Poster #PT15

# Comparing the impact of random forest vs Bayesian G-computation on matching-adjusted indirect comparisons of treatments between trials: A simulation study

**Hoora Moradian[1], Gabriel Tremblay[1], Bart Heeg[1]**
[1]Cytel, Inc., Waltham, MA, US

**G-computation provides a reliable alternative matching-adjusted indirect comparison (MAIC) method when high imbalances lead to poor effect modifier (EM) overlap and when MAIC with a propensity score (PS) weighting approach is not feasible given the small resulting effective sample size.**

## Background

- MAIC is a common method of population-adjusted indirect treatment comparison between two studies. It uses a PS weighting approach which is sensitive to poor EM overlap and small sample size of the index trial as reweighting often leads to significantly smaller effective sample sizes.[1]
- G-computation[2] is a marginalization method that can achieve more accurate estimates than MAIC when EM overlap is poor.
- Random forest[3] (RF) is a non-parametric ensemble technique that averages outcomes from multiple decision trees and can weight patient characteristics based how many times any pair of subjects ends up in the same terminal nodes.

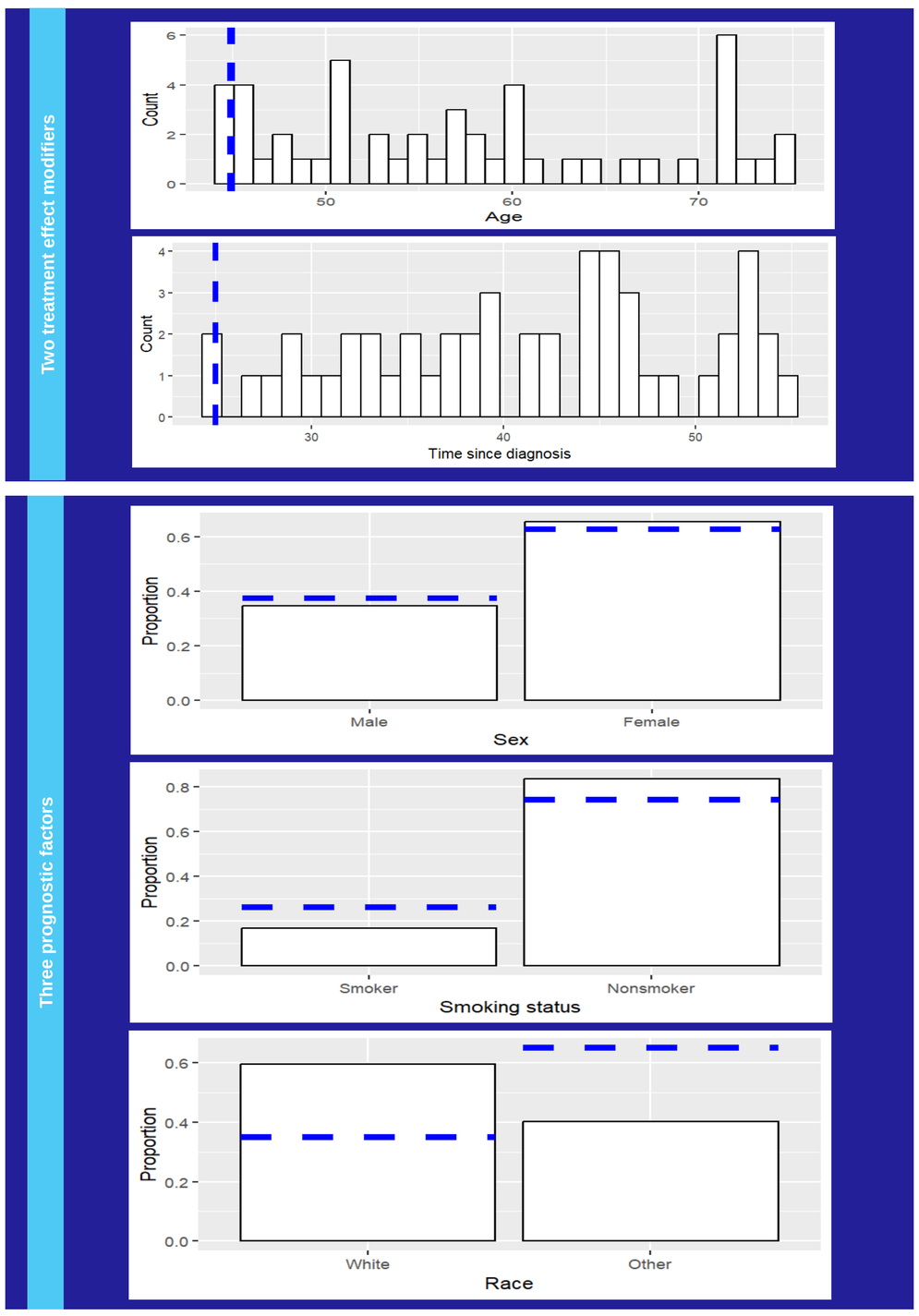
## Objective

- This study aimed to evaluate and compare the convergence and fitting of MAIC with RF and g-computation when EM overlap is poor in varied sample sizes for the index trial.
- We investigated the impact of a high level of imbalance in EMs between studies in different scenarios which were varied based on the sample size.
- Our hypothesis was that the MAIC using RF approach weight estimates were significantly more accurate than those by the g-computation approach particularly when the latter's performance deteriorates, when the index trial has a small sample size, thereby improving the results of MAICs. At large sample size, however, g-computations accuracy became significantly better than MAIC with RF.

