Machine Learning Insights into Cemented or Uncemented Hemiarthroplasty for Intracapsular Hip Fracture: A Causal Forest Analysis of the WHiTE 5 Trial

Corneliu Bolbocean, NDPCHS, University of Oxford, Oxford, UK
Zaid Hattab, J.E. Cairnes School of Business and Economics, University of Galway, Ireland
Stephen O'Neill, DHSRP, LSHTM, London, UK
Matthew L. Costa, NDORMS, University of Oxford, Oxford, UK

Coresponding author: Corneliu Bolbocean, Nuffield Department of Primary Care Health Sciences, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, Email: Corneliu.Bolbocean@phc.ox.ac.uk.

Highlights:

1. With regard to cemented versus uncemented hemiarthroplasty for hip fracture, our CF analysis indicates that treatment effects appear to be homogeneous by subgroup and timepoint.
2. Cemented hemiarthroplasty is expected to increase health-related quality of life compared with modern uncemented hemiarthroplasty for the great majority of patients having surgery for a displaced intracapsular fracture of the hip.
3. These insights are reassuring for clinical decision and health policy makers, ensuring that resource allocation within the health care system is both effective and equitable.
4. This study highlights the potential of causal forest analysis to investigate different treatment effects in subgroups of participants and across different timepoints.

What is new?

1. **Key findings:** With regard to cemented versus uncemented hemiarthroplasty for hip fracture, our CF analysis indicates that treatment effects appear to be homogeneous by subgroup and timepoint.
2. **What this adds to what is known:** This study highlights the potential of causal forest analysis to investigate different treatment effects in subgroups of participants and across different timepoints.
3. **Implications:** Cemented hemiarthroplasty is expected to increase health-related quality of life compared with modern uncemented hemiarthroplasty for the great majority of patients having surgery for a displaced intracapsular fracture of the hip.

Declarations of interest: none.

Abstract

Aim: Cemented hemiarthroplasty has been recently shown to be an effective treatment in patients with an intracapsular hip fracture. However, it remains unclear which patient groups benefit most from the use of cemented hemiarthroplasty. Knowledge about treatment effect heterogeneity is crucial for decision makers to target interventions towards specific subgroups that have the greatest benefit. We evaluate heterogeneity in the treatment effect of cemented hemiarthroplasty in the WHiTE 5 multicentre, randomized, controlled trial conducted in England and Wales using a machine learning approach. Causal Forest (CF) analysis was used to compare cemented with modern, uncemented hemiarthroplasty in patients 60 years of age or older with an intracapsular hip fracture.

Methods: We used CF to estimate subgroup- and individual-level treatment effects to compare cemented with modern, uncemented hemiarthroplasty. We used the EuroQol Group 5-Dimension (EQ-5D) multi-attribute utility scores as the main outcome measure at 1 month, 4 and 12 months follow up.

Results: Our analysis revealed a complex landscape of response to cemented hemiarthroplasty over a 12-month period. Findings suggest greater variability in treatment effects at the 1-month mark than at subsequent follow-up periods, with particular regard to subgroups based on age. Results showed that conclusions regarding heterogeneity of effects with respect to baseline characteristics, including age, health status, and lifestyle factors like alcohol consumption depend on the timepoint considered. However, in almost all cases the overall effect estimates lies within the confidence intervals for subgroups estimates.

Conclusion:

With regard to cemented versus uncemented hemiarthroplasty for hip fracture, treatment effects appear to be homogeneous by subgroup and timepoint. This study highlights the potential of causal forest analysis to investigate different treatment effects in subgroups of participants and across different timepoints.

Keywords: Heterogeneity of treatment effects, machine learning, causal forests, hemiarthroplasty, precision medicine.

1 Introduction

Hip fractures among the elderly represent a significant issue that compromises health-related quality of life (HRQoL) and imposes a considerable economic strain on healthcare systems worldwide[1–3]. The most common type of hip fracture is the displaced intracapsular fracture which is usually treated with a hemiarthroplasty. There is ongoing debate regarding the optimal method for securing the hemiarthroplasty to the bone of the femur. Evidence from a meta-analysis of randomized controlled trials indicates that bone cement fixation leads to less pain after surgery and improved mobility compared to early versions of uncemented ”press-fit” implants[4]. However, the use of bone cement has been linked to negative patient outcomes such as a decrease in blood pressure during surgery and rare instances of cardiovascular collapse and death[5].

A recent randomised trial found that cemented hemiarthroplasty resulted in a modest but statistically significantly better HRQoL than modern, hydroxyapatite-coated, uncemented hemiarthroplasty on a sample of patients aged 60 years or older with an intracapsular hip fracture[6]. However, it remains unknown whether cemented hemiarthroplasty may benefit certain patient groups more than others, i.e., whether the treatment effect is heterogeneous. The value of analyzing heterogeneous effects to support clinicians and decision makers has been acknowledged for a long time, yet studies still mainly focus on average treatment effects[1–3,6,7]. To address this when analysing the cemented hemiarthroplasty intervention, we report treatment effect heterogeneity for relevant subgroups in addition to previously published results of average treatment effects[6].

For the evaluation of heterogeneous treatment effects, several theoretical frameworks have been suggested. However, each framework has its limitations. Traditional parametric methods that employ interaction terms provide a direct way to estimate heterogeneous treatment effects. However, these methods are limited because of the interdependence of variables, especially when several interaction terms are used. This issue can reduce the depth and usefulness of the analysis[8]. The robustness of results obtained from interaction analysis can be compromised by model mis-specification[9–11]. Subgroup analysis is prone to producing inaccurate conclusions due to its tendency to be underpowered[9,12] and its susceptibility to misinterpretion of random variation as significant treatment effects[13,14]. Finally, the practice of retrospective ’effect fishing’ across multiple subgroups typically results in a spurious findings[15–17], leading to a proliferation of false-positive subgroup findings and is characterized by sampling

bias[18,19].

The causal forest was designed to address the drawbacks of traditional modelling as an approach grounded in machine learning (ML) for causal inference[20]. The causal forest's key strengths in estimating heterogeneous treatment effects include managing complex, high-dimensional interactions among a multitude of input variables without necessitating parametric assumptions by the researcher[20]. It algorithmically segments data according to variation in treatment effects across individuals[18,19] and is capable of generating confidence intervals for the estimated treatment effects[20]. Furthermore, it employs cross-fitting, or 'honesty,' as a critical element of sound statistical inference, incorporating a safeguard against overfitting through the estimation of treatment effects using out-of-bag samples[20,21]. The utility of causal forests has been demonstrated in diverse fields and increasingly in healthcare decision making[22–24].

The objective of this study was to assess heterogeneity of treatment effects on HRQoL of cemented vs uncemented hemiarthroplasty on a sample of patients of 60 years of age or older with a displaced intracapsular hip fracture using data from the WHiTE 5 trial.

2 Data - WHiTE 5 Trial

WHiTE 5 was a multicentre, randomized, controlled trial; the protocol has been published previously[7,25] and results have reported in New England Journal of Medicine[6]. Briefly, this was a randomized controlled trial comparing cemented and uncemented hemiarthroplasty in patients over 60 years old with intracapsular hip fractures. The primary outcome considered here is the health-related quality of life, assessed using the EQ-5D utility scores at 1, 4 and 12 months post-randomization.

A total of 1225 patients were enrolled in the study, with 876 (71.6%) of them providing follow-up data at 4 months. Outcome data were also available at 1 month (N=927) and 1 year (N=876). The demographic spread and baseline characteristics are consistent with the population typically affected by intracapsular hip fractures[26]. The present study found that among patients 60 years of age or older with an intracapsular hip fracture, cemented hemiarthroplasty resulted in a modest but significantly better quality of life and a lower risk of periprosthetic fracture than uncemented hemiarthroplasty, at lower cost.

