target population for a future comparative study. *Journal of the American Statistical Association* **108**(502), 527–539 (2013). DOI 10.1080/01621459.2013.770705. URL <https://doi.org/10.1080/01621459.2013.770705>. PMID: 24058223
- [33] Zhu, J., Gallego, B.: Targeted estimation of heterogeneous treatment effect in observational survival analysis. *Journal of Biomedical Informatics* **107**, 103474 (2020)



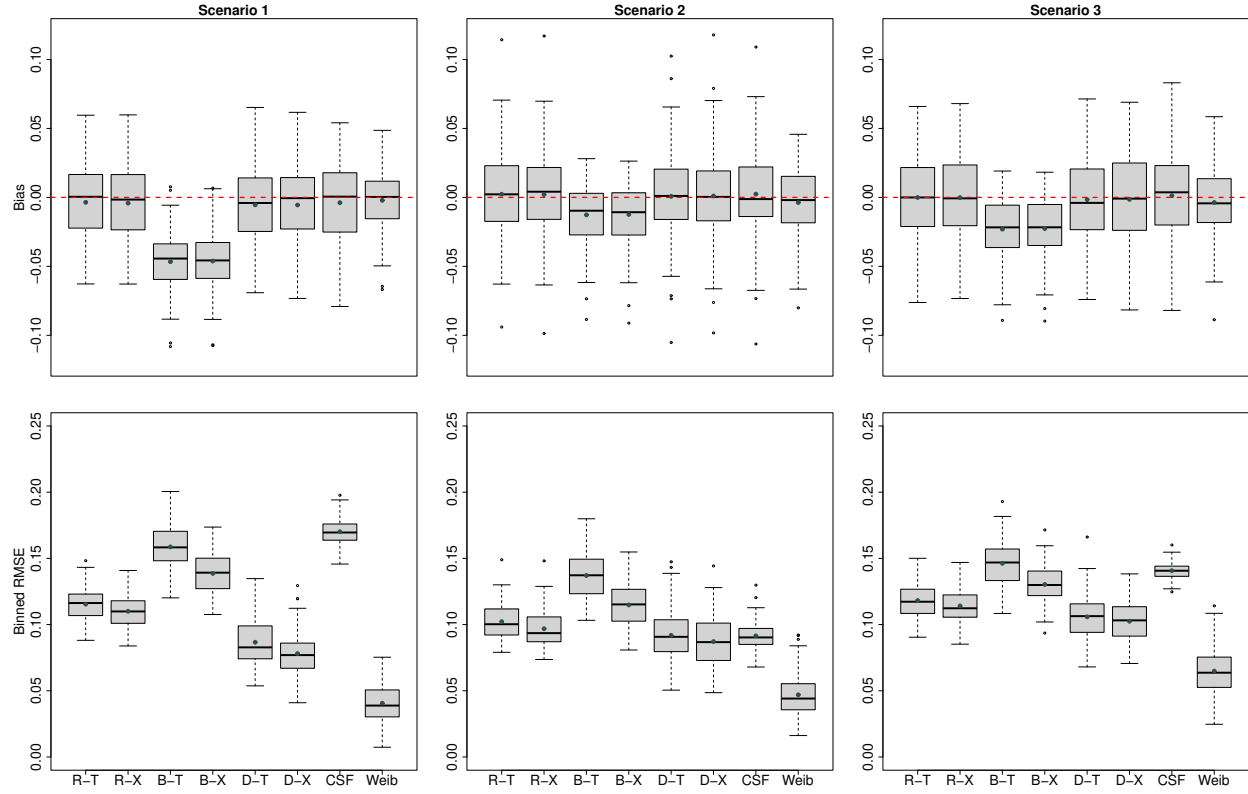
Web Figure 1: Simulation results when truths are generated from a **Weibull model**: box plots of biases (upper panel) and binned RMSEs (lower panel) to compare the performance of ITE estimates under **unbalanced design** at the median failure time with sample size (of the training dataset) as 4000. The ITE estimates from the following combinations of the meta-algorithms and base learners were examined: RSF with T-learner (R-T), RSF with X-learner (R-X), BAFT with T-learner (B-T), BAFT with X-learner (B-X), DNNSurv with T-learner (D-T) and DNNSurv with X-learner (D-X). ‘CSF’ represents the causal survival forest method, and ‘Weib’ represents the true Weibull model as the best case for ITE estimates.



