

Your Title

YOUR NAME

2025-05-23

Abstract

Background: (Brief few sentences) **Objectives:** 1. Estimate the causal effect of YOUR EXPOSURE on YOUR OUTCOMES measured one year later. 2. Evaluate whether these effects vary across the population. 3. Provide policy guidance on which individuals might benefit most. **Method:** We conducted a three-wave retrospective cohort study (waves XX-XXX, October XXXX–October XXXX) using data *SIMULATED* from the New Zealand Attitudes and Values Study, a nationally representative panel. Participants were eligible if they participated in the NZAVS in the baseline wave (XXXX,...). We defined the exposure as (XXXX > NUMBER on a 1-7 Likert Scale (1 = yes, 0 = no)). To address attrition, we applied inverse probability of censoring weights; to improve external validity, we applied weights to the population distribution of Age, Ethnicity, and Gender. We computed expected mean outcomes for the population in each exposure condition (high XXXX/low XXXXX). Under standard causal assumptions of unconfoundedness, the contrast provides an unbiased average treatment effect. We then used causal forests to detect heterogeneity in these effects and employed policy tree algorithms to identify individuals (“strong responders”) likely to experience the greatest benefits. **Results:** Increasing XXXXX leads to XXXXX. Heterogeneous responses to (e.g. *Forgiveness*, *Personal Well-Being*, and *Life-Satisfaction*...) reveal structural variability in subpopulations... **Implications:** (Brief few sentences) **Keywords:** *Causal Inference*; *Cross-validation*; *Distress*; *Employment*; *Longitudinal*; *Machine sLearning*; *Religion*; *Semi-parametric*; *Targeted Learning*.

Introduction

Your place to shine here

Method

Target Population

The data are simulated from the New Zealand Attitudes and Values Study. For the purposes of this assessment, the target population for this study comprises New Zealand residents as represented in the 2018 of the New Zealand Attitudes and Values Study (NZAVS) during the years 2018 weighted by New Zealand Census weights for age, gender, and ethnicity (refer to Sibley (2021)). The NZAVS is a national probability study designed to reflect the broader New Zealand population accurately. Despite its comprehensive scope, the NZAVS has some limitations in its demographic representation. Notably, it tends to under-sample males and individuals of Asian descent while over-sampling females and Māori (the indigenous peoples of New Zealand). To address these disparities and enhance the accuracy of our findings, we apply New Zealand Census survey weights to the sample data.

Eligibility Criteria

To be included in the analysis of this study, participants needed to participate in the 2018 of the study and respond to the baseline measure of Extraversion.

Participants may have been lost to follow-up at the end of the study if they met eligibility criteria at 2018. We adjusted for attrition and non-response using censoring weights, described below.

A total of 39,635 individuals met these criteria and were included in the study.

Average treatment effect

To learn how the outcome would shift if everyone received a different exposure, we emulate a **target trial** (Hernán & Robins, 2016). Making the hypothetical experiment explicit fixes the estimand, the data requirements, and the assumptions.

Our guiding question is:

How would the outcomes change if, for every individual, we set the exposure to >4, scale range 1-7 rather than <=4, scale range 1-7, conditional on their baseline characteristics?

We compare two interventions:

1. **1** – every participant is set to >4, scale range 1-7.
2. **0** – every participant is set to <=4, scale range 1-7.

The difference in population means defines the **average treatment effect (ATE)**. Figure Figure 1 plots the exposure distribution and its dichotomisation; the centre dashed line marks the mean, flanked by one standard deviation.

Because we test several outcomes, we adjust the ATE confidence intervals for multiplicity with bonferroni at $\alpha = 0.05$.

The longitudinal design and rich baseline covariates allow us – under the standard identification assumptions of consistency, positivity, and no unmeasured confounding – to attribute the differences between the two exposure means as a causal effect. Conditioning on demographics, personality traits, and other pretreatment factors renders exposure assignment ignorable (Rosenbaum & Rubin, 1983).

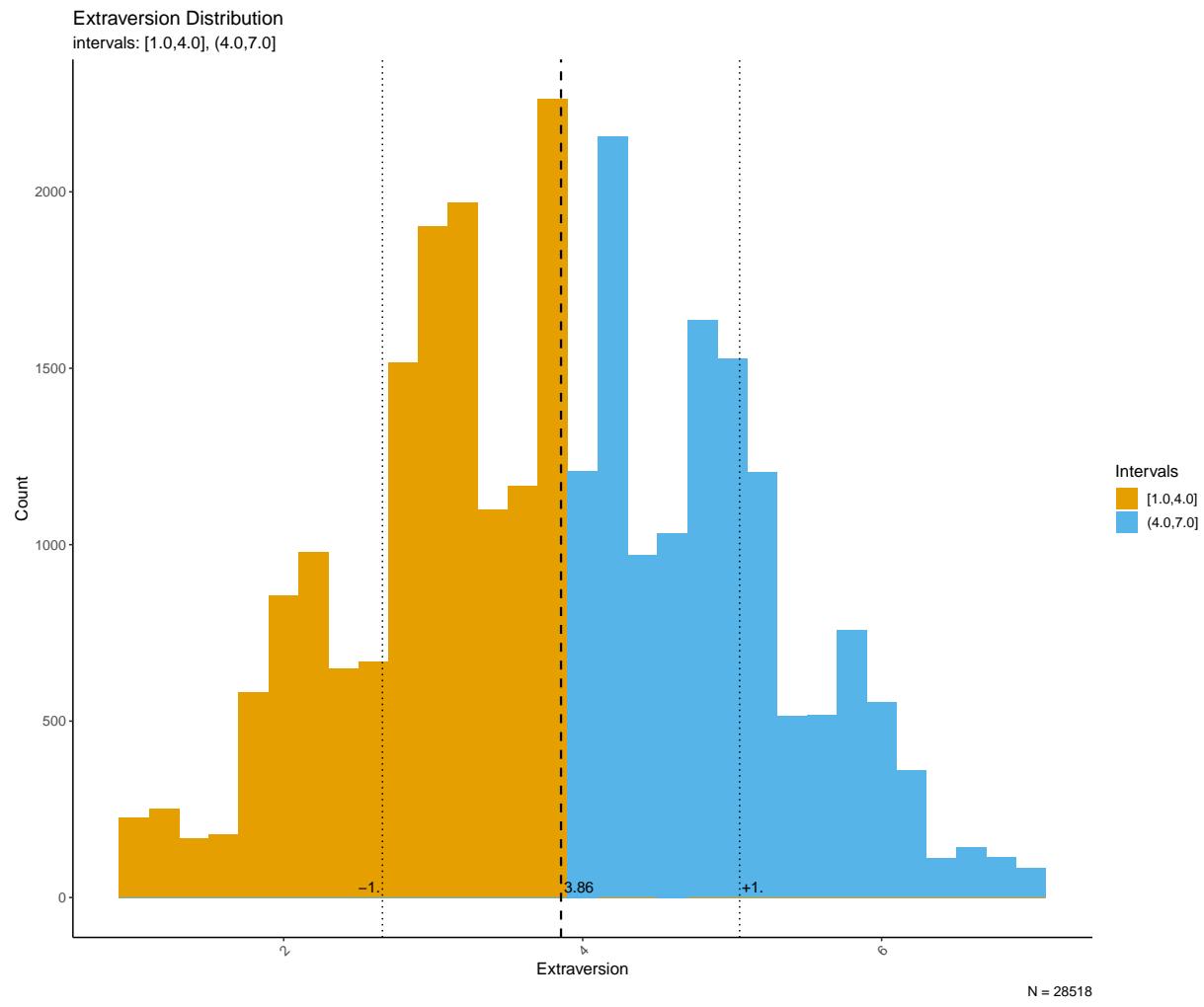


Figure 1: Histogram of exposure with binary groupin

Moderators and treatment policies

We pursued two complementary objectives: (i) to test whether personalised targeting, based on individual conditional average treatment effects $\hat{\tau}(x)$, yields welfare gains, and (ii) to convert any such gains into transparent, practitioner-ready decision rules.

Pre-processing and honest model training

Our protocol is as follows: in settings where the exposure is positive, outcomes for which lower is better are sign-flipped so that larger values always index improvement. Similarly, where the exposure is negative, outcomes for which ‘higher is better’ are sign flipped. Here, we flipped: Anxiety, Depression, Rumination.

Each model estimate used an honest 70/30 split: the training fold built the causal forest with grf ([Tibshirani et al., 2024](#)), while the held-out fold powered all diagnostic checks and provided data for fitting policy trees. This separation curbs over-fitting yet keeps the workflow simple.

Budget-based screening with Qini curves.

Before constructing rules we asked a budget question: If resources allow treatment of only the top 20% or 50% of individuals ranked by $\hat{\tau}(x)$, what uplift is purchased relative to treating everyone? Qini curves quantified the incremental gain; outcomes whose 95% confidence intervals excluded zero at either spend level were labelled actionable, signalling meaningful heterogeneity in benefit.

Deriving transparent decision rules.

For each actionable outcome we trained depth-2 policy trees with policytree ([Athey & Wager, 2021a; Sverdrup et al., 2024](#)) on the validation data. The resulting if–then statements maximise expected welfare under the same budget cap and remain auditable by domain experts.

Global heterogeneity tests (supplementary).