3 Statistical Analysis

We utilised frequency distributions, and measures of central tendency and dispersion, such as means and standard deviations to describe the baseline characteristics of study participants. We assessed the covariates balance across the treatment arms using student t-test for continuous variables, and Pearson's chi-squared test for categorical variables across the following variables: age group, sex, proxy consent as a marker of cognitive impairment, smoking status, chronic renal failure, diabetes, alcohol consumption, residence status before injury, home ownership, residential care status, nursing care status, EQ-5D index scores and VAS scores.

We utilized the Causal Forest algorithm[20], a machine learning technique, to estimate patient-level treatment effects and then identify factors that drive the heterogeneity of these effects regarding the trial intervention. The CF method is a generalization of random forest[27] tailored to the estimation of treatment effects. A random forest comprises an ensemble of decision trees that iteratively split the dataset based on the response variable such that the groups' outcomes are as different as possible until a set stopping criterion is met. This procedure is repeated multiple times over random data subsets, which mitigates the risk of overfitting that plagues single decision trees. In causal forests, splits are determined based on expected effects rather than outcomes.

The utilization of random forests has been popular in economics, health, and environmental science due to their robust predictive capabilities and their robustness to potential confounding effects[28]. Comparative studies have demonstrated that random forests can yield comparable or superior predictions relative to traditional methods such as ordinary least squares and logistic regression[29]. This advantage stems from the model's flexibility in handling both linear and non-linear relationships and intricate inter-variable interactions, all without the need for predefined model structures. This method is implemented in the generalized random forest R package *grf*[30]. We estimate conditional average treatment effects (CATEs) for our pre-specified subgroups by taking the estimated patient-level treatment effects and plugging them into an augmented inverse propensity weighting AIPW estimator[31] of group average treatment effects[32]. Appendix A provides additional details regarding the CF and its implementation in this study.

Heterogeneity of treatment effects was assessed using existing data, informed by relevant literature[7] and clinical judgement. We considered the following pre-specified subgroup variables: sex (male,

female), age group (≤ 69), 70-79, 80-89, ≥ 90, age group (≤ 80) and > 90, smoking status, chronic renal failure, diabetes, and alcohol consumption. Statistical analysis was implemented using R software. We provide details regarding the calibration and tuning parameters in our implementation of CF in Appendix B.

4 Results

4.1 Baseline Characteristics

Table 1 shows that the baseline characteristics were balanced across the randomized arms. Specifically, this table compared baseline characteristics between patients receiving uncemented vs cemented hemiarthroplasty. The mean age for the uncemented group was 84.7 years, and for the cemented group it was 85.0 years, with no significant difference between the groups (p-value = 0.544). The mean EQ-5D multi-attribute utility scores at baseline for the uncemented group was 0.569 compared to 0.593 for the cemented group; the difference was not statistically significant (p-value = 0.223). Moreover, results show that there was no significant difference in baseline EQ-5D VAS score (p-value = 0.523) by treatment arm, nor for proxy consent (p-value = 0.104), smoking status (p-value = 0.064), chronic renal failure (p-value = 0.808), diabetes (p-value = 0.882), alcohol consumption (p-value = 0.616), or residence status before injury (p-value = 0.116) at a 5% level of significance. Overall evidence suggests that covariate balance following randomisation was achieved. This implies that the observed outcomes at follow-up times can be attributed to the type of hemiarthroplasty rather than pre-existing differences.

4.2 Heterogeneous treatment effects in EQ-5D index and VAS scores by pre-defined subgroups

Figure 1 illustrates the estimated treatment effects at the patient level for both the EQ-5D index and VAS score outcomes across three time points (1 month, 4 months, and 12 months), ordered by their magnitude. It should be noted that the sample at each time point differs due to loss to follow up, most notably at 1 year. The caterpillar plots suggest some heterogeneity at the patient level for the VAS score outcome, while the heterogeneity is less clear the EQ-5D index outcome. Figures 2 & 3 illustrates the effects across the pre-defined subgroups and over time. Findings are broadly similar for both outcomes.

Generally, the effect estimates suggest weak evidence of variation in effects, with the confidence intervals of the subgroup effect estimates including the overall effect in most cases. Notably, there is scant evidence of differences in effects among gender, diabetes, and smoking subgroups at each time point. While, some evidence of heterogeneity is detected within age and alcohol consumption subgroups, this heterogeneity varies by time point and outcome. Multiple testing concerns, and the absence of clear relationships suggests caution is warranted in viewing this as representing truly heterogeneous effects. Treatment effects on EQ-5D for the age subgroups at 1 month suggest a differential response to treatment in the short term; with older subgroups (≥ 90) experiencing higher effects compared to younger subgroups (80-90 years , 70-79 years, and ≤ 69 years). By the 4-month follow-up, the distribution of treatment effects by age group converged towards the overall effect, indicating a more uniform response to the treatment across the age subgroups. At 12 months, a similar pattern of variation re-emerges to some degree, albeit the sign flips for those aged 69 years or younger. Again, we caution that attrition means the estimates are not directly comparable. Turning to the results for subgroups based on the alcohol consumption variable, we see some evidence of heterogeneity in effects at 4 months on EQ-5D, but the pattern of heterogeneity differs in other time points and for the VAS score. Overall the results do not provide a strong justification for making different treatment decisions for the subgroups considered.

5 Discussion

This is the first study that utilised novel machine learning methods to study the heterogeneity of treatment effects in a common orthopaedic trauma surgery, leveraging the power of causal forest models to examine how different patient demographics respond to treatment. Our analysis, did not reveal statistically significant differences across age groups, providing reassurance to clinicians and policymakers that cemented hemiarthroplasty is the preferred intervention for all subgroups of patients over 60 years of age with a displaced intracapsular fracture of the hip.

With regard to treatment effects over time, at 12 months, there appears to be more variability in effect sizes across age groups compared with earlier time points. The effect size for the youngest age group ($<$ 69 years) at 12 months shows a markedly larger positive effect compared to earlier time points, although the confidence interval is wide and crosses zero. This suggests that younger patients might experience a greater benefit over time, which could be clinically relevant as these patients are

likely to live longer than older patients. The older subgroups (ages 70-79, 80-89, $\geq$90) do not exhibit a consistent trend, with effect sizes fluctuating around zero.

Across all follow-up times, females tend to have a higher positive effect size compared to males. However, the confidence intervals are wide and include zero, so these effects are not statistically significant. Nevertheless, the consistent direction of the effect might suggest significant effects by gender may be detected with a larger sample size.

Overall, with regard to cemented versus uncemented hemiarthroplasty for hip fracture, treatment effects appear to be homogeneous by subgroup and timepoint. This study highlights the potential of causal forest analysis to investigate different treatment effects in subgroups of participants and across different timepoints.

6 Conclusion

The CF approach provided estimates of individual-level treatment effects that suggest that for most patients in the WHiTE 5 Trial, cemented hemiarthroplasty is expected to increase health-related quality of life compared with modern uncemented hemiarthroplasty. The subgroup effects analysis revealed no statistically significant differences at any individual time point. However, the observed trends suggest a nuanced landscape of treatment efficacy. These insights are reassuring for clinical decision and health policy makers, ensuring that resource allocation within the health care system is both effective and equitable.

References

[1] Campion Cooper, G Campion, and LJ 3rd Melton. Hip fractures in the elderly: a world-wide projection. *Osteoporosis international*, 2:285–289, 1992.

[2] Olof Johnell and JA Kanis. An estimate of the worldwide prevalence, mortality and disability associated with hip fracture. *Osteoporosis international*, 15:897–902, 2004.

[3] XL Griffin, N Parsons, J Achten, M Fernandez, and ML Costa. Recovery of health-related quality of life in a united kingdom hip fracture population: the warwick hip trauma evaluation-a prospective cohort study. *The bone & joint journal*, 97(3):372–382, 2015.