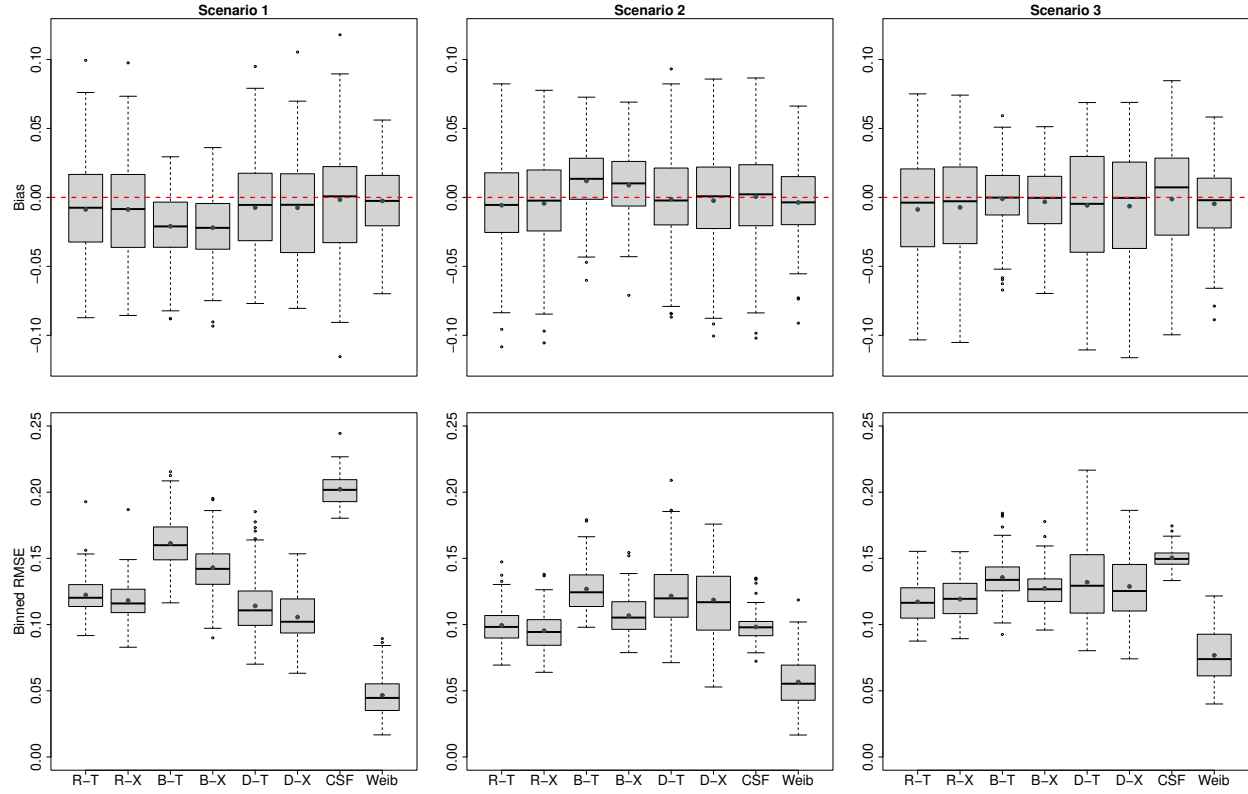
Web Figure 2: Simulation results when truths are generated from a **Weibull model**: box plots of biases and binned RMSEs to compare the performance of ITE estimates under **dependent design** at the median failure time.