## Methods

- The *wakefield*[4] package in R 4.2.0[5] was used to simulate data for an anchored two-study comparison with logit link function with three treatment levels included five covariates, two EMs (age and time since diagnosis) and three prognostic variables (sex, race, and smoking status).
- Weights were estimated to match the EM distributions between the two trials using RF approach and MAICs were applied over 1,000 iterations.
- All scenarios were explored at three different sample size values for the index trial (N=50, N=150, and N=300).
- The *randomForestSRC*[6] package was used to calculate the RF weights and the *maic*[7] package was employed for MAIC analyses.
- In MAIC with PS, weights were estimated with the logistic PS weighting model provided in the *maic* package. In MAIC with RF, weights were calculated by counting the number of times the comparator summary data observation fell into the same terminal node as each observation in the index trial[8].
- The R code, generously made available by Remiro-Azócar et al., 2022,[2] was used for g-computation approach.
- Figure 1 illustrates the distributions of the variables of the index trial (AB) where we have individual patient data, compared with the reported average in a comparator trial (AC) where we only have summary data.

**Figure 1. Variables distribution of the AB trial compared with the reported aggregate summary in the AC trial shown by dashed blue line in scenarios**




## Results

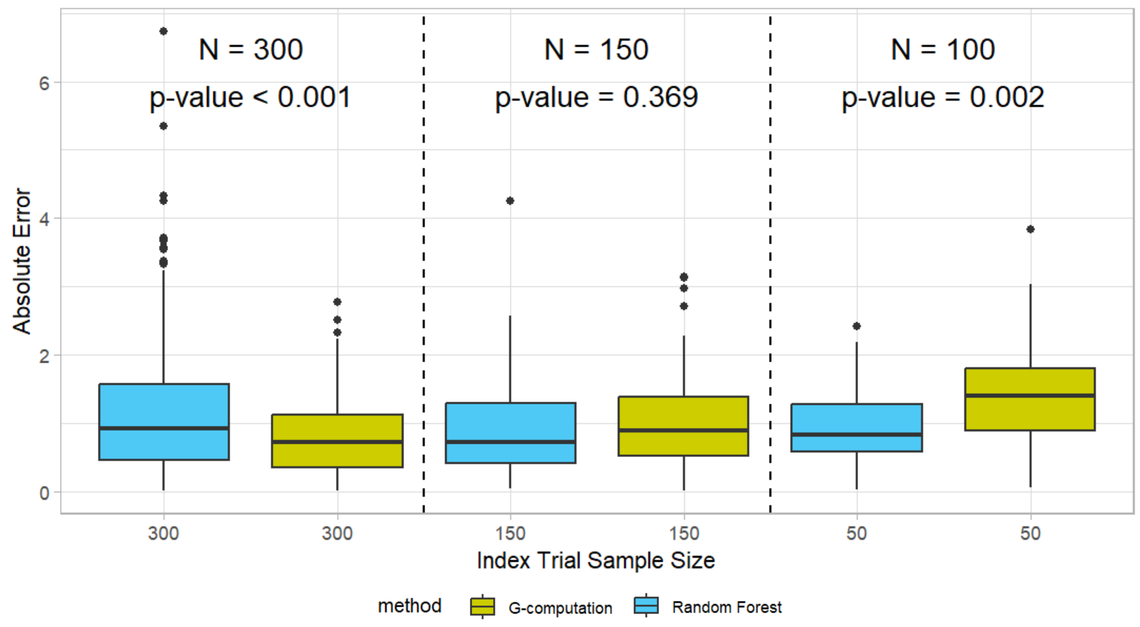
- Due to the very poor overlap, MAIC with PS did not converge at any iteration with any sample size value (N=50, N=150, and N=300). G-computation convergence increased as the sample size went from 50 to 300. MAIC with RF converged at all iterations (N=50, N=150, and N=300).
- Mean absolute error (MAE; i.e., absolute difference between the point estimates of the log odds ratio of treatment C vs. B) was significantly lower with g-computation than MAIC with RF at N=300 (MAE=0.79 vs. 1.16, p-value <0.001).
- At N=150, there was no significant difference between the two approaches (p-value=0.369).
- Notably, at N=100, MAIC with RF had significantly higher accuracy than g-computations (MAE=0.94 vs. 1.42, p-value=0.002).