[4] Martyn J Parker, Kurinchi Selvan Gurusamy, and Shin Azegami. Arthroplasties (with and without bone cement) for proximal femoral fractures in adults. *Cochrane database of systematic reviews*, (6), 2010.

[5] Nakulan Nantha Kumar, Setor K Kunutsor, Miguel A Fernandez, Elizabeth Dominguez, Nick Parsons, Matt L Costa, and Michael R Whitehouse. Effectiveness and safety of cemented and uncemented hemiarthroplasty in the treatment of intracapsular hip fractures: a systematic review and meta-analysis of randomized controlled trials. *The bone & joint journal*, 102(9):1113–1121, 2020.

[6] Miguel A Fernandez, Juul Achten, Nicholas Parsons, Xavier L Griffin, May-Ee Png, Jenny Gould, Alwin McGibbon, and Matthew L Costa. Cemented or uncemented hemiarthroplasty for intracapsular hip fracture. *New England Journal of Medicine*, 386(6):521–530, 2022.

[7] Miguel Antonio Fernandez, Juul Achten, Robin Gillmore Lerner, Katy Mironov, Nicholas Parsons, Melina Dritsaki, May E Png, Alwin McGibbon, Jenny Gould, Xavier Griffin, et al. Randomised controlled trial comparing hydroxyapatite coated uncemented hemiarthroplasty with cemented hemiarthroplasty for the treatment of displaced intracapsular hip fractures: a protocol for the white 5 study. *BMJ open*, 9(12):e033957, 2019.

[8] Péter Elek and Anikó Bíró. Regional differences in diabetes across europe–regression and causal forest analyses. *Economics & Human Biology*, 40:100948, 2021.

[9] Jens Hainmueller, Jonathan Mummolo, and Yiqing Xu. How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis*, 27 (2):163–192, 2019.

[10] Anning Hu. Heterogeneous treatment effects analysis for social scientists: A review. *Social Science Research*, 109:102810, 2023.

[11] David AA Baranger, Megan C Finsaas, Brandon L Goldstein, Colin E Vize, Donald R Lynam, and Thomas M Olino. Tutorial: Power analyses for interaction effects in cross-sectional regressions. *Advances in Methods and Practices in Psychological Science*, 6(3):25152459231187531, 2023.

[12] Mark Petticrew, Peter Tugwell, Elizabeth Kristjansson, Sandy Oliver, Erin Ueffing, and Vivian Welch. Damned if you do, damned if you don't: subgroup analysis and equity. *J Epidemiol Community Health*, 66(1):95–98, 2012.

[13] Peter M Rothwell. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186, 2005.

[14] Jonathan MV Davis and Sara B Heller. Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550, 2017.

[15] Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

[16] David I Cook, Val J Gebski, and Anthony C Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289, 2004.

[17] Ewout W Steyerberg and EW Steyerberg. *Applications of prediction models*. Springer, 2009.

[18] Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

[19] Jan-Michael Becker, Arun Rai, Christian M Ringle, and Franziska Völckner. Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS quarterly*, pages 665–694, 2013.

[20] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[21] Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

[22] Xuejing Jin, Fatima Al Sayah, Arto Ohinmaa, Deborah A Marshall, and Jeffrey A Johnson. Responsiveness of the eq-5d-3l and eq-5d-5l in patients following total hip or knee replacement. *Quality of Life Research*, 28:2409–2417, 2019.

[23] Carl Bonander and Mikael Svensson. Using causal forests to assess heterogeneity in cost-effectiveness analysis. *Health Economics*, 30(8):1818–1832, 2021.

[24] Zia Sadique, Richard Grieve, Karla Diaz-Ordaz, Paul Mouncey, Francois Lamontagne, and Stephen O'Neill. A machine-learning approach for estimating subgroup-and individual-level treatment effects: an illustration using the 65 trial. *Medical Decision Making*, 42(7):923–936, 2022.

[25] May E Png, Stavros Petrou, Miguel A Fernandez, Juul Achten, Nicholas Parsons, Alwin McGibbon, Jenny Gould, Xavier L Griffin, Matthew L Costa, et al. Cost-utility analysis of cemented hemiarthroplasty versus hydroxyapatite-coated uncemented hemiarthroplasty for the treatment of displaced intracapsular hip fractures: the world hip trauma evaluation 5 (white 5) trial. *The Bone & Joint Journal*, 104(8):922–928, 2022.

[26] D Metcalfe, ML Costa, NR Parsons, J Achten, J Masters, ME Png, SE Lamb, and XL Griffin. Validation of a prospective cohort study of older adults with hip fractures. *The bone & joint journal*, 101(6):708–714, 2019.

[27] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[28] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

[29] Susanne Dandl, Andreas Bender, and Torsten Hothorn. Heterogeneous treatment effect estimation for observational data using model-based forests. *arXiv preprint arXiv:2210.02836*, 2022.

[30] Julie Tibshirani, Susan Athey, Stefan Wager, R Friedberg, L Miner, and M Wright. grf: Generalized random forests. *R package version*, 1(0):7–3, 2020.

[31] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[32] Noemi Kreif, Karla DiazOrdaz, Rodrigo Moreno-Serra, Andrew Mirelman, Taufik Hidayat, and Marc Suhrcke. Estimating heterogeneous policy impacts using causal machine learning: a case study of health insurance reform in indonesia. *Health Services and Outcomes Research Methodology*, pages 1–36, 2021.

[33] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[34] Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[35] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[36] Susanne Dandl, Torsten Hothorn, Heidi Seibold, Erik Sverdrup, Stefan Wager, and Achim Zeileis. What makes forest-based heterogeneous treatment effect estimators work? *arXiv preprint arXiv:2206.10323*, 2022.

[37] Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5(2):37–51, 2019.

[38] Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56, 2010.

[39] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

[40] Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv preprint arXiv:2111.07966*, 2021.

Table 1: **Baseline Characteristics**

Covariate	Uncemented (N=484)	Cemented (N=472)	P-value	Overall (N=956)
Age				
Mean (SD)	84.7 (7.40)	85.0 (7.56)	0.544	84.9 (7.48)
Median [Min, Max]	86.0 [62.0, 101]	86.0 [61.0, 103]		86.0 [61.0, 103]
EQindex0				
Mean (SD)	0.569 (0.308)	0.593 (0.299)	0.223	0.581 (0.304)
Median [Min, Max]	0.636 [-0.386, 0.989]	0.650 [-0.272, 0.989]		0.640 [-0.386, 0.989]
EQ-5D VAS score				
Mean (SD)	62.6 (21.2)	61.7 (21.2)	0.523	62.2 (21.2)
Median [Min, Max]	65.0 [1.00, 100]	60.0 [0, 100]		60.0 [0, 100]
Proxy Consent n (%)	227 (46.9%)	195 (41.3%)	0.104	422 (44.1%)
No	238 (49.2%)	256 (54.2%)		494 (51.7%)
Yes	19 (3.9%)	21 (4.4%)		40 (4.2%)
Gender				
Male	326 (67.4%)	325 (68.9%)	0.668	651 (68.1%)
Female	158 (32.6%)	147 (31.1%)		305 (31.9%)
Current Smoker n (%)	450 (93.0%)	424 (89.8%)	0.064	874 (91.4%)
No	32 (6.6%)	48 (10.2%)		80 (8.4%)
Yes	2 (0.4%)	0 (0%)		2 (0.2%)
Chronic renal failure n (%)	440 (90.9%)	433 (91.7%)	0.808	873 (91.3%)
No	42 (8.7%)	38 (8.1%)		80 (8.4%)
Yes	2 (0.4%)	1 (0.2%)		3 (0.3%)
Diabetes n (%)	399 (82.4%)	388 (82.2%)	0.882	787 (82.3%)
No	82 (16.9%)	83 (17.6%)		165 (17.3%)
Yes	3 (0.6%)	3 (0.6%)		4 (0.4%)
Alcohol consumption n (%)	442 (91.3%)	427 (90.5%)	0.616	869 (90.9%)
0-7 units/wk	20 (4.1%)	25 (5.3%)		45 (4.7%)
8-14 units/wk	9 (1.9%)	7 (1.5%)		16 (1.7%)
15-21 units/wk	8 (1.7%)	12 (2.5%)		20 (2.1%)
$>$21 units/wk	5 (1.0%)	1 (0.2%)		6 (0.6%)
Residence status before injury n (%)	354 (73.1%)	367 (77.8%)	0.116	721 (75.4%)
Own home/sheltered housing	59 (12.2%)	57 (12.1%)		116 (12.1%)
Residential care	69 (14.3%)	47 (10.0%)		116 (12.1%)
Nursing Care	2 (0.4%)	1 (0.2%)		3 (0.3%)