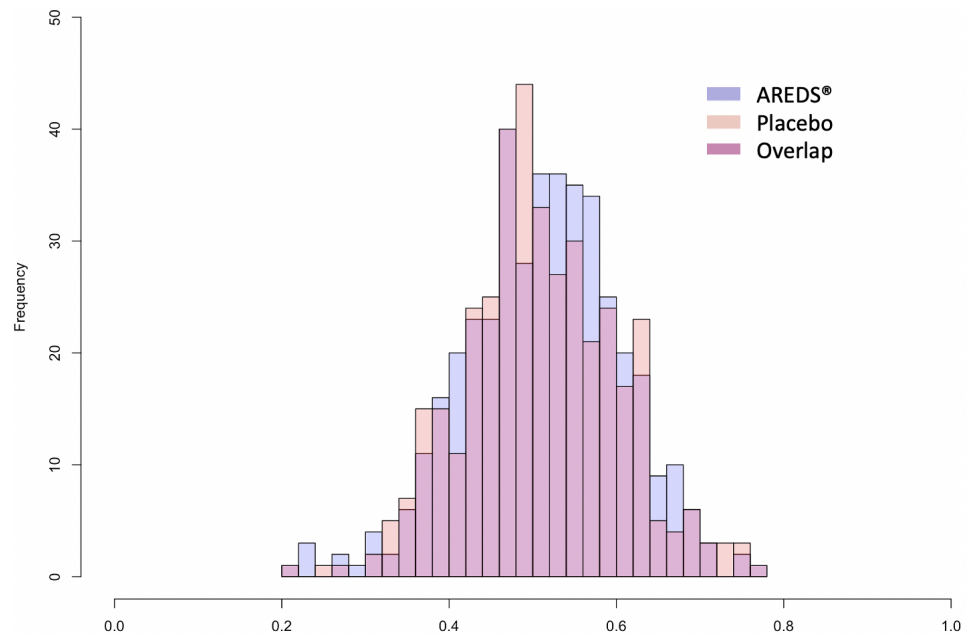
Web Figure 3: Simulation results at the **75% percentile of survival times** when truths are generated from a **Weibull model**: box plots of biases (upper panel) and binned RMSEs (lower panel) to compare the performance of ITE estimates under **balanced design**.