Appendix D reports RATE-AUTOC and RATE-Qini statistics, asking whether any covariate information can beat a uniform policy. These tests, controlled for false discovery at $q=0.1$ via Benjamini–Hochberg false-discovery-rate adjustment, are informative but not required for the budget-first pipeline.

Overall, combining Qini curves to screen and shallow policy trees to act isolates budget-relevant treatment heterogeneity and distils it into actionable rules, avoiding the chase for spurious complexity. Full technical details appear in Appendix E.

Exposure Indicator

The New Zealand Attitudes and Values Study assesses Extraversion using the following question:

Mini-IPIP6 Extraversion dimension: (i) I am the life of the party. (ii) I don’t talk a lot. (r) (iii) I keep in the background. (r) (iv) I talk to a lot of different people at parties. (Refer to [Appendix A](#)).

Causal Identification Assumptions

This study relies on the following identification assumptions for estimating the causal effect of Extraversion:

1. **Consistency:** the observed outcome under the observed Extraversion is equal to the potential outcome under that exposure level. As part of consistency, we assume no interference: the potential outcomes for one individual are not affected by the Extraversion status of other individuals.
2. **No unmeasured confounding:** all variables that affect both Extraversion and the outcome have been measured and accounted for in the analysis.

3. **Positivity:** there is a non-zero probability of receiving each level of Extraversion for every combination of values of Extraversion and confounders in the population. Positivity is the only fundamental causal assumption that can be evaluated with data (refer to Appendix C).

Confounding Control

To manage confounding in our analysis, we implement VanderWeele (2019)'s *modified disjunctive cause criterion* by following these steps:

1. **Identified all common causes** of both the treatment and outcomes.
2. **Excluded instrumental variables** that affect the exposure but not the outcome. Instrumental variables do not contribute to controlling confounding and can reduce the efficiency of the estimates.
3. **Included proxies for unmeasured confounders** affecting both exposure and outcome. According to the principles of d-separation Pearl (2009), using proxies allows us to control for their associated unmeasured confounders indirectly.
4. **Controlled for baseline exposure and baseline outcome.** Both are used as proxies for unmeasured common causes, enhancing the robustness of our causal estimates, refer to VanderWeele et al. (2020).

Statistical Estimation

We estimate heterogeneous treatment effects with Generalized Random Forests (GRF) (Tibshirani et al., 2024). GRF extends random forests for causal inference by focusing on conditional average treatment effects (CATE). It handles complex interactions and non-linearities without explicit model specification, and it provides ‘honest’ estimates by splitting data between model-fitting and inference. GRF is doubly robust because it remains consistent if either the outcome model or the propensity model is correct. We evaluate policies with the policytree package (Athey & Wager, 2021a; Sverdrup et al., 2024) and visualise results with margot (Bulbulia, 2024a). (Refer to Appendix E for a detailed explanation of our approach.)

Missing Data

The GRF package accepts missing values at baseline. To obtain valid inference for missing responses we computed inverse probability of censoring weights for censoring of the exposure, given that systematic censoring following the baseline wave may lead to selection bias that limit generalisation to the baseline target population (Bulbulia, 2024b). See Appendix E.

Sensitivity Analysis

We perform sensitivity analyses using the E-value metric (Linden et al., 2020; VanderWeele & Ding, 2017). The E-value represents the minimum association strength (on the risk-ratio scale) that an unmeasured confounder would need with both exposure and outcome—after adjusting for measured covariates—to explain away the observed association (Linden et al., 2020; VanderWeele et al., 2020). Confidence intervals for each E-value were derived from the multiplicity-adjusted confidence intervals of the corresponding coefficient estimates (bonferroni, $\alpha = 0.05$), so the sensitivity analysis obeys the same error-control framework as the main results.

Results

Average Treatment Effects

ATE Effects of Extraversion on Mult

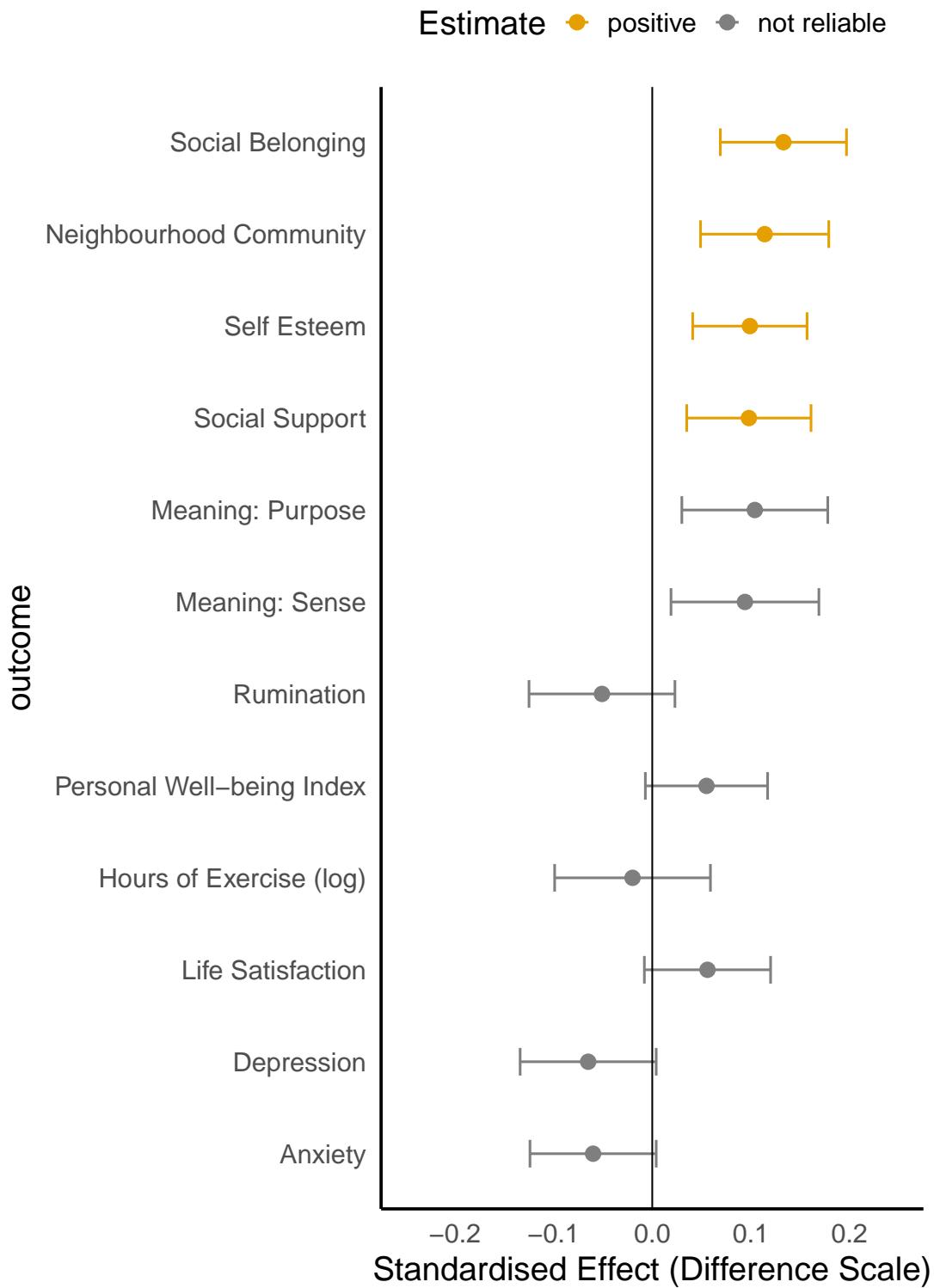


Figure 2: Average Treatment Effects on Multi-dimensional Wellbeing

Table 1: Average Treatment Effects on Multi-dimensional Wellbeing

Outcome	ATE	2.5 %	97.5 %	E-Value	E-Value bound
Social Belonging	0.133	0.09	0.177	1.51	1.39
Neighbourhood Community	0.114	0.07	0.159	1.458	1.328
Self Esteem	0.099	0.06	0.139	1.415	1.299
Social Support	0.098	0.055	0.141	1.413	1.284
Meaning: Purpose	0.104	0.053	0.154	1.43	1.278
Meaning: Sense	0.094	0.043	0.145	1.401	1.244
Depression	-0.065	-0.112	-0.017	1.315	1.146
Anxiety	-0.06	-0.104	-0.016	1.3	1.141
Life Satisfaction	0.056	0.012	0.1	1.287	1.121
Personal Well-being Index	0.055	0.013	0.098	1.284	1.116
Rumination	-0.051	-0.102	-0.001	1.271	1.012
Hours of Exercise (log)	-0.02	-0.074	0.034	1.155	1

Confidence intervals were adjusted for multiple comparisons using bonferroni correction ($\alpha = 0.05$). E-values were also adjusted using bonferroni correction ($\alpha = 0.05$).