Notes:

Table 2: **Outcomes Characteristics**

Outcome	Uncemented (N=595)	Cemented (N=592)	P-value	Overall (N=1187)
EQindex Baseline				
Mean (SD)	0.569 (0.308)	0.593 (0.299)	0.223	0.581 (0.304)
Median [Min, Max]	0.636 [-0.386, 0.989]	0.650 [-0.272, 0.989]		0.640 [-0.386, 0.989]
Missing	111 (18.7%)	120 (20.3%)		231 (19.5%)
EQindex Month 1				
Mean (SD)	0.337 (0.349)	0.401 (0.351)	0.009	0.369 (0.351)
Median [Min, Max]	0.418 [-0.415, 0.988]	0.527 [-0.409, 0.989]		0.481 [-0.415, 0.989]
Missing	187 (31.4%)	173 (29.2%)		360 (30.3%)
EQindex Month 4				
Mean (SD)	0.405 (0.365)	0.449 (0.368)	0.148	0.427 (0.367)
Median [Min, Max]	0.521 [-0.529, 0.989]	0.573 [-0.387, 0.989]		0.550 [-0.529, 0.989]
Missing	310 (52.1%)	298 (50.3%)		608 (51.2%)
EQindex Month 12				
Mean (SD)	0.421 (0.367)	0.454 (0.339)	0.492	0.439 (0.352)
Median [Min, Max]	0.523 [-0.293, 0.988]	0.581 [-0.272, 0.989]		0.544 [-0.293, 0.989]
Missing	491 (82.5%)	475 (80.2%)		966 (81.4%)
EQ5DVAS Month 1				
Mean (SD)	56.5 (23.6)	59.0 (22.9)	0.112	57.8 (23.3)
Median [Min, Max]	60.0 [0, 100]	60.0 [0, 100]		60.0 [0, 100]
Missing	182 (30.6%)	173 (29.2%)		355 (29.9%)
EQ5DVAS Month 4				
Mean (SD)	57.9 (23.2)	60.2 (23.1)	0.232	0.591 (23.2)
(SD)	57.9 (23.2)	60.2 (23.1)	0.232	59.1 (23.2)
Median [Min, Max]	60.0 [0, 100]	65.0 [0, 100]		60.0 [0, 100]
Missing	308 (51.8%)	299 (50.5%)		607 (51.1%)
EQ5DVAS Month 12				
Mean (SD)	60.5 (23.1)	60.5 (23.9)	0.983	60.5 (23.4)
Median [Min, Max]	60.0 [0, 100]	65.0 [8.00, 100]		60.0 [0, 100]
Missing	489 (82.2%)	477 (80.6%)		966 (81.4%)

Notes:

Table 3: **Covariates Month 1**

Covariate	Uncemented (N=408)	Cemented (N=419)	P-value	Overall (N=827)
Age				
Mean (SD)	84.5 (7.56)	84.7 (7.55)	0.665	84.6 (7.55)
Median [Min, Max]	86.0 [62.0, 101]	85.0 [61.0, 102]		86.0 [61.0, 102]
EQindex1				
Mean (SD)	0.573 (0.307)	0.605 (0.296)	0.132	0.589 (0.302)
Median [Min, Max]	0.633 [-0.386, 0.989]	0.670 [-0.272, 0.989]		0.650 [-0.386, 0.989]
EQ-5D VAS score	20 (4.9%)	26 (6.2%)		46 (5.6%)
Mean (SD)	63.4 (21.0)	62.8 (21.3)	0.688	63.1 (21.1)
Median [Min, Max]	65.0 [1.00, 100]	60.0 [0, 100]		65.0 [0, 100]
Proxy Consent n (%)	182 (44.6%)	156 (37.2%)	0.076	338 (40.9%)
No	192 (47.1%)	216 (51.6%)		408 (49.3%)
Yes	34 (8.3%)	47 (11.2%)		81 (9.8%)
Gender				
Male	273 (66.9%)	300 (71.6%)	0.166	573 (69.3%)
Female	135 (33.1%)	119 (28.4%)		254 (30.7%)
Current Smoker n (%)	377 (92.4%)	375 (89.5%)	0.11	752 (90.9%)
No	24 (5.9%)	38 (9.1%)		62 (7.5%)
Yes	7 (1.7%)	6 (1.4%)		13 (1.6%)
Chronic renal failure n (%)	373 (91.4%)	389 (92.8%)	0.888	762 (92.1%)
No	28 (6.9%)	27 (6.4%)		55 (6.7%)
Yes	7 (1.7%)	3 (0.7%)		10 (1.2%)
Diabetes n (%)	336 (82.4%)	344 (82.1%)	0.74	680 (82.2%)
No	64 (15.7%)	71 (16.9%)		135 (16.3%)
Yes	8 (2.0%)	4 (1.0%)		12 (1.5%)
Alcohol consumption n (%)	364 (89.2%)	371 (88.5%)	0.792	735 (88.9%)
0-7 units/wk	18 (4.4%)	21 (5.0%)		39 (4.7%)
8-14 units/wk	9 (2.2%)	8 (1.9%)		17 (2.1%)
15-21 units/wk	7 (1.7%)	11 (2.6%)		18 (2.2%)
$>$21 units/wk	10 (2.5%)	8 (1.9%)		18 (2.2%)
Residence status before injury n (%)				
Own home/sheltered housing	299 (73.3%)	330 (78.8%)	0.172	629 (76.1%)
Residential care	50 (12.3%)	43 (10.3%)		93 (11.2%)
Nursing Care	59 (14.5%)	46 (11.0%)		105 (12.7%)

Notes:

Table 4: **Covariates Month 4**

Covariate	Uncemented (N=285)	Cemented (N=294)	P-value	Overall (N=579)
Age				
Mean (SD)	84.0 (7.82)	84.3 (7.83)	0.744	84.2 (7.82)
Median [Min, Max]	85.0 [62.0, 101]	85.0 [61.0, 102]		85.0 [61.0, 102]
EQindex0				
Mean (SD)	0.580 (0.301)	0.601 (0.294)	0.414	0.591 (0.297)
Median [Min, Max]	0.636 [-0.343, 0.989]	0.651 [-0.272, 0.989]		0.644 [-0.343, 0.989]
EQ-5D VAS score				
Mean (SD)	64.7 (20.6)	64.2 (20.5)	0.813	64.4 (20.5)
Median [Min, Max]	70.0 [2.00, 100]	65.0 [0, 100]		67.0 [0, 100]
Proxy Consent n (%)	115 (40.4%)	100 (34.0%)	0.127	215 (37.1%)
No	135 (47.4%)	157 (53.4%)		292 (50.4%)
Yes	35 (12.3%)	37 (12.6%)		72 (12.4%)
Gender				
Male	192 (67.4%)	219 (74.5%)	0.072	411 (71.0%)
Female	93 (32.6%)	75 (25.5%)		168 (29.0%)
Current Smoker n (%)	257 (90.2%)	259 (88.1%)	0.292	516 (89.1%)
No	20 (7.0%)	29 (9.9%)		49 (8.5%)
Yes	8 (2.8%)	6 (2.0%)		14 (2.4%)
Chronic renal failure n (%)	258 (90.5%)	271 (92.2%)	0.771	529 (91.4%)
No	21 (7.4%)	19 (6.5%)		40 (6.9%)
Yes	6 (2.1%)	4 (1.4%)		10 (1.7%)
Diabetes n (%)	232 (81.4%)	245 (83.3%)	0.752	477 (82.4%)
No	47 (16.5%)	45 (15.3%)		92 (15.9%)
Yes	6 (2.1%)	4 (1.4%)		10 (1.7%)
Alcohol consumption n (%)	252 (88.4%)	254 (86.4%)	0.425	506 (87.4%)
0-7 units/wk	216 (5.6%)	18 (6.1%)		34 (5.9%)
8-14 units/wk	3 (1.1%)	6 (2.0%)		9 (1.6%)
15-21 units/wk	4 (1.4%)	9 (3.1%)		13 (2.2%)
$>$21 units/wk	10 (3.5%)	7 (2.4%)		17 (2.9%)
Residence status before injury n (%)				
Own home/sheltered housing	217 (76.1%)	235 (79.9%)	0.486	452 (78.1%)
Residential care	34 (11.9%)	32 (10.9%)		66 (11.4%)
Nursing Care	34 (11.9%)	27 (9.2%)		61 (10.5%)

Notes:

Table 5: **Covariates Month 12**

Covariate	Uncemented (N=104)	Cemented (N=117)	P-value	Overall (N=221)
Age				
Mean (SD)	83.5 (7.84)	84.4 (8.04)	0.394	83.9 (7.95)
Median [Min, Max]	85.0 [62.0, 101]	85.0 [65.0, 100]		85.0 [62.0, 101]
EQindex0				
Mean (SD)	0.579 (0.300)	0.589 (0.303)	0.828	0.584 (0.301)
Median [Min, Max]	0.617 [-0.266, 0.989]	0.654 [-0.135, 0.989]		0.633 [-0.266, 0.989]
EQ-5D VAS score				
Mean (SD)	62.9 (19.6)	64.5 (22.3)	0.591	63.8 (21.0)
Median [Min, Max]	65.0 [10.0, 100]	70.0 [2.00, 100]		65.0 [2.00, 100]
Proxy Consent n (%)	45 (43.3%)	28 (23.9%)	0.004	73 (33.0%)
No	45 (43.3%)	70 (59.8%)		115 (52.0%)
Yes	14 (13.5%)	19 (16.2%)		33 (14.9%)
Gender				
Male	70 (67.3%)	86 (73.5%)	0.389	156 (70.6%)
Female	34 (32.7%)	31 (26.5%)		65 (29.4%)
Current Smoker n (%)	98 (94.2%)	101 (86.3%)	0.074	199 (90.0%)
No	4 (3.8%)	13 (11.1%)		17 (7.7%)
Yes	2 (1.9%)	3 (2.6%)		5 (2.3%)
Chronic renal failure n (%)	94 (90.4%)	113 (96.6%)	0.041	207 (93.7%)
No	9 (8.7%)	2 (1.7%)		11 (5.0%)
Yes	1 (1.0%)	2 (1.7%)		3 (1.4%)
Diabetes n (%)	89 (85.6%)	101 (86.3%)	0.913	190 (86.0%)
No	14 (13.5%)	14 (12.0%)		28 (12.7%)
Yes	1 (1.0%)	2 (1.7%)		3 (1.4%)
Alcohol consumption n (%)	95 (91.3%)	99 (84.6%)	0.272	194 (87.8%)
0-7 units/wk	4 (3.8%)	9 (7.7%)		13 (5.9%)
8-14 units/wk	0 (0%)	2 (1.7%)		2 (0.9%)
15-21 units/wk	2 (1.9%)	4 (3.4%)		6 (2.7%)
$>$21 units/wk	3 (2.9%)	3 (2.6%)		6 (2.7%)
Residence status before injury n (%)				
Own home/sheltered housing	84 (80.8%)	98 (83.8%)	0.441	182 (82.4%)
Residential care	9 (8.7%)	12 (10.3%)		21 (9.5%)
Nursing Care	11 (10.6%)	7 (6.0%)		18 (8.1%)

Notes:

Figure 1: Caterpillar Plots.This graph shows the individualized effect

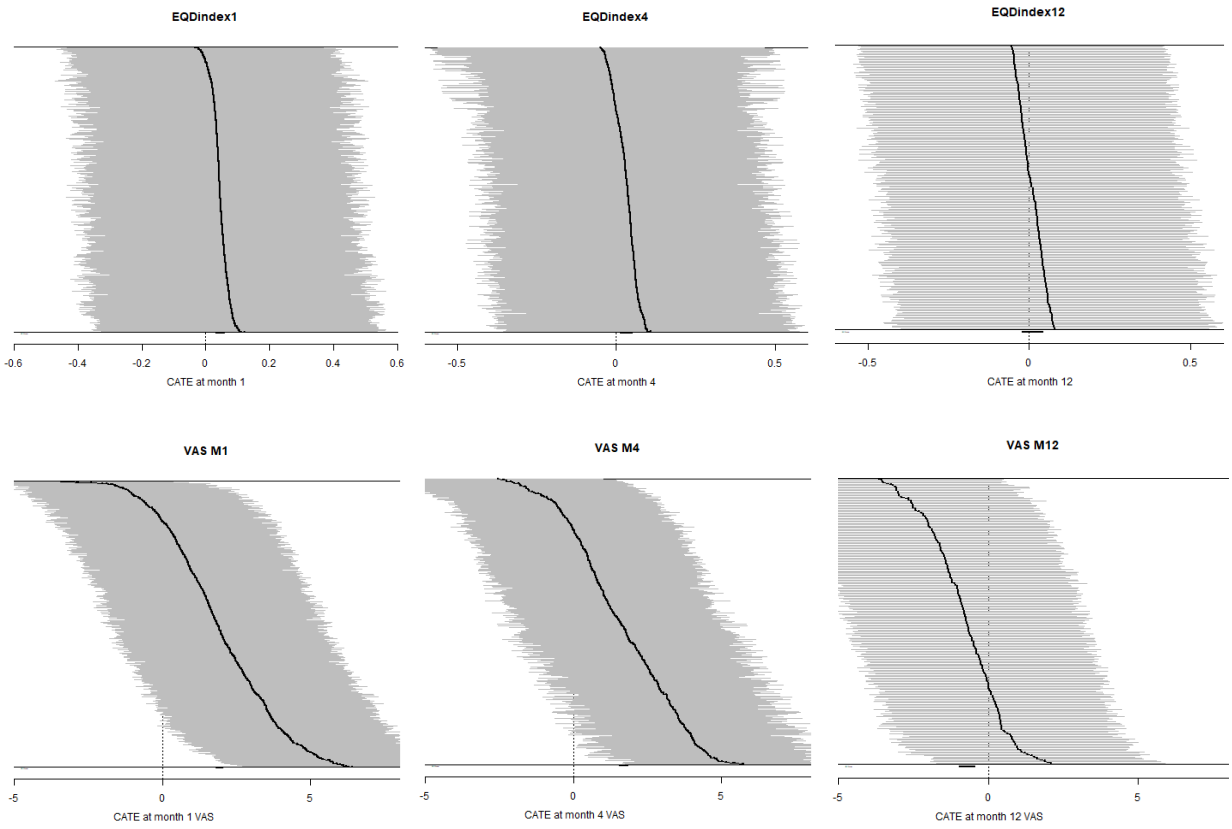
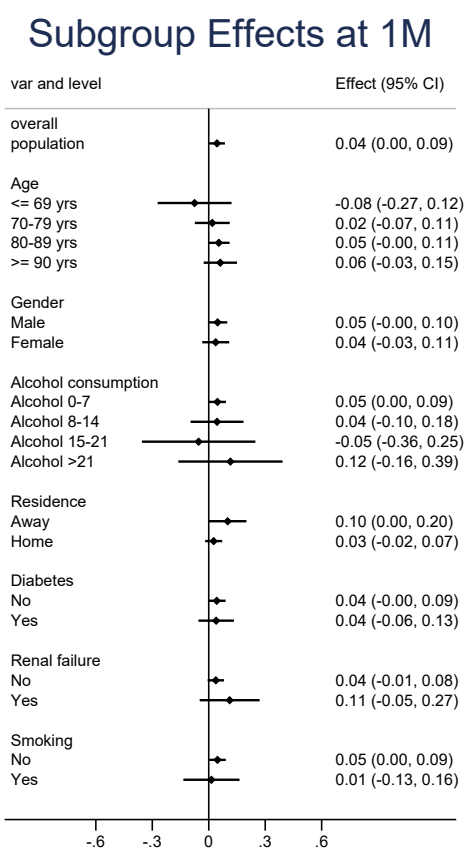
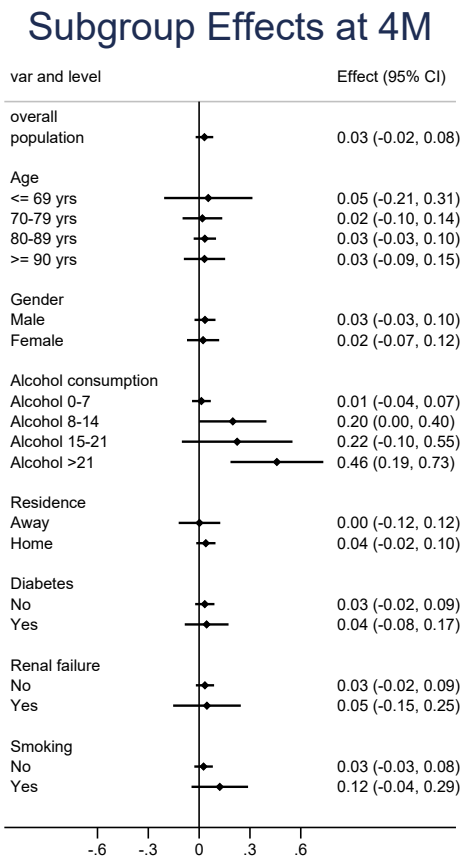