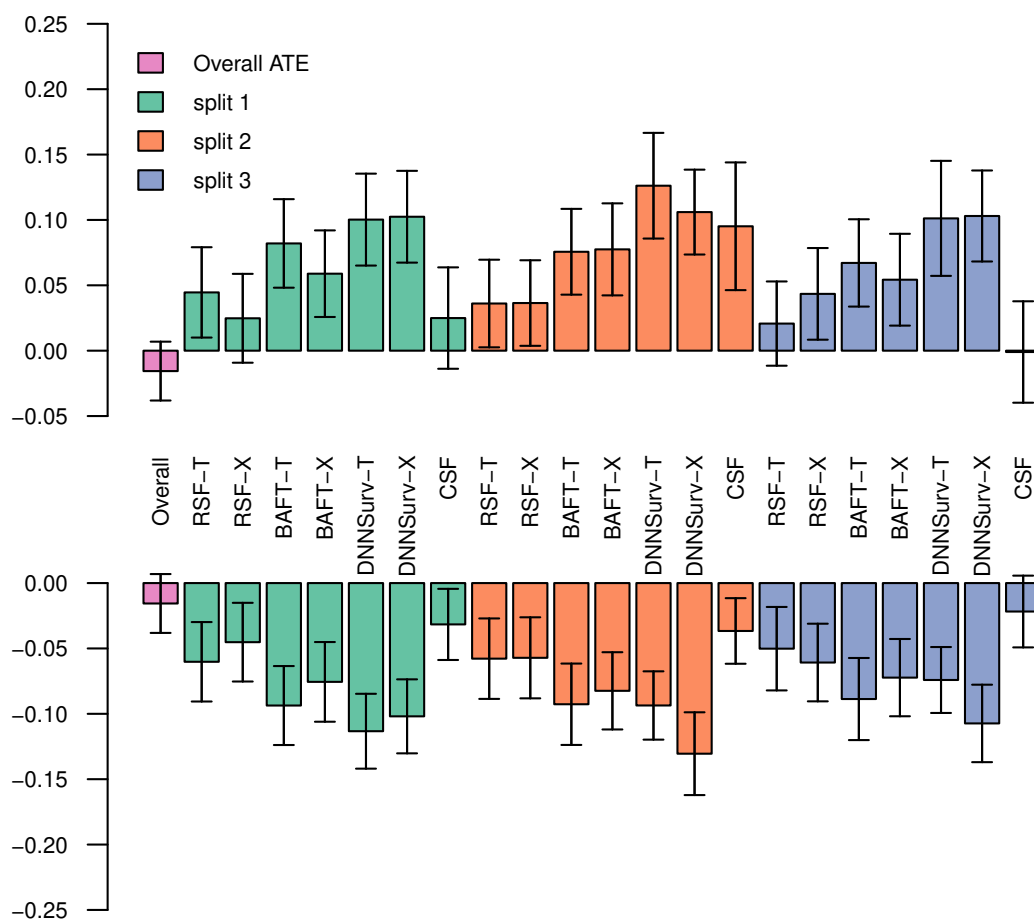
Web Figure 4: Simulation results at the **75% percentile of survival times** when truths are generated from a **Weibull model**: box plots of biases (upper panel) and binned RMSEs (lower panel) to compare the performance of ITE estimates under **dependent design**.