The following outcomes showed reliable causal evidence (E-value lower bound > 1.2): - Social Belonging: 0.133(0.069,0.197); on the original scale, 0.145 (0.075,0.215). E-value bound = 1.329 - Neighbourhood Community: 0.114(0.049,0.179); on the original scale, 0.179 (0.077,0.281). E-value bound = 1.264 - Self Esteem: 0.099(0.041,0.157); on the original scale, 0.126 (0.052,0.2). E-value bound = 1.238 - Social Support: 0.098(0.035,0.161); on the original scale, 0.11 (0.039,0.18). E-value bound = 1.216

Planned Subgroup Comparisons (Optional)

Based on theoretical findings we expected that the effects of {name_exposure} would vary by age...@fig-planned-comparison and Table 2

Younger vs Older

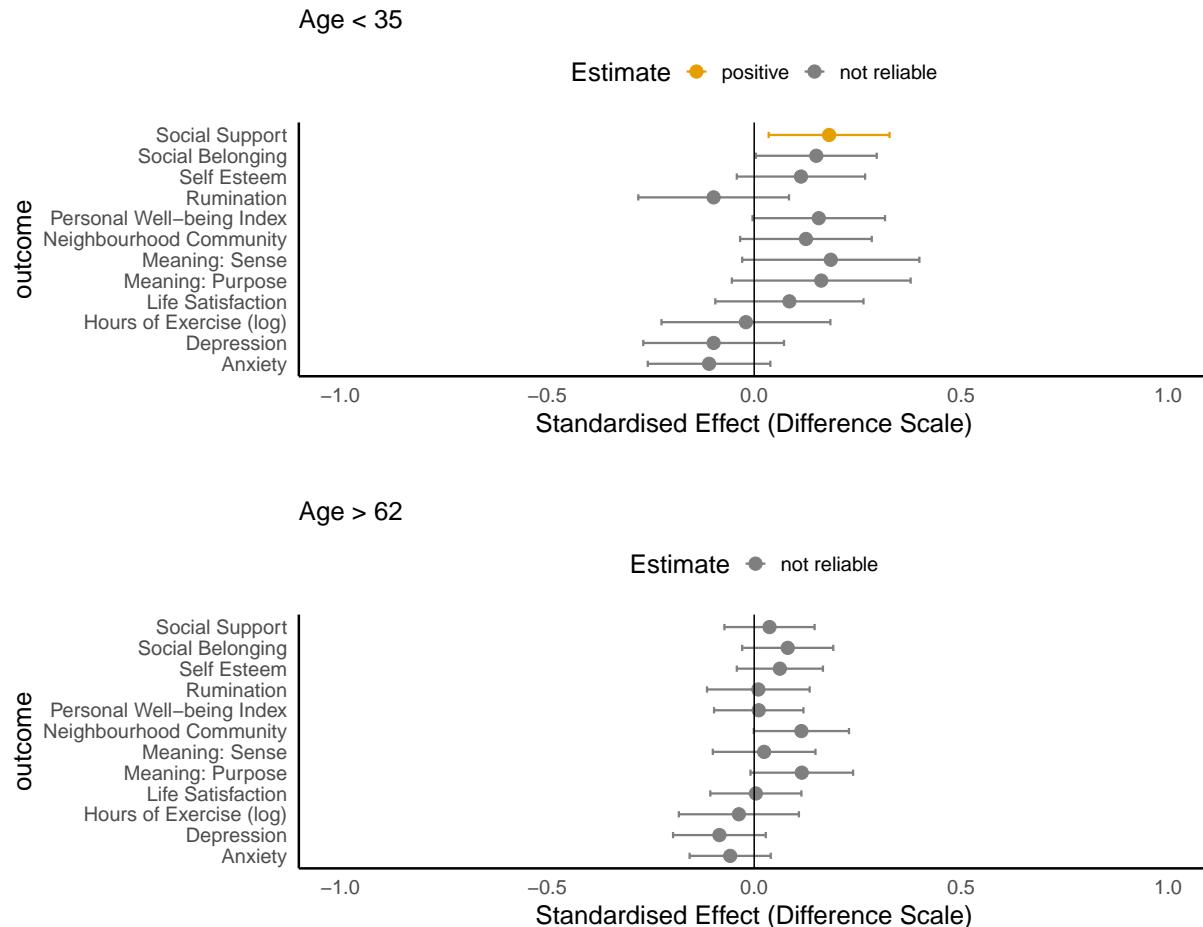


Figure 3: Planned Comparison Plot

Table 2: Planned Comparison Table

Outcomes	Group Differences
Social Support	-0.144 [-0.269, -0.019]
Social Belonging	-0.069 [-0.194, 0.056]
Self Esteem	-0.051 [-0.179, 0.077]
Rumination	0.108 [-0.043, 0.259]
Personal Well-being Index	-0.145 [-0.277, -0.013]
Neighbourhood Community	-0.011 [-0.146, 0.124]
Meaning: Sense	-0.161 [-0.330, 0.008]
Meaning: Purpose	-0.047 [-0.218, 0.124]

Table 2: Planned Comparison Table

Outcomes	Group Differences
Life Satisfaction	-0.081 [-0.225, 0.063]
Hours of Exercise (log)	-0.017 [-0.188, 0.154]
Depression	0.014 [-0.125, 0.153]
Anxiety	0.051 [-0.070, 0.172]

We found reliable treatment-effect differences comparing People Over 62 Years Old to People Under 35 Years Old for Social Support (People Over 62 Years Old vs People Under 35 Years Old): $\delta = -0.144$ [-0.269, -0.019] and Personal Well-being Index (People Over 62 Years Old vs People Under 35 Years Old): $\delta = -0.145$ [-0.277, -0.013]. We did not find reliable differences for all other outcomes.

Heterogeneous Treatment Effects

Qini Curves

The Qini curve shows the cumulative **gain** as we expand a targeting rule down the CATE ranking.

- **Beneficial exposure:** we add individuals from the top positive CATEs downward; the baseline is ‘expose everyone.’
- **Detrimental exposure:** we first flip outcome direction (so higher values represent **more harm**; see Anxiety, Depression, Rumination), then *add* the exposure starting with individuals whose CATEs show the **greatest harm**, gradually including those predicted to be more resistant to harm; the baseline is ‘expose everyone.’ The curve therefore quantifies the harm by when those most susceptible to harm are exposed.

If the Qini curve stays above its baseline, a targeted policy increases the outcome more than a one-size-fits-all alternative. (Outcome directions were flipped where needed—Anxiety, Depression, Rumination—so the positively valenced exposures always have positively valanced outcomes and negative exposures always have negatively valenced outcomes.)

We computed cumulative gains from prioritising individuals by CATE at 20% and 50% spend levels, comparing against a no-prioritisation baseline.

Social Belonging No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.

Anxiety (reversed) No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.

Depression (reversed) No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.

Life Satisfaction At 20% spend: CATE prioritisation is beneficial (diff: 0.09 [95% CI 0.05, 0.13]). At 50% spend: CATE prioritisation is beneficial (diff: 0.09 [95% CI 0.03, 0.14]).

Hours of Exercise (log) No benefits for priority investments as measured by the Qini curve at the twenty or fifty percent spend levels.

Meaning: Purpose At 20% spend: CATE prioritisation is beneficial (diff: 0.08 [95% CI 0.05, 0.12]). At 50% spend: CATE prioritisation is beneficial (diff: 0.08 [95% CI 0.03, 0.13]).

Meaning: Sense At 20% spend: CATE prioritisation is beneficial (diff: 0.08 [95% CI 0.04, 0.12]). At 50% spend: CATE prioritisation is beneficial (diff: 0.07 [95% CI 0.02, 0.12]).

Neighbourhood Community At 20% spend: CATE prioritisation is beneficial (diff: 0.09 [95% CI 0.05, 0.12]). At 50% spend: No reliable benefits from CATE prioritisation.

Personal Well-being Index At 20% spend: CATE prioritisation is beneficial (diff: 0.10 [95% CI 0.06, 0.13]). At 50% spend: CATE prioritisation is beneficial (diff: 0.09 [95% CI 0.03, 0.14]).

Rumination (reversed) At 20% spend: No reliable benefits from CATE prioritisation. At 50% spend: CATE prioritisation worsens outcomes compared to ATE.

Self Esteem At 20% spend: CATE prioritisation is beneficial (diff: 0.06 [95% CI 0.03, 0.10]). At 50% spend: CATE prioritisation is beneficial (diff: 0.07 [95% CI 0.02, 0.11]).

Social Support At 20% spend: CATE prioritisation is beneficial (diff: 0.10 [95% CI 0.06, 0.13]). At 50% spend: CATE prioritisation is beneficial (diff: 0.07 [95% CI 0.02, 0.13]).

Table 3 presents results for our Qini curve analysis at different spend rates.