Figure 2: Graphs showing the conditional treatment effects on EQ-5D Index at Months 1, 4, and 12 for the subgroups.





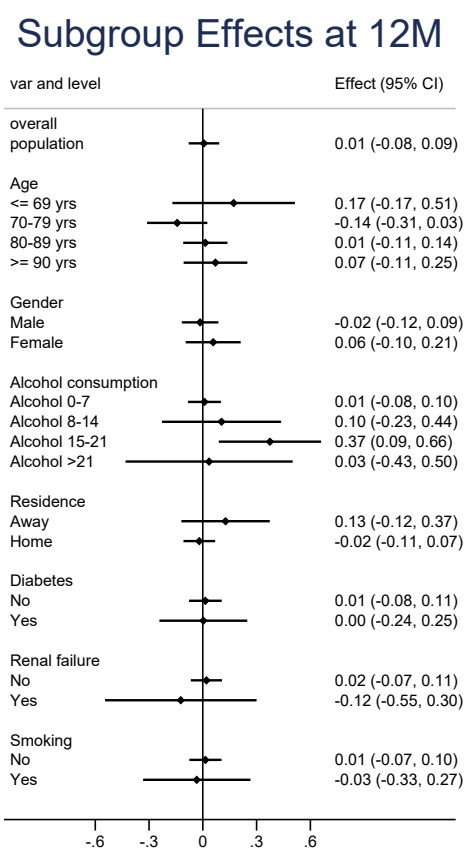
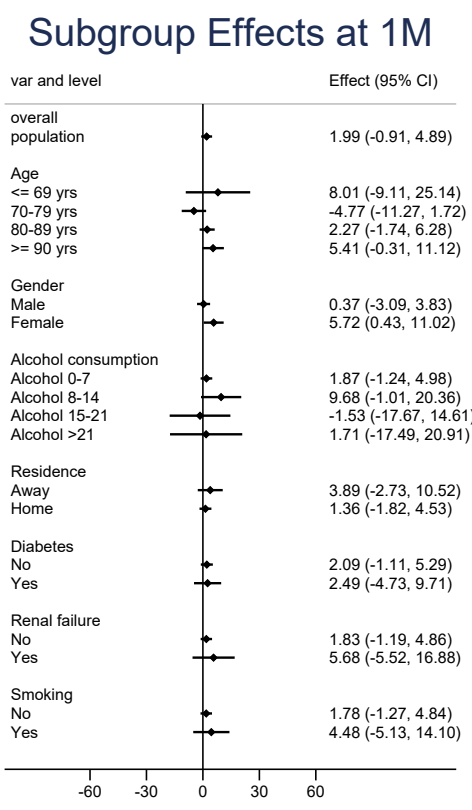
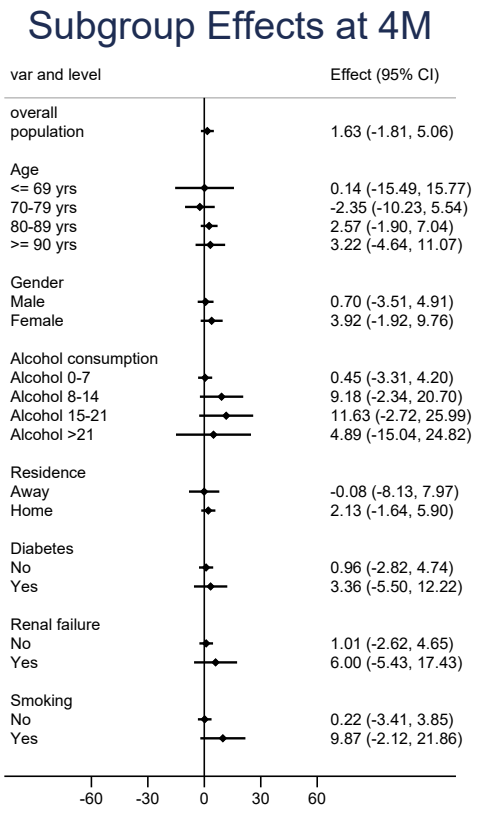
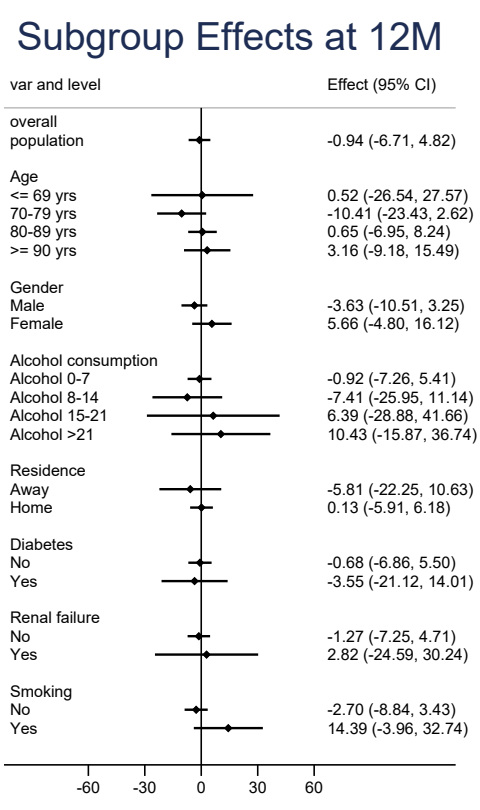


Figure 3: Graphs graphs showing the conditional treatment effects on EQ-5D VAS at Months 1, 4, and 12 for subgroups.