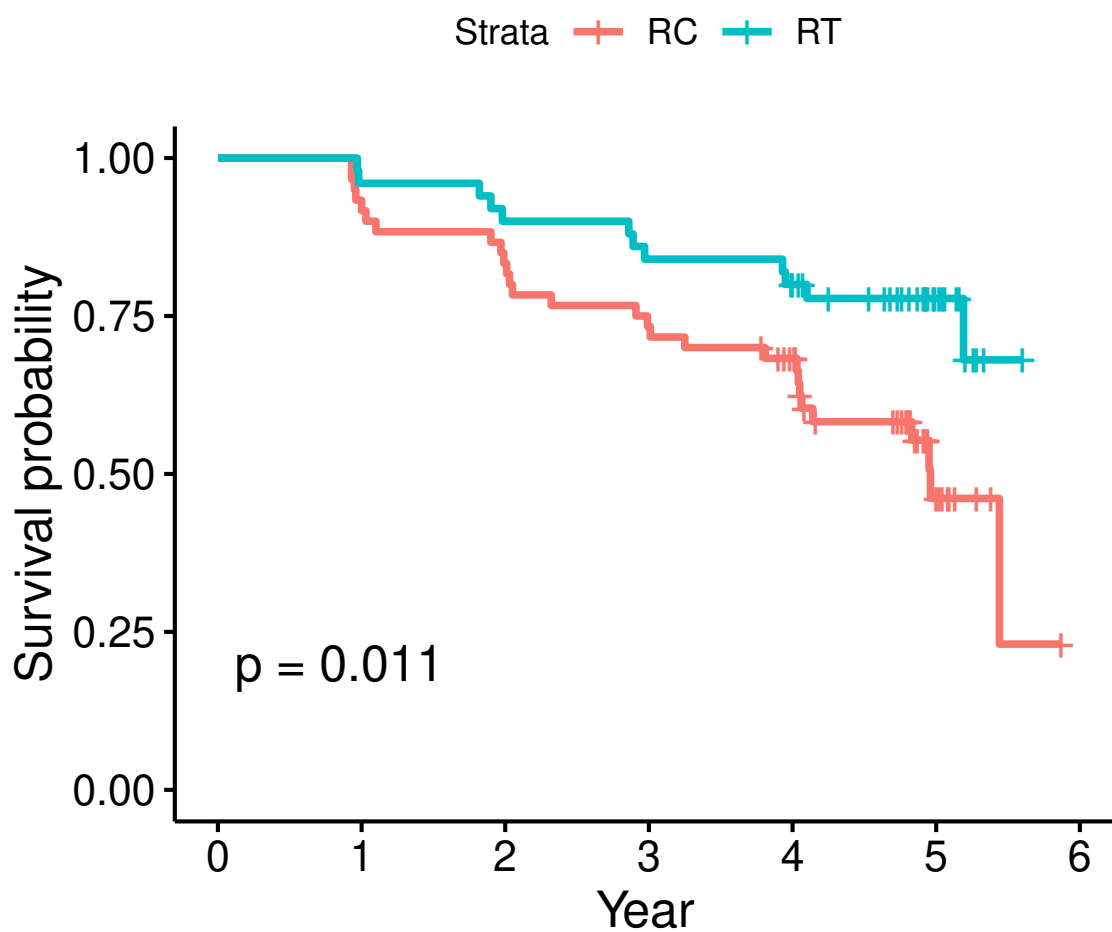
Web Figure 5: Simulation results at the **75% percentile of survival times** when truths are generated from a **Weibull model**: box plots of biases (upper panel) and binned RMSEs (lower panel) to compare the performance of ITE estimates under **unbalanced design** with sample size of 4000.