Table 3: Qini Curve Results

Model	Spend 20%	Spend 50%
Social Belonging	0.03 [-0.00, 0.07]	0.00 [-0.05, 0.05]
Anxiety (reversed)	-0.00 [-0.01, 0.01]	-0.00 [-0.01, 0.01]
Depression (reversed)	-0.00 [-0.01, 0.01]	-0.00 [-0.01, 0.01]
Life Satisfaction	0.09 [0.05, 0.13]	0.09 [0.03, 0.14]
Hours of Exercise (log)	0.01 [-0.02, 0.04]	-0.02 [-0.07, 0.03]
Meaning: Purpose	0.08 [0.05, 0.12]	0.08 [0.03, 0.13]
Meaning: Sense	0.08 [0.04, 0.12]	0.07 [0.02, 0.12]
Neighbourhood Community	0.09 [0.05, 0.12]	0.03 [-0.02, 0.08]
Personal Well-being Index	0.10 [0.06, 0.13]	0.09 [0.03, 0.14]
Rumination (reversed)	-0.01 [-0.03, 0.01]	-0.04 [-0.06, -0.01]
Self Esteem	0.06 [0.03, 0.10]	0.07 [0.02, 0.11]
Social Support	0.10 [0.06, 0.13]	0.07 [0.02, 0.13]

Figure 4 presents results for reliable Qini results

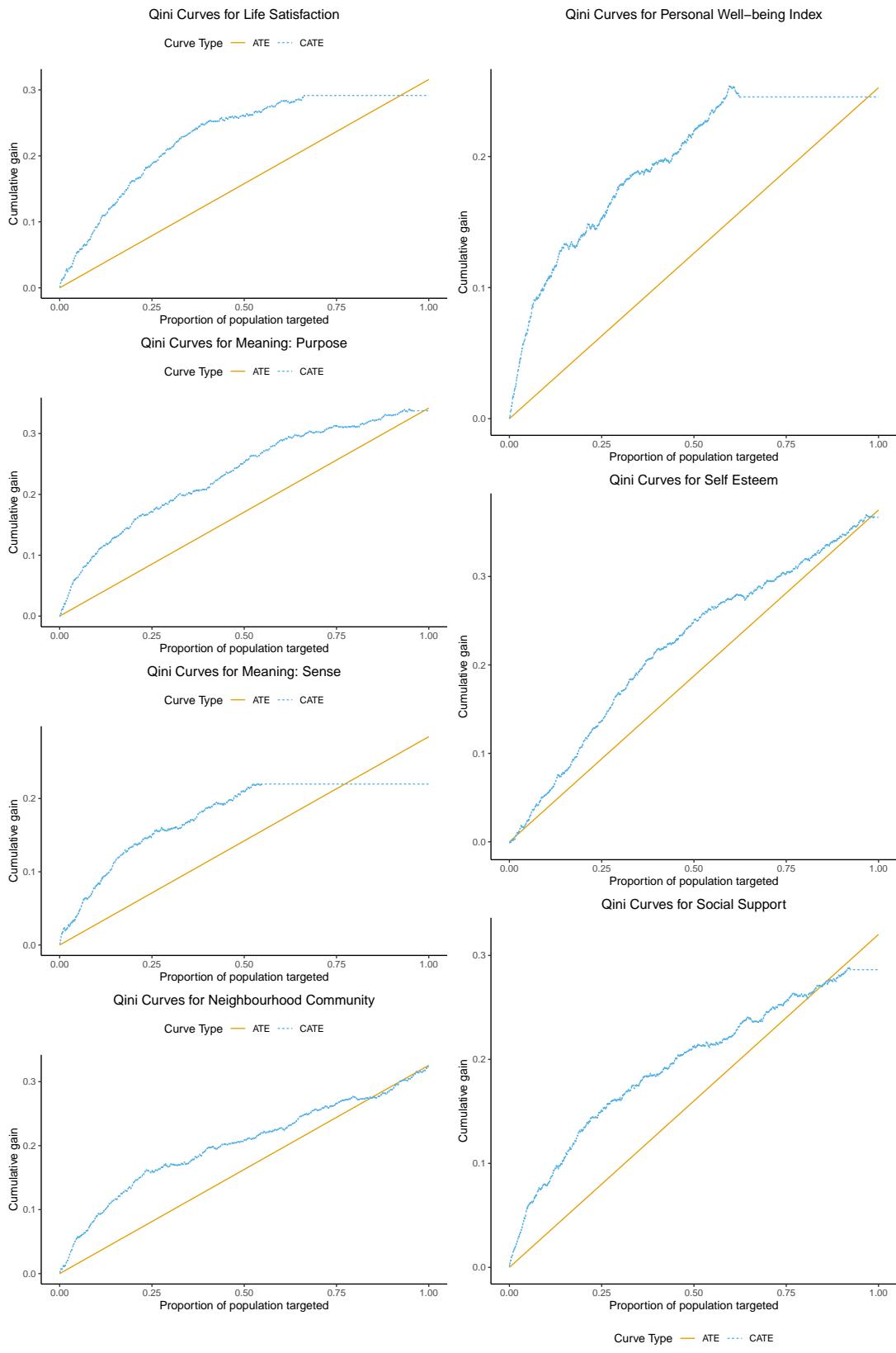


Figure 4: Qini Graphs

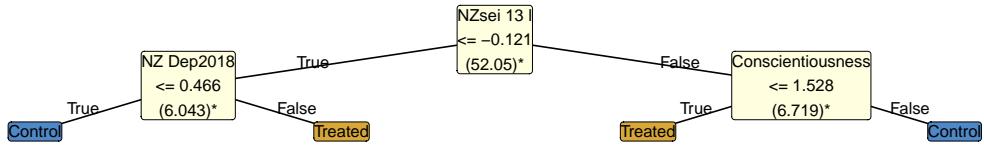
Decision Rules (Who is Most Sensitive to Treatment?)

Policy Trees

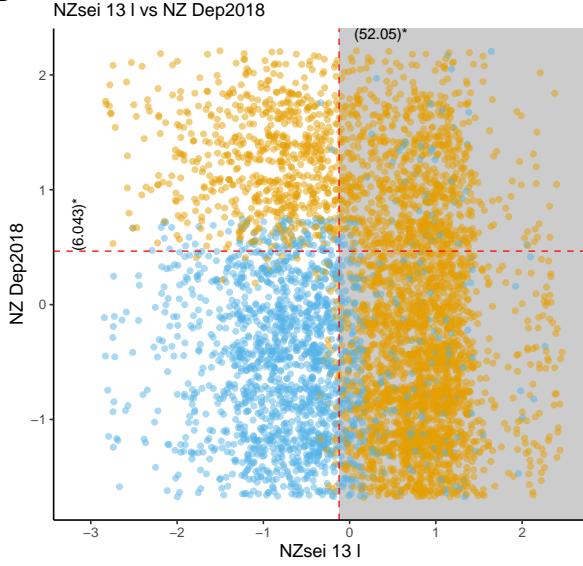
We used policy trees ([Athey & Wager, 2021b, 2021a](#); [Sverdrup et al., 2024](#)) to find straightforward ‘if-then’ rules for who benefits most from treatment, based on participant characteristics. Because we flipped some measures, a higher predicted effect always means greater improvement. Policy trees can uncover small but important subgroups whose treatment responses stand out, even when the overall differences might be modest.

A

Life Satisfaction



B



C

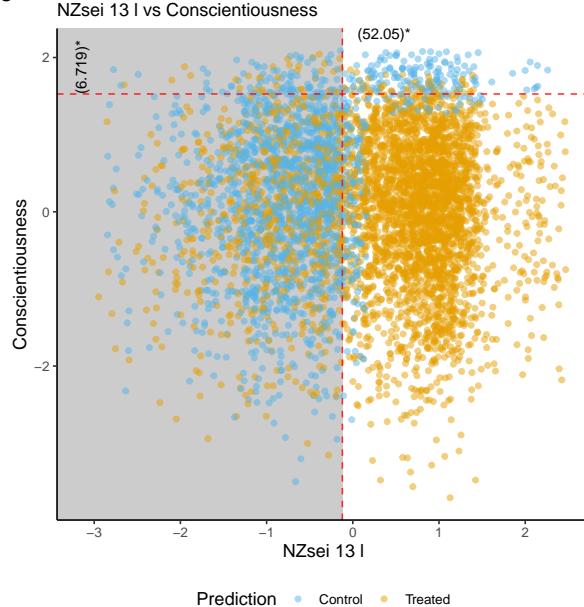


Figure 5: Decision Tree: {glued_policy_names_1}

Findings for Life Satisfaction:

Split 1: $\text{NZsei 13 l} \leq -0.121$ (original: 52.05). Within that subgroup, split 2a: $\text{NZ Dep2018} \leq 0.466$ (original: 6.043), → **Control**; $\text{NZ Dep2018} > 0.466$ (original: 6.043) → **Treated**.

Split 2: $\text{NZsei 13 l} > -0.121$ (original: 52.05). Within that subgroup, split 2b: $\text{Conscientiousness} \leq 1.528$ (original: 6.719), → **Treated**; $\text{Conscientiousness} > 1.528$ (original: 6.719) → **Control**.