**Figure 2. MAE for the true log odds ratio, number of times (out of 1,000 iterations) MAIC with PS weighting did not converge at any sample size**

| Sample size | 300 | 150 | 50 |
|---|---|---|---|
| RF mean absolute error | 1.16 | 0.91 | 0.93 |
| G-computation mean absolute error | 0.79 | 1.00 | 1.42 |
| Propensity mean absolute error | NA | NA | NA |
| Number of non-convergence runs with RF | 0 | 0 | 0 |
| Number of non-convergence runs with g-computation | 649 | 894 | 966 |
| Number of non-convergence runs with propensity | 1,000 | 1,000 | 1,000 |

Abbreviations: NA, not applicable; RF, random forest

**Figure 3. Absolute error estimating the true log odds ratio in MAICs with weights calculated with RF compared with g-computation technique over 1,000 iterations**




## Conclusions

- This simulated study demonstrated that when high imbalances lead to poor overlap, and MAIC with PS is no longer feasible given the small resulting effective sample size, g-computation provides a reliable alternative MAIC method.
- MAIC with RF is a robust alternative when in addition to poor overlap between the two studies, the index trial has a small sample size.
- This was expected since fewer observations would be likely to be incorrect given higher similarity weights when the sample size is small; this is due to the fact that the main idea behind the RF weighting algorithm is to assign large weights for observations in the index trial which are "similar" to the summary comparator data point and a weight of zero in the opposite case.
- Additional simulation and patient-level data studies should be conducted to explore results with number of EMs and medium sample size of the index trial.
- Further simulation studies should be conducted to explore results with other methods like MAIC with exact logistic regression.[9]

## References

1. Phillippo et al. (2019). Population Adjustment Methods for Indirect Comparisons: A Review of National Institute for Health and Care Excellence Technology Appraisals. International journal of technology assessment in health care, 35(3), 221–228; 2. Remiro-Azócar A, Heath A, Baio G. (2022). Parametric G-computation for compatible indirect treatment comparisons with limited individual patient data. Research Synthesis Methods, 13(6), 716-744; 3. Breiman L. (2001). Random forests. Machine learning, 45, 5-32; 4. Rinker TW. (2018). wakefield: Generate Random Data. version 0.3.3. Buffalo, New York. https://github.com/trinker/wakefield; 5. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/; 6. Ishwaran H and Kogalur UB. (2022). Fast Unified Random Forests for Survival, Regression, and Classification (RF- SRC), R package version 3.1.1. Ishwaran H and Kogalur UB. (2007). Random survival forests for R. R News 7(2), 25—31; Ishwaran H, Kogalur UB, Blackstone EH. Lauer MS. (2008). Random survival forests. Ann. Appl. Statist. 2(3), 841—860; 7. Young R (2022). maic: Matching-Adjusted Indirect Comparison_. R package version 0.1.4, https://CRAN.R-project.org/package=maic; 8. Moradian H et al. (2023). MSR18 Evaluating the Impact of Random Forest (RF) on Matching Adjusted Indirect Comparisons (MAICS) of Treatments between Trials: A Simulation Study. Value in Health, 26(6), S281; 9. Moradian H et al. (2022). MSR11 Evaluating the Impact of Exact Logistic Regression on Matching Adjusted Indirect Comparisons (MAICs) of Treatments Between Trials: A Simulation Study. Value in Health, 25(12), S351-S352. comparisons with limited individual patient data. Research Synthesis Methods. 2022;13(6):716-744.

## Acknowledgements

The authors would like to thank Colleen Dumont of Cytel for editorial review.