Web Figure 6: Histograms of the estimated propensity scores of getting AREDS formula in the AREDS dataset.



Web Figure 7: The mean treatment effect (KM estimates) at year three of participants in the RT cohort (recommended for taking the treatment) (upper panel) and in the RC cohort (recommended for taking the placebo) (lower panel) from each method in each split of data.



Web Figure 8: The Kaplan Meier curves of RT and RC cohorts in AREDS2 where the treatment recommendation rule is based on the D-X result on AREDS data at year five. Here ‘RT’ means ‘recommended for taking the AREDS formula’, and ‘RC’ means ‘not recommended for taking the AREDS formula’. The p-value is from the log-rank test.

Web Table 1: Simulation results when truths are generated from a **Weibull model**: prediction accuracy under **unbalanced design** at the median failure time with sample size as 4000. Six metrics are summarized with mean (SD) for each method under each scenario: overall accuracy (ACC), positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity and F-score.

	R-T	R-X	B-T	B-X	D-T	D-X	CSF	Weibull
Scenario 1								
ACC	89.87 (2.10)	90.40 (2.08)	86.38 (2.38)	88.98 (2.11)	90.27 (2.16)	91.18 (2.03)	79.24 (7.85)	95.75 (1.80)
PPV	92.06 (3.74)	92.63 (3.88)	90.45 (2.29)	92.02 (2.56)	94.10 (3.01)	94.64 (2.88)	79.07 (9.05)	97.12 (2.35)
NPV	86.33 (7.06)	86.96 (7.15)	77.53 (5.32)	82.53 (5.51)	83.14 (6.62)	84.75 (6.50)	95.58 (7.39)	93.28 (5.45)
Sensitivity	93.94 (4.22)	94.10 (4.37)	90.14 (3.12)	92.38 (3.14)	92.05 (4.00)	92.82 (3.84)	98.17 (3.71)	96.89 (2.77)
Specificity	80.39 (10.46)	81.77 (10.71)	77.62 (6.11)	81.05 (6.80)	86.11 (7.80)	87.34 (7.50)	35.11 (32.37)	93.09 (5.88)
F-score	92.84 (1.50)	93.20 (1.52)	90.25 (1.76)	92.14 (1.53)	92.95 (1.65)	93.62 (1.54)	87.15 (4.15)	96.95 (1.31)
Scenario 2								
ACC	80.06 (3.85)	80.48 (3.87)	74.88 (4.69)	77.42 (4.91)	78.53 (5.64)	78.56 (5.48)	75.16 (6.53)	89.22 (4.67)
PPV	89.58 (4.22)	89.31 (4.70)	86.55 (3.65)	87.69 (4.23)	87.91 (5.39)	87.86 (5.53)	77.79 (7.25)	93.91 (4.25)
NPV	64.71 (8.55)	65.90 (8.61)	55.80 (6.91)	59.82 (7.94)	62.53 (10.51)	62.75 (10.41)	79.54 (15.50)	82.25 (12.17)
Sensitivity	81.92 (7.77)	83.03 (7.93)	76.69 (6.27)	79.62 (6.63)	81.50 (8.25)	81.69 (8.41)	93.67 (12.86)	91.11 (7.76)
Specificity	75.52 (12.66)	74.23 (14.43)	70.43 (9.43)	72.01 (11.39)	71.25 (15.59)	70.90 (16.24)	29.80 (29.66)	84.57 (11.88)
F-score	85.22 (3.45)	85.66 (3.41)	81.16 (3.98)	83.26 (4.03)	84.21 (4.56)	84.26 (4.47)	83.88 (7.12)	92.18 (3.81)
Scenario 3								
ACC	84.41 (2.64)	84.04 (2.68)	81.82 (2.87)	82.94 (2.72)	84.31 (3.80)	84.61 (3.70)	74.16 (3.75)	91.15 (2.65)
PPV	90.01 (3.57)	88.54 (3.97)	89.50 (2.56)	88.83 (2.70)	90.93 (3.80)	90.73 (3.82)	76.97 (6.95)	94.99 (2.77)
NPV	73.36 (7.49)	75.03 (8.27)	66.58 (5.58)	68.99 (6.14)	71.65 (8.16)	72.78 (8.76)	84.06 (19.55)	83.68 (7.73)
Sensitivity	88.28 (5.60)	89.66 (5.91)	84.55 (4.10)	86.47 (4.37)	86.95 (5.75)	87.66 (5.77)	93.61 (10.81)	92.62 (4.38)
Specificity	74.77 (11.24)	70.05 (13.17)	75.03 (7.26)	72.56 (8.19)	77.74 (10.83)	77.01 (11.07)	25.74 (29.93)	87.47 (7.47)
F-score	88.93 (2.19)	88.85 (2.14)	86.87 (2.25)	87.53 (2.13)	88.71 (2.97)	88.98 (2.86)	83.65 (3.03)	93.69 (2.02)

Web Table 2: Simulation results when truths are generated from a **Weibull model**: prediction accuracy under the **dependent design** at the median failure time.