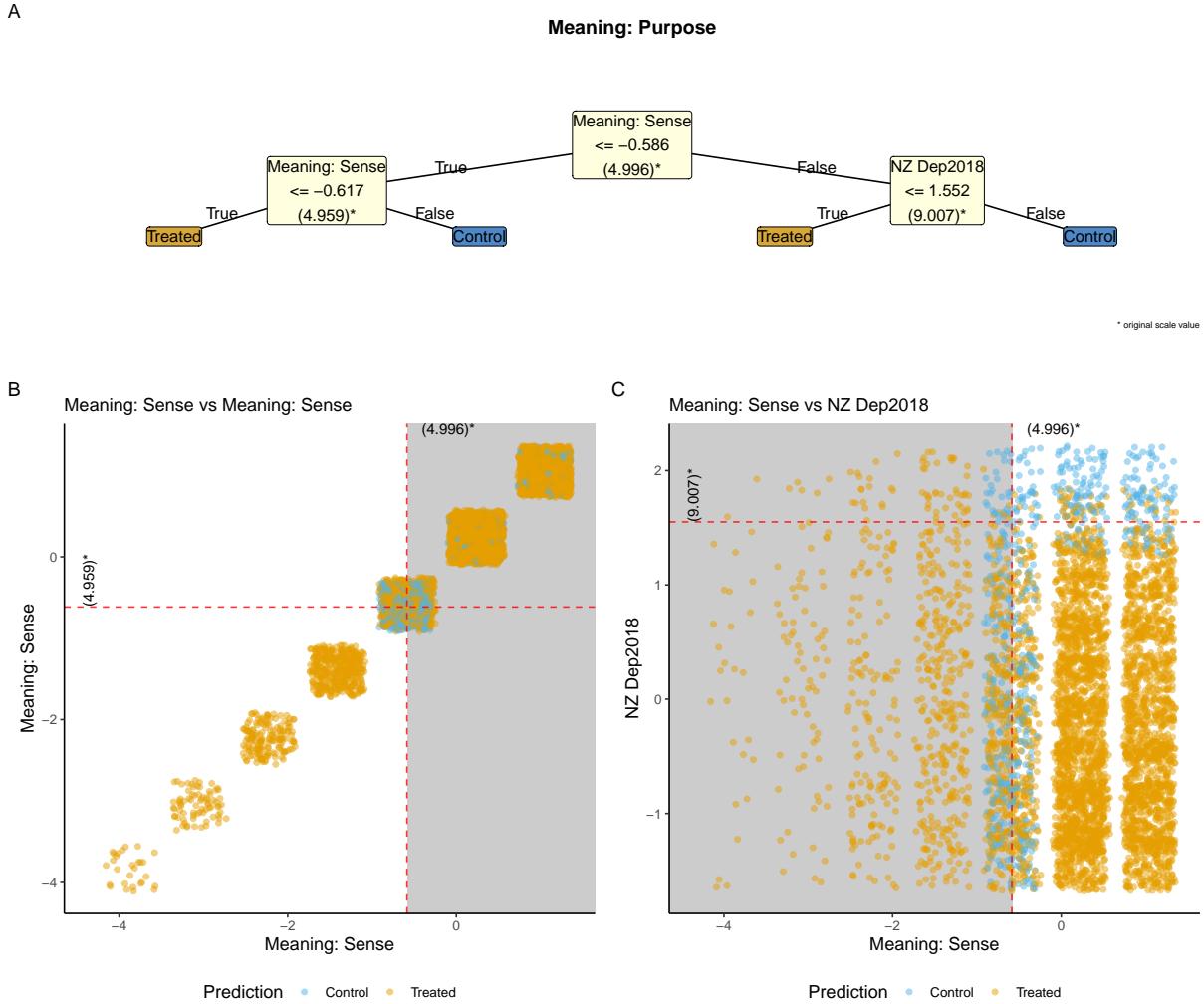


Figure 6: Decision Tree: {glued_policy_names_2}

Findings for Meaning: Purpose:

Split 1: Meaning: Sense ≤ -0.586 (original: 4.996). Within that subgroup, split 2a: Meaning: Sense ≤ -0.617 (original: 4.959), \rightarrow **Treated**; Meaning: Sense > -0.617 (original: 4.959) \rightarrow **Control**.

Split 2: Meaning: Sense > -0.586 (original: 4.996). Within that subgroup, split 2b: NZ Dep2018 ≤ 1.552 (original: 9.007), \rightarrow **Treated**; NZ Dep2018 > 1.552 (original: 9.007) \rightarrow **Control**.

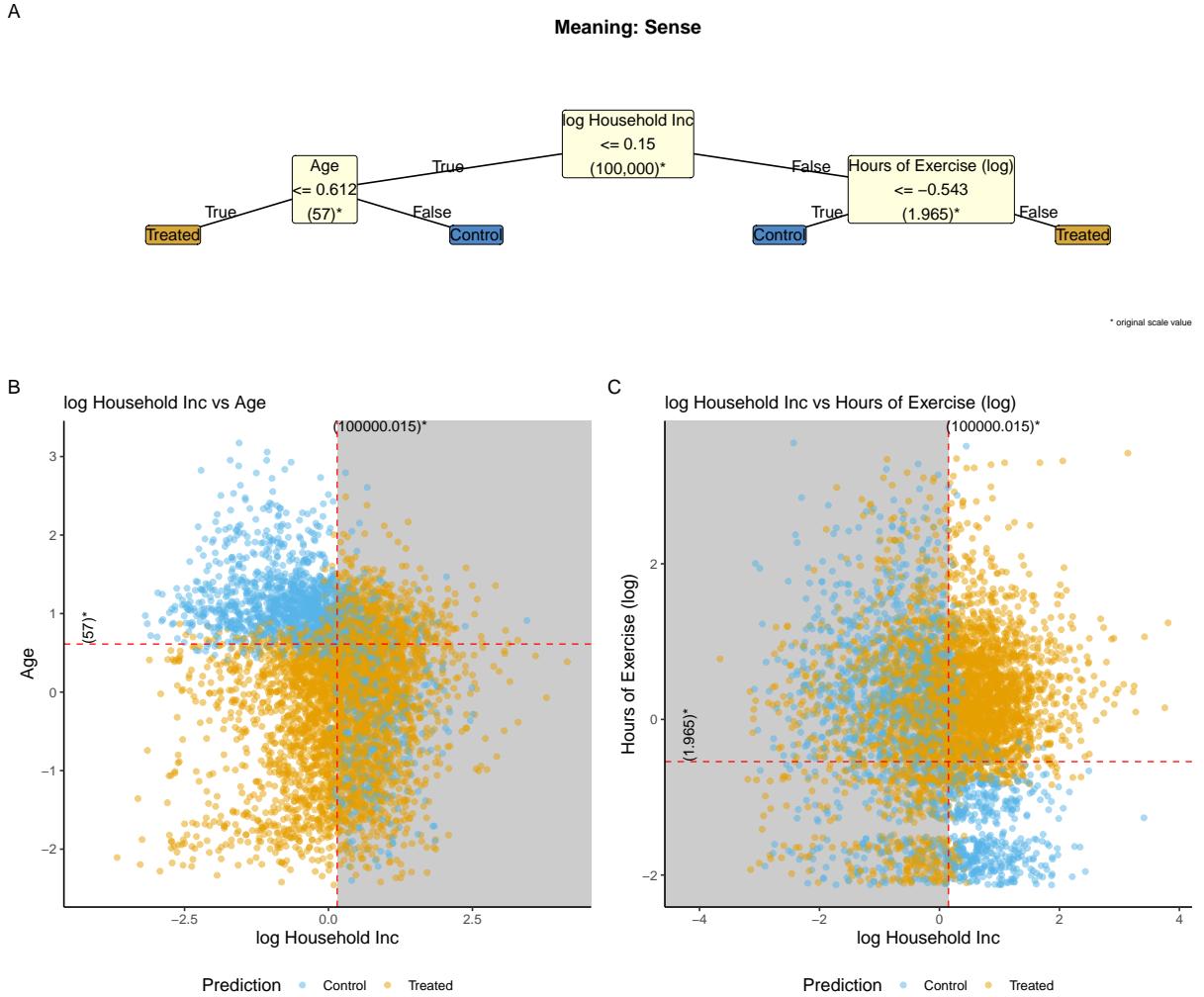


Figure 7: Decision Tree: {glued_policy_names_3}

Findings for Meaning: Sense:

Split 1: $\log \text{Household Inc} \leq 0.15$ (original: 100000.015). Within that subgroup, split 2a: $\text{Age} \leq 0.612$ (original: 57), → **Treated**; $\text{Age} > 0.612$ (original: 57) → **Control**.

Split 2: $\log \text{Household Inc} > 0.15$ (original: 100000.015). Within that subgroup, split 2b: $\text{Hours of Exercise} (\log) \leq -0.543$ (original: 1.965), → **Control**; $\text{Hours of Exercise} (\log) > -0.543$ (original: 1.965) → **Treated**.