A Causal Forest Method Details

For each patient $i = 1, \ldots, 1187$ in the dataset, we observe a binary treatment indicator D_i ($1 =$ *cemented hemiarthroplasty*, $0 =$ *uncemented hemiarthroplasty*), a matrix of baseline characteristics of patient i denoted by X_i that includes the 20 covariates that may act as treatment modifiers (see Table1) and a set of outcomes, $Y_{i,j}$, where j indexes the outcomes. The outcomes are listed in Table .... We consider a generic outcome Y_i for the methods' description. We describe the method using the Neyman-Rubin potential outcomes framework[33,34].

Theoretically, two potential outcomes are possible for each patient i: $Y_i(0)$ corresponding to the scenario where patient i is assigned to the uncemented hemiarthroplasty group, and $Y_i(1)$ signifying the outcome had patient i been assigned to the cemented group. However, the fundamental problem of causal inference[35] manifests since at most one of the two potential outcomes is ever observed for each patient. The observed outcome Y_i can be represented as $Y_i = D_i \times Y_i(1) + (1 - D_i) \times Y_i(0)$, and the effect of the intervention on the outcome for patient i will be $\tau_i = Y_i(1) - Y_i(0)$.

In this study, the estimands of interest can be obtained by aggregating the τ_i's: the Average Treatment Effect (ATE) quantifies the overall effect on the population, and the Conditional Average Treatment Effects (CATE) quantifies the average patient-level effect given their baseline characteristics $(X_i = x)$, which can then be aggregated for subgroups of interest.

$$ATE = E(\tau_i) \tag{1}$$

$$CATE(x) = \tau_i(x) = E(\tau_i | X_i = x) \tag{2}$$

When incorporating the covariates X into the model, a reformulation of the observed outcome Y can be expressed as follows[36].

$$Y_i = \mu_i(X) + D_i \times \tau_i(X) + E \tag{3}$$

Where $\mu_i(X)$ represents the prognostic effect that is resulted from the impact of a subset of covariates X, while the subset of treatment moderators are included in $\tau_i(X)$. If the treatment assignment is assumed to be non-deterministic, the conditional mean of Y will be represented as[36]:

$$E(Y_i | X_i = x) = \mu_i(x) + e_i(x) \times \tau_i(x) = m_i(x) \tag{4}$$

where $e_i(x)$ is the propensity score that is estimated by regressing the treatment on the covariates, and $m_i(x)$ is referred to as the marginal mean.

A.1 Causal forest

To estimate the $CATE(x)$, we apply the Causal Forest method[20], which is a generalization of the random forest of Breiman[27] to the estimation of treatment effects. Athey & Imbens[21] modified the classification and regression tree (CART) prediction approach to construct a 'causal tree' which focuses on estimating the expected conditional treatment effects, $\tau_i(x)$, rather than predicting the outcome (Y_i), as is done in a traditional CART. To achieve this, equation (3) is rewritten as[20]:

$$\begin{aligned}(Y_i | X_i = x) &= m_i(x) - m_i(x) + \mu_i(X) + D_i \times \tau_i(X) + E \\ &= m_i(x) + \tau_i(X)(D_i - e_i(x)) + E\end{aligned} \tag{5}$$

This representation enables the estimation of the treatment effects $\hat{\tau}_i(\boldsymbol{x})$ through a two-step process initiating by regression of outcome and treatment on covariates to obtain estimates of marginal mean $\hat{m}_i(\boldsymbol{x})$ and the propensity $\hat{e}_i(\boldsymbol{x})$, respectively. Subsequently, the estimates of interest $\hat{\tau}_i(\boldsymbol{x})$ are derived by selecting $\hat{\tau}_i(\mathbf{X})$ which minimizes the loss function as defined by Equation (6)[36]:

$$\frac{1}{2}\left[Y_i - \hat{m}_i(x) - \hat{\tau}_i(x)(D_i - \hat{e}_i(x))\right]^2 \tag{6}$$

This local centering algorithm enhances the model's robustness to potential confounding effects[28].

Furthermore, an 'honest' estimation is implemented where partitioning and estimating the effects are conducted on distinct subsamples to prevent overfitting and provide correct inference. That is, the splitting criterion of the causal tree aims to minimize the expected mean squared error (EMSE) of the treatment effects, is defined as[21]:

$$\begin{aligned} -\widehat{\mathsf{EMSE}}_\tau(S^{tr}, N^{est}, T) = \frac{1}{N^{tr}} \sum_{i \in S^{tr}} \left(\hat{\tau}^2(X_i|S^{tr}, T)\right) \\ - \left(\frac{1}{N^{tr}} + \frac{1}{N^{est}}\right) \sum_{L \in T} \left(\frac{S^2_{S^{tr}_{\mathsf{cemented}}(L)}}{p} + \frac{S^2_{S^{tr}_{\mathsf{uncemented}}(L)}}{1-p}\right) \end{aligned} \tag{7}$$

where, $\mathcal{S}^{\mathsf{train}}$ is the training subsample that is used to construct the tree T, $\mathcal{S}^{\mathsf{est}}$ is the estimation subsample which is different from the training subsample, N^{est} is the number of patients in the estimation sample, N^{tr} is the number of patients in the training subsample, L is a 'leaf' (i.e. a subgroup defined by the splits) in tree T, $S^2_{S^{tr}_{\mathsf{cemented}}(l)}$ and $S^2_{S^{tr}_{\mathsf{uncemented}}(l)}$ are the within-leaf variances of outcomes for the patients at the two treatment arms, and p is the marginal treatment probability $P(D_i = 1)$ which is constant and does not depend on $\mathbf{X_i}$ in fully randomized experiments such as the WHITE 5 trial considered here.

This splitting criterion is constructed to prefer leaves exhibiting heterogeneous effects by maximizing the first term of equation (7), and simultaneously, leaves with a good fit by minimizing the within-leaf variance. However, an individual tree can be too noisy. To overcome this, Wager & Athey (2018)[20] proposed the CF which generates an ensemble of B causal trees, each of which produces an estimate $\hat{\tau}_b(X)$, which are then aggregated to obtain a CATE estimate, $\hat{\tau}(X)$. The $\hat{\tau}_i(X)$ estimates are estimated using an adaptive locally weighted estimator[37] such that:

$$\hat{\tau}_i(x) = \frac{\sum_{i=1}^n \alpha_i(x)(Y_i - \hat{m}^{(-i)}(X_i))(D_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^n \alpha_i(x)(D_i - \hat{e}^{(-i)}(X_i))^2} \tag{8}$$

where the superscript $(-i)$ denotes the out-of-bag predictions which are obtained from the subsample of trees where observation i was not used to determine the splits, $\hat{m}(x)$ is the estimated conditional mean outcome $E[Y_i|X_i = x]$ obtained by fitting a regression forest, $\hat{e}(x)$ is the estimated conditional propensity score $P[D_i = 1|X_i = x]$ obtained by fitting a binary regression forest, and $\hat{\alpha}(x)$ is the weight given to observation i which measures how often observation i is assigned to the same leaf that the point (x) lies within[37]. This method is implemented in the generalized random forest R package *grf*[30]. We estimate CATEs for our pre-specified subgroups by taking the estimated patient-level treatment effects and plugging them into an augmented inverse propensity weighting AIPW estimator[31] of group average treatment effects[32].