	R-T	R-X	B-T	B-X	D-T	D-X	CSF	Weibull
Scenario 1								
ACC	89.86 (1.67)	90.41 (1.72)	85.64 (2.18)	88.21 (1.94)	92.23 (2.34)	92.99 (2.25)	81.54 (6.55)	96.29 (1.88)
PPV	91.47 (3.07)	92.10 (3.17)	90.78 (1.95)	92.34 (2.16)	95.25 (2.38)	95.99 (2.26)	80.66 (7.35)	97.71 (2.15)
NPV	87.17 (6.46)	87.68 (6.65)	75.12 (5.09)	79.78 (5.28)	86.68 (6.90)	87.56 (6.88)	95.93 (6.34)	93.70 (5.39)
Sensitivity	94.54 (3.64)	94.65 (3.75)	88.53 (3.38)	90.78 (3.31)	93.70 (4.28)	94.03 (4.24)	98.45 (3.00)	97.06 (2.78)
Specificity	78.94 (8.77)	80.55 (8.90)	78.88 (5.23)	82.23 (5.76)	88.80 (6.03)	90.55 (5.72)	42.11 (26.37)	94.52 (5.37)
F-score	92.87 (1.21)	93.24 (1.25)	89.59 (1.71)	91.49 (1.50)	94.37 (1.89)	94.90 (1.82)	88.38 (3.50)	97.34 (1.38)
Scenario 2								
ACC	79.03 (4.68)	79.71 (4.88)	71.93 (4.35)	74.76 (4.55)	81.79 (5.62)	82.24 (5.27)	74.71 (5.68)	90.45 (4.82)
PPV	86.59 (4.76)	86.97 (5.00)	84.66 (3.60)	86.11 (4.19)	90.64 (5.18)	90.56 (5.33)	77.41 (7.72)	94.37 (4.80)
NPV	64.76 (9.15)	66.18 (9.46)	51.51 (5.89)	55.52 (6.62)	68.61 (11.08)	69.38 (10.42)	82.58 (16.74)	84.63 (11.40)
Sensitivity	84.06 (8.20)	84.74 (8.48)	74.03 (5.63)	77.16 (5.95)	83.65 (10.09)	84.50 (9.45)	93.98 (12.25)	92.51 (7.18)
Specificity	66.72 (15.36)	67.38 (16.19)	66.79 (9.37)	68.89 (11.39)	77.21 (15.06)	76.70 (16.01)	27.50 (29.71)	85.42 (13.33)
F-score	84.91 (4.08)	85.41 (4.29)	78.85 (3.72)	81.20 (3.75)	86.42 (5.48)	86.88 (5.07)	83.76 (6.03)	93.14 (3.76)
Scenario 3								
ACC	83.68 (2.26)	83.98 (2.29)	79.55 (2.84)	81.49 (2.68)	86.94 (2.81)	86.83 (2.39)	74.29 (3.57)	91.94 (2.33)
PPV	86.71 (3.52)	86.62 (3.72)	87.94 (2.74)	88.44 (2.99)	91.85 (3.64)	91.09 (3.77)	74.23 (4.40)	95.36 (2.97)
NPV	76.81 (7.44)	78.42 (7.88)	62.95 (5.13)	66.93 (5.49)	77.71 (8.41)	78.71 (8.11)	95.58 (8.45)	85.25 (6.87)
Sensitivity	91.48 (4.82)	92.16 (4.88)	82.83 (4.17)	85.40 (4.28)	90.00 (5.38)	90.77 (5.11)	98.89 (3.24)	93.42 (3.91)
Specificity	64.24 (12.09)	63.61 (12.88)	71.38 (7.99)	71.76 (9.04)	79.34 (11.18)	77.02 (11.80)	13.04 (18.90)	88.25 (8.20)
F-score	88.86 (1.66)	89.12 (1.63)	85.21 (2.26)	86.78 (2.08)	90.72 (2.18)	90.74 (1.83)	84.64 (1.64)	94.28 (1.74)

Web Table 3: Baseline characteristics of the AREDS data used in the analysis.

Number of subjects	All (<i>n</i> = 806)	Placebo (<i>n</i> = 391)	Antioxidants and Zinc (<i>n</i> = 415)	<i>p</i> -value*
Age				0.4905
Mean (SD)	68.77 (5.05)	68.90 (5.17)	68.66 (4.93)	
Median (Range)	68.60 (55.30-81.00)	68.50 (55.30-81.00)	68.70 (55.50-79.50)	
Sex (n, %)				0.8236
Female	466 (57.82)	224 (57.29)	242 (58.31)	
Male	340 (42.18)	167 (42.71)	173 (41.69)	
Smoking (n, %)				0.6877
Never Smoked	393 (48.76)	194 (49.62)	199 (47.95)	
Former/Current Smoker	413 (51.24)	197 (50.38)	216 (52.05)	
AREDS AMD categories (n, %)				0.5474
2	312 (38.71)	158 (40.41)	154 (37.11)	
3	457 (56.70)	214 (54.73)	243 (58.55)	
4	37 (4.59)	19 (4.86)	18 (4.34)	
Baseline AREDS AMD severity score				0.6303
Mean (SD)	4.09 (2.06)	4.13 (2.06)	4.06 (2.07)	
Median (Range)	4 (1-8)	4.00 (1-8)	4 (1-8)	

**p*-value is based on two-sample *t*-test or Chi-square test for continuous or categorical variables.