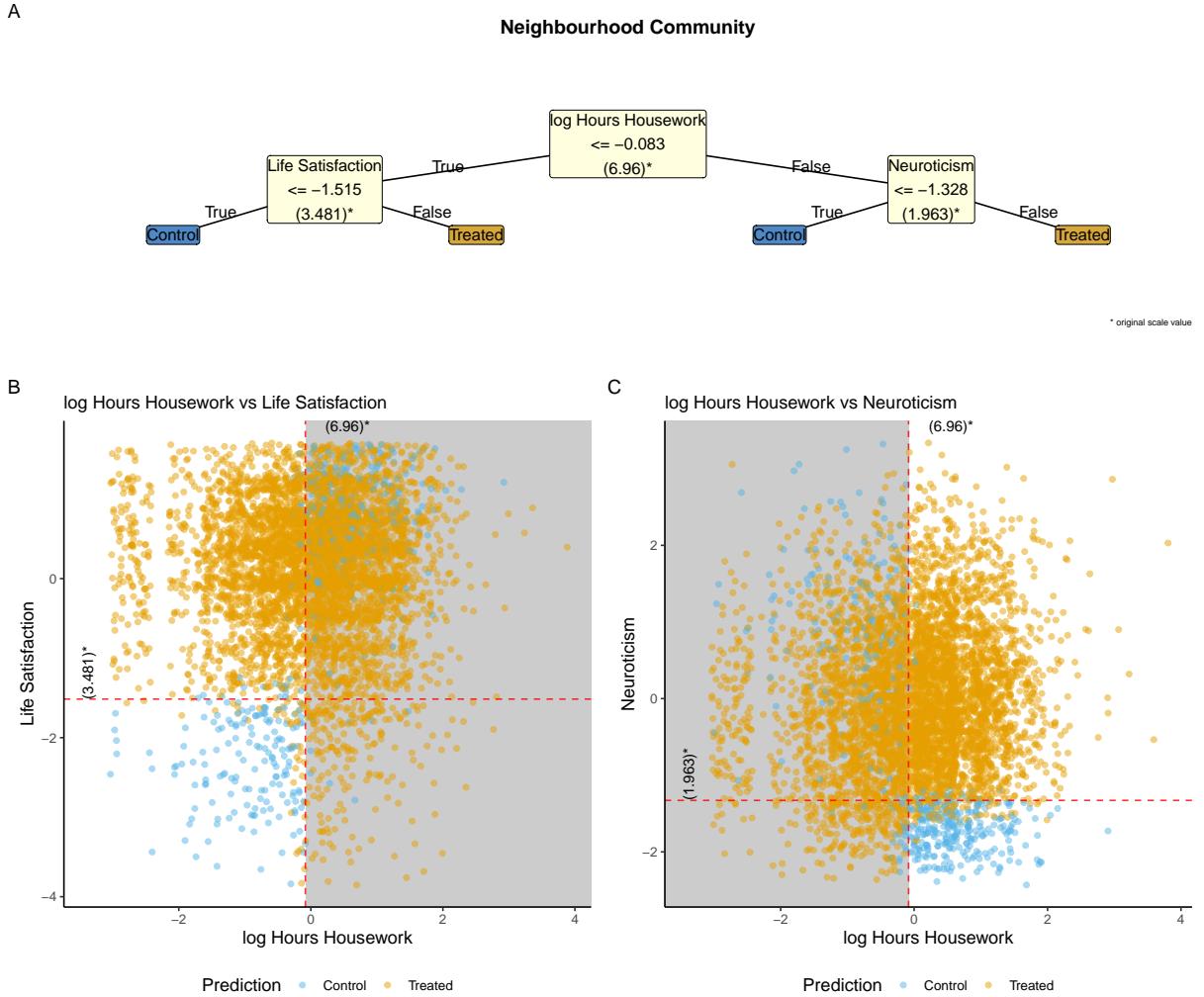


Figure 8: Decision Tree: {glued_policy_names_4}

Findings for Neighbourhood Community:

Split 1: $\log \text{Hours Housework} \leq -0.083$ (original: 6.96). Within that subgroup, split 2a: $\text{Life Satisfaction} \leq -1.515$ (original: 3.481), → **Control**; $\text{Life Satisfaction} > -1.515$ (original: 3.481) → **Treated**.

Split 2: $\log \text{Hours Housework} > -0.083$ (original: 6.96). Within that subgroup, split 2b: $\text{Neuroticism} \leq -1.328$ (original: 1.963), → **Control**; $\text{Neuroticism} > -1.328$ (original: 1.963) → **Treated**.

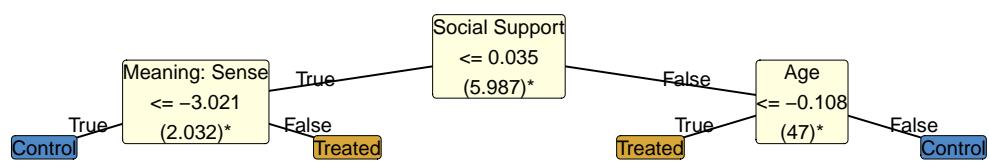
Findings for Personal Well-being Index:

Split 1: Social Support ≤ 0.035 (original: 5.987). Within that subgroup, split 2a: Meaning: Sense ≤ -3.021 (original: 2.032), → **Control**; Meaning: Sense > -3.021 (original: 2.032) → **Treated**.

Split 2: Social Support > 0.035 (original: 5.987). Within that subgroup, split 2b: Age ≤ -0.108 (original: 47), → **Treated**; Age > -0.108 (original: 47) → **Control**.

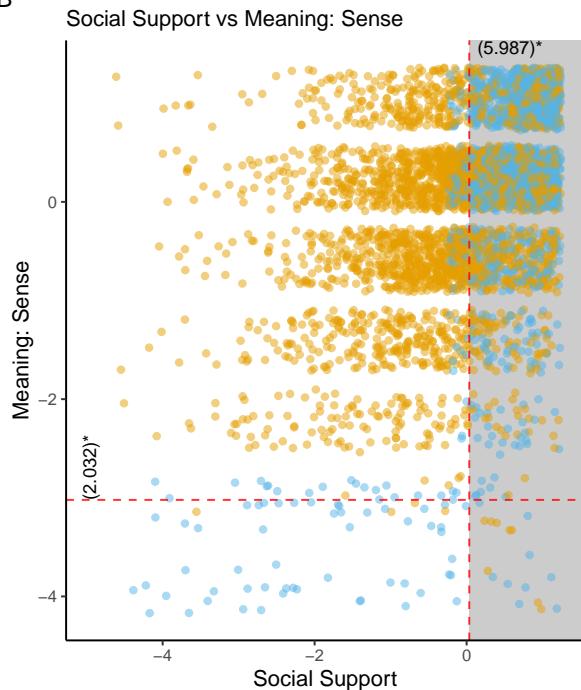
A

Personal Well-being Index

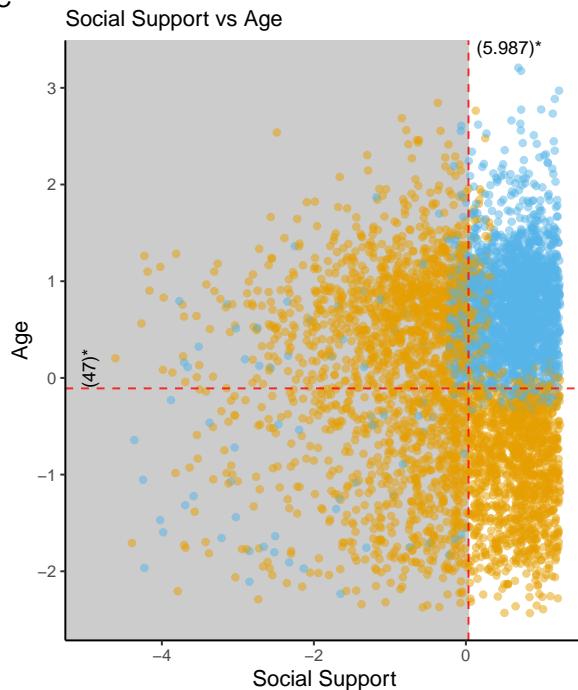


* original scale value

B



C

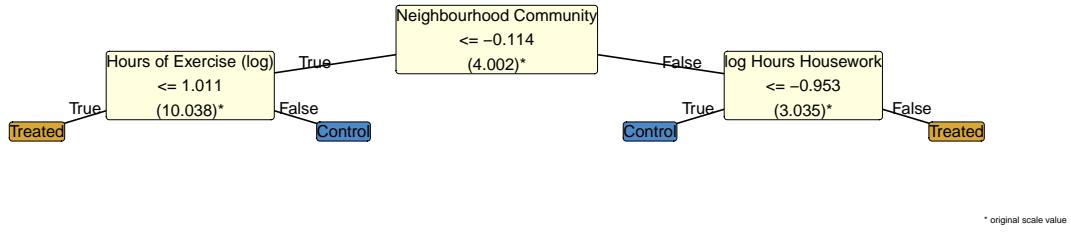


Prediction ● Control ● Treated

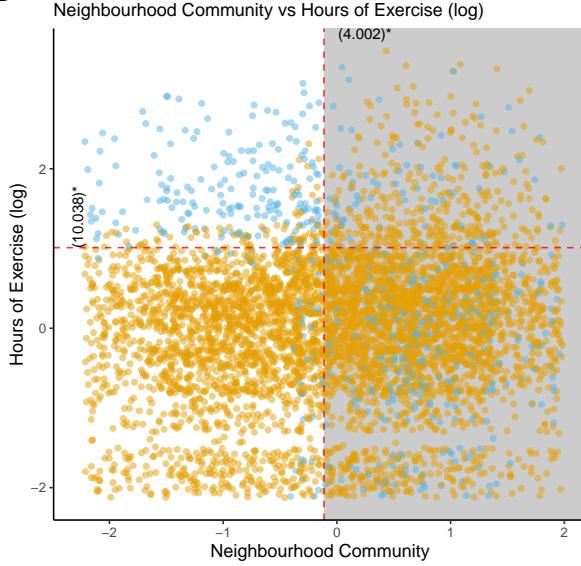
Prediction ● Control ● Treated

Figure 9: Decision Tree: {glued_policy_names_5}

A

Self Esteem

B



C

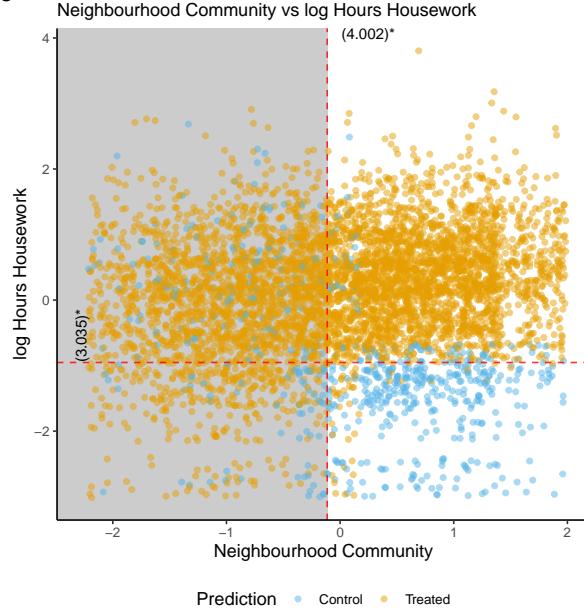


Figure 10: Decision Tree: {glued_policy_names_6}

Findings for Self Esteem:

Split 1: Neighbourhood Community ≤ -0.114 (original: 4.002). Within that subgroup, split 2a: Hours of Exercise (log) ≤ 1.011 (original: 10.038), \rightarrow **Treated**; Hours of Exercise (log) > 1.011 (original: 10.038) \rightarrow **Control**.