A.2 AIPW estimator

The strength of the AIPW estimator[31] stems from its double robustness property which means that the estimates of the average treatment effects of the population and the subgroups remain consistent even if one of the propensity or outcome regression forests is miss specified[38]. Glynn and Quinn citep glynn2010 provided a theoretical and experimental evidence of its superiority over other estimators such as: regression estimator, inverse propensity weighted (IPW) estimator, and propensity score matching estimator.

In our study, the AIPW scores that are averaged to obtain the ATE and CATE estimates are obtained using the following formula[39]:

$$\hat{\gamma}_i = \hat{m}_i(X_i, 1) - \hat{m}_i(X_i, 0) + \frac{(Y_i - \hat{m}_i(X_i, D_i))(D_i - \hat{e}(X_i))}{\hat{e}(X_i)(1 - \hat{e}(X_i))} \tag{9}$$

where $\hat{m}_i(x, d) = \mathbb{E}[Y_i(d)|X_i = x]$ denotes the nonparametric estimate of the conditional mean of the treatment group.

B Application of CF Approaches to Estimate Group ATEs in the WHiTE 5 Trial

We implement the CF for each outcome using 20,000 trees. This number of trees is large enough to make the perturbation error - which results from fitting different forests – negligible to the variances of the estimated CATEs[30]. All other tuning hyperparameters (sample fraction used to build each tree, number of variables tried for each split, minimum number of individuals in each tree leaf, honesty fraction, and parameters which determine the imbalance of the splits) are determined using cross-validation.

The forests were fitted in two stages[24]. During the first stage, the model is fitted over all covariates. The second stage considers only the most important covariates, i.e., those whose importance exceeds 20% of the average importance (see Figure B.1), where importance is defined as the simple weighted sum of how many times each covariate was used to determine the sample split at each depth in the forest[37]. Then, we regressed the estimated CATEs on the most important covariates, and obtained the estimates of best linear projection along with coefficient standard errors (see Figure B.1).

To test for heterogeneity, omnibus heterogeneity tests were performed, and their results are presented in the supplement material (see Table B.1). This test yields two parameters: ATE parameter to test the null hypothesis of good calibration of the ATE, where a value of 1 indicates a correct mean forest. The second parameter is the Heterogeneity parameter, also with a value of 1 indicating well-calibrated estimates of heterogeneity within the forest. If the Heterogeneity parameter is positive, its associated p-value indicates the strength of evidence supporting the null hypothesis of no heterogeneity[37]. However, the calibration tests indicate the absence of heterogeneity, since the heterogeneity parameter is negative for the six outcomes.

Furthermore, we applied the Rank-Weighted Average Treatment Effect (RATE) metric proposed by[40] to test to examine the presence of substantial heterogeneity, and to assess the strength of our CATE estimates are at distinguishing subpopulations with different treatment effects. Particularly, we aim to measure the benefit there is to prioritizing cemented therapy provision based on the heterogeneity that is identified by our causal forest. This approach assigns, based on the estimated CATEs, a higher score to patients estimated to benefit more from cemented therapy and a lower score to those with lower benefit compared to uncemented one. The benefit refers to the expected increase in outcomes when providing the cemented therapy to a fraction of the population with the highest prioritization scores as opposed to giving the therapy to a randomly selected fraction of the same size. The figures (see Figure B.1) depict the Target Operator Characteristic (TOC) curves on the outcomes. These curves chop the population up into groups defined by above mentioned scores, then plot this over all groups where each group is the top q-th fraction of patients with the largest score.

Table B.1: **Calibration Tests**

Outcome	ATE parameter	Standard error	*P*-val	Heterogeneity parameter	Standard error	*P*-val
EQ-5D Index Month 1	1.01	0.479	0.02	-1.61	-1.76	0.96
EQ-5D VAS Month 1	1.01	0.756	0.09	-1.14	0.828	0.92
EQ-5D Index Month 4	1.01	0.844	0.12	-0.74	0.760	0.84
EQ-5D VAS Month 4	1.04	1.174	0.19	-1.69	0.979	0.96
EQ-5D Index Month 12	1.26	4.789	0.40	-1.50	1.052	0.92
EQ-5D VAS Month 12	0.94	2.186	0.33	-18.27	2.203	1.00

Notes:

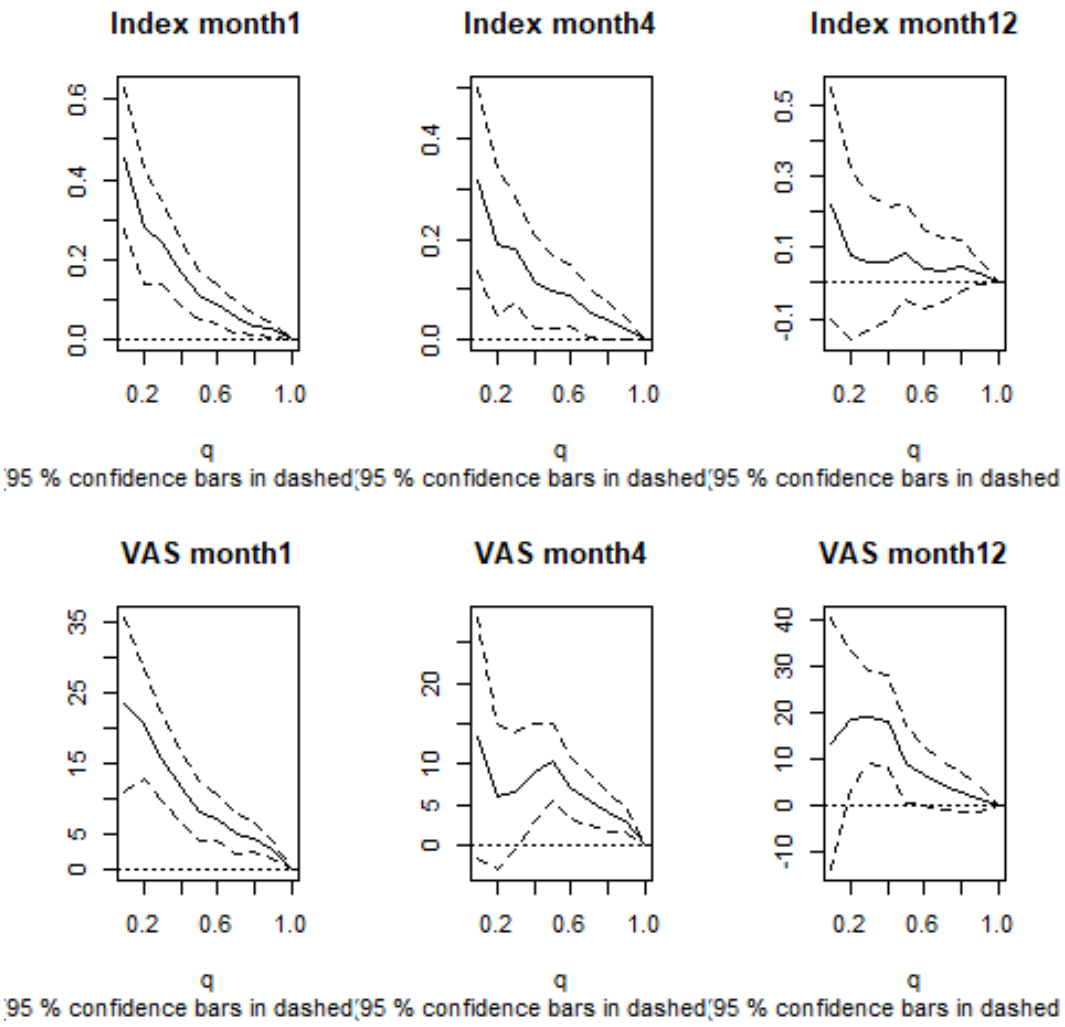


Figure B.1: CF Approaches to Estimate Group ATEs