Split 2: Neighbourhood Community > -0.114 (original: 4.002). Within that subgroup, split 2b: log Hours Housework ≤ -0.953 (original: 3.035), \rightarrow **Control**; log Hours Housework > -0.953 (original: 3.035) \rightarrow **Treated**.

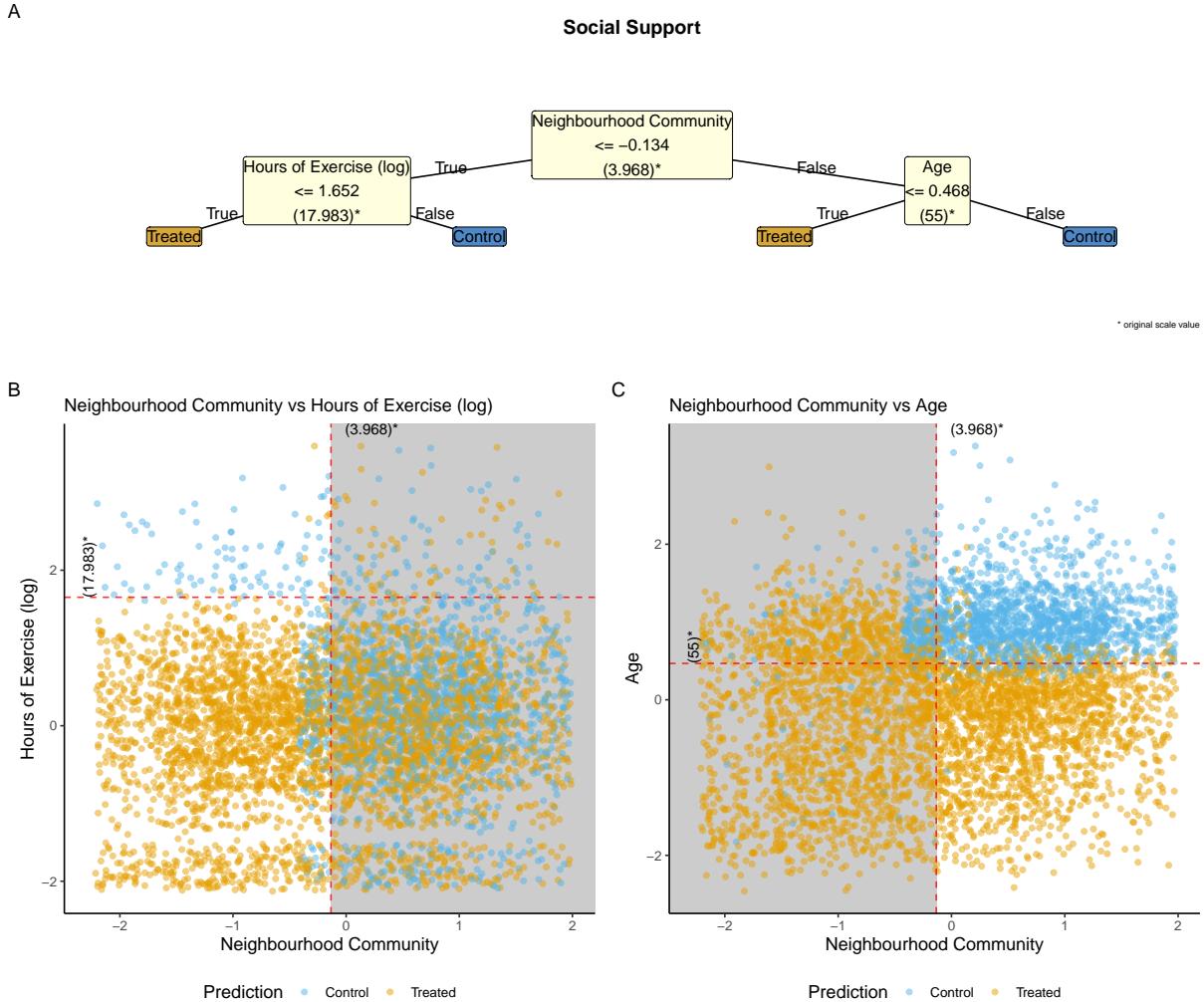


Figure 11: Decision Tree: {glued_policy_names_7}

Findings for Social Support:

Split 1: Neighbourhood Community ≤ -0.134 (original: 3.968). Within that subgroup, split 2a: Hours of Exercise (log) ≤ 1.652 (original: 17.983), \rightarrow **Treated**; Hours of Exercise (log) > 1.652 (original: 17.983) \rightarrow **Control**.

Split 2: Neighbourhood Community > -0.134 (original: 3.968). Within that subgroup, split 2b: Age ≤ 0.468 (original: 55), \rightarrow **Treated**; Age > 0.468 (original: 55) \rightarrow **Control**.

A shallow policy tree recommends actions based on two splits for depth=2, or one split for depth=1. We trained on 50% of the data and evaluated on the rest.

Discussion

Appendix A: Measures

Measures

Baseline Covariate Measures

Baseline Covariates

Age

What is your date of birth?

We asked participants' ages in an open-ended question ("What is your age?" or "What is your date of birth"). ([Sibley, 2021](#))

Agreeableness

I sympathize with others' feelings. I am not interested in other people's problems. I feel others' emotions. I am not really interested in others (reversed).

Mini-IPIP6 Agreeableness dimension: (i) I sympathize with others' feelings. (ii) I am not interested in other people's problems. (r) (iii) I feel others' emotions. (iv) I am not really interested in others. (r) ([Sibley et al., 2011](#))

Alcohol Frequency

"How often do you have a drink containing alcohol?"

Participants could chose between the following responses: '(1 = Never - I don't drink, 2 = Monthly or less, 3 = Up to 4 times a month, 4 = Up to 3 times a week, 5 = 4 or more times a week, 6 = Don't know)' ([Health, 2013](#))

Alcohol Intensity

"How many drinks containing alcohol do you have on a typical day when drinking alcohol? (number of drinks on a typical day when drinking)"

Participants responded using an open-ended box. ([Health, 2013](#))

Social Belonging

Know that people in my life accept and value me. Feel like an outsider (reversed). Know that people around me share my attitudes and beliefs.

We assessed felt belongingness with three items adapted from the Sense of Belonging Instrument (Hagerty & Patusky, 1995): (1) "Know that people in my life accept and value me"; (2) "Feel like an outsider"; (3) "Know that people around me share my attitudes and beliefs". Participants responded on a scale from 1 (Very Inaccurate) to 7 (Very Accurate). The second item was reversely coded. ([Hagerty & Patusky, 1995](#))

Born in Nz

Where were you born? (please be specific, e.g., which town/city?)

Coded binary (1 = New Zealand; 0 = elsewhere.) ([Sibley, 2021](#))

Conscientiousness

I get chores done right away. I like order. I make a mess of things. I often forget to put things back in their proper place.

Mini-IPIP6 Conscientiousness dimension: (i) I get chores done right away. (ii) I like order. (iii) I make a mess of things. (r) (iv) I often forget to put things back in their proper place. (r) ([Sibley et al., 2011](#))

Education Level

What is your highest level of qualification?

We asked participants, “What is your highest level of qualification?”. We coded participants highest finished degree according to the New Zealand Qualifications Authority. Ordinal-Rank 0-10 NZREG codes (with overseas school qualifications coded as Level 3, and all other ancillary categories coded as missing) ([Sibley, 2021](#))

Employed

Are you currently employed (This includes self-employed or casual work)?

Binary response: (0 = No, 1 = Yes) ([Statistics New Zealand, 2017](#))

Ethnicity

Which ethnic group(s) do you belong to?

Coded string: (1 = New Zealand European; 2 = Māori; 3 = Pacific; 4 = Asian) ([Statistics New Zealand, 2017](#))

Disability Status

Do you have a health condition or disability that limits you and that has lasted for 6+ months?

We assessed disability with a one-item indicator adapted from Verbrugge (1997). It asks, “Do you have a health condition or disability that limits you and that has lasted for 6+ months?” (1 = Yes, 0 = No). ([Verbrugge, 1997](#))

Log Hours with Children

Hours spent...looking after children.

We took the natural log of the response + 1. ([Sibley et al., 2011](#))

Log Hours Commuting

Hours spent...travelling/commuting.

We took the natural log of the response + 1. ([Sibley, 2021](#))

Log Hours of Exercise

Hours spent...exercising/physical activity.

We took the natural log of the response + 1. ([Sibley et al., 2011](#))

Log Hours on Housework

Hours spent...housework/cooking.

We took the natural log of the response + 1. ([Sibley et al., 2011](#))

Log Household Income

Please estimate your total household income (before tax) for the year XXXX.

We took the natural log of the response + 1. ([Sibley, 2021](#))

Male

We asked participants' gender in an open-ended question: "what is your gender?"

Here, we coded all those who responded as Male as 1, and those who did not as 0. ([Fraser et al., 2020](#))

Neuroticism

I have frequent mood swings. I am relaxed most of the time (reversed). I get upset easily. I seldom feel blue (reversed).

Mini-IPIP6 Neuroticism dimension: (i) I have frequent mood swings. (ii) I am relaxed most of the time. (r) (iii) I get upset easily. (iv) I seldom feel blue. (r) ([Sibley et al., 2011](#))

Non Heterosexual

How would you describe your sexual orientation? (e.g., heterosexual, homosexual, straight, gay, lesbian, bisexual, etc.)

Open-ended question, coded as binary (not heterosexual = 1). ([Greaves et al., 2017](#))

Nz Deprivation Index

New Zealand Deprivation - Decile Index - Using 2018 Census Data

Numerical: (1-10) ([Atkinson et al., 2019](#))

Occupational Prestige Index

We assessed occupational prestige and status using the New Zealand Socio-economic Index 13 (NZSEI-13).

This index uses the income, age, and education of a reference group, in this case, the 2013 New Zealand census, to calculate a score for each occupational group. Scores range from 10 (Lowest) to 90 (Highest). This list of index scores for occupational groups was used to assign each participant a NZSEI-13 score based on their occupation. ([Fahy et al., 2017](#))

Openness

I have a vivid imagination. I have difficulty understanding abstract ideas (reversed). I do not have a good imagination (reversed). I am not interested in abstract ideas (reversed).

Mini-IPIP6 Openness to Experience dimension: (i) I have a vivid imagination. (ii) I have difficulty understanding abstract ideas. (r) (iii) I do not have a good imagination. (r) (iv) I am not interested in abstract ideas. (r) ([Sibley et al., 2011](#))

Parent

If you are a parent, in which year was your eldest child born?

Parents were coded as 1, while the others were coded as 0. ([Sibley, 2021](#))

Has Partner

What is your relationship status? (e.g., single, married, de-facto, civil union, widowed, living together, etc.)

Coded as binary (has partner = 1). ([Sibley, 2021](#))

Political Conservatism

Please rate how politically liberal versus conservative you see yourself as being.

Ordinal response: (1 = Extremely Liberal, 7 = Extremely Conservative) ([Jost, 2006](#))

Major Religions

Do you identify with a religion and/or spiritual group? -> (If yes...)-> What religion or spiritual group?

Open-ended (string). Coded from New Zealand Census Categories. Levels are: “Not Religious”, “Anglican”, “Buddhist”, “Catholic”, “Christian (Non-Denominational)”, “Christian (Other Denominations)”, “Hindu”, “Jewish”, “Muslim”, “Presbyterian, Congregational, Reformed”, “Other Religions”. ([Sibley, 2021](#))

Religious Identification

How important is your religion to how you see yourself?

Ordinal response: (1 = Not Important, 7 = Very Important) ([Sibley, 2021](#))

Rural Classification

High Urban Accessibility = 1, Medium Urban Accessibility = 2, Low Urban Accessibility = 3, Remote = 4, Very Remote = 5.

“Participants residence locations were coded according to a five-level ordinal categorisation ranging from Urban to Rural.” ([Whitehead et al., 2023](#))

Sample Frame Opt in

Participant was not randomly sampled from the New Zealand Electoral Roll.

Code string (Binary): (0 = No, 1 = Yes) ([Sibley, 2021](#))

Short Form Health

In general, would you say your health is...

Ordinal response: (1 = Poor, 7 = Excellent) ([Instrument Ware Jr & Sherbourne, 1992](#))

Smoker

Do you currently smoke tobacco cigarettes?

Binary smoking indicator (0 = No, 1 = Yes). ([Sibley, 2021](#))

Exposure Measures

Exposure Variable

Extraversion

I am the life of the party. I don't talk a lot (reversed). I keep in the background (reversed). I talk to a lot of different people at parties.

Mini-IPIP6 Extraversion dimension: (i) I am the life of the party. (ii) I don't talk a lot. (r) (iii) I keep in the background. (r) (iv) I talk to a lot of different people at parties. ([Sibley et al., 2011](#))

Outcome Measures

Outcome Variables

Social Belonging

Know that people in my life accept and value me. Feel like an outsider (reversed). Know that people around me share my attitudes and beliefs.

We assessed felt belongingness with three items adapted from the Sense of Belonging Instrument (Hagerty & Patusky, 1995): (1) “Know that people in my life accept and value me”; (2) “Feel like an outsider”; (3) “Know that people around me share my attitudes and beliefs”. Participants responded on a scale from 1 (Very Inaccurate) to 7 (Very Accurate). The second item was reversely coded. ([Hagerty & Patusky, 1995](#))

Anxiety

During the past 30 days, how often did...you feel restless or fidgety? During the past 30 days, how often did...you feel that everything was an effort? During the past 30 days, how often did...you feel nervous?

Ordinal response: (0 = None Of The Time; 1 = A Little Of The Time; 2= Some Of The Time; 3 = Most Of The Time; 4 = All Of The Time) ([Kessler et al., 2002](#))

Depression

During the past 30 days, how often did...you feel hopeless? During the past 30 days, how often did...you feel so depressed that nothing could cheer you up? During the past 30 days, how often did...you feel you feel restless or fidgety?

Ordinal response: (0 = None Of The Time; 1 = A Little Of The Time; 2= Some Of The Time; 3 = Most Of The Time; 4 = All Of The Time) ([Kessler et al., 2002](#))

Life Satisfaction

I am satisfied with my life. In most ways my life is close to ideal.

Ordinal response (1 = Strongly Disagree to 7 = Strongly Agree). ([Diener et al., 1985](#))

Log Hours of Exercise

Hours spent...exercising/physical activity.

We took the natural log of the response + 1. ([Sibley et al., 2011](#))

Meaning Purpose

My life has a clear sense of purpose

Ordinal response (1 = Strongly Disagree to 7 = Strongly Agree). ([Steger et al., 2006](#))

Meaning Sense

I have a good sense of what makes my life meaningful.

Ordinal response (1 = Strongly Disagree to 7 = Strongly Agree). ([Steger et al., 2006](#))

Neighbourhood Community

I feel a sense of community with others in my local neighbourhood.

Ordinal response (1 = Strongly Disagree to 7 = Strongly Agree). ([Sengupta et al., 2013](#))

Personal Well Being Index

no information available for this variable.

Rumination

During the last 30 days, how often did...you have negative thoughts that repeated over and over?

Ordinal responses: 0 = None of The Time, 1 = A little of The Time, 2 = Some of The Time, 3 = Most of The Time, 4 = All of The Time. ([Nolen-hoeksema & Morrow, 1993](#))

Self Esteem

On the whole am satisfied with myself. Take a positive attitude toward myself. Am inclined to feel that I am a failure (reversed).

Ordinal response (1 = Very inaccurate to 7 = Very accurate). ([Rosenberg, 1965](#))

Social Support

There are people I can depend on to help me if I really need it. There is no one I can turn to for guidance in times of stress (reversed). I know there are people I can turn to when I need help.

Ordinal response: (1 = Strongly Disagree, 7 = Strongly Agree) ([Cutrona & Russell, 1987](#))

Appendix B: Sample Characteristics

Sample Statistics: Baseline Covariates

Table 4 presents sample demographic statistics.

Table 4: Demographic statistics for New Zealand Attitudes and Values Cohort: {baseline_wave_glued}.

	2018
	(N=39635)
Age	
Mean (SD)	48.5 (13.9)
Median [Min, Max]	51.0 [18.0, 99.0]
Agreeableness	
Mean (SD)	5.35 (0.988)
Median [Min, Max]	5.47 [1.00, 7.00]
Missing	9 (0.0%)
Alcohol Frequency	
Mean (SD)	2.16 (1.34)
Median [Min, Max]	2.00 [0, 5.00]
Missing	1342 (3.4%)
Alcohol Intensity	
Mean (SD)	2.15 (2.09)
Median [Min, Max]	2.00 [0, 15.0]
Missing	2348 (5.9%)
Belong	
Mean (SD)	5.14 (1.07)
Median [Min, Max]	5.31 [1.00, 7.00]
Missing	7 (0.0%)
Born in NZ	
0	8510 (21.5%)
1	30670 (77.4%)
Missing	455 (1.1%)
Conscientiousness	
Mean (SD)	5.10 (1.06)
Median [Min, Max]	5.23 [1.00, 7.00]
Education Level	
no_qualification	1003 (2.5%)
cert_1_to_4	13801 (34.8%)
cert_5_to_6	4953 (12.5%)
university	10400 (26.2%)
post_grad	4220 (10.6%)
masters	3297 (8.3%)
doctorate	930 (2.3%)
Missing	1031 (2.6%)
Employed	
0	8111 (20.5%)
1	31475 (79.4%)
Missing	49 (0.1%)
Ethnicity	
euro	31454 (79.4%)

	2018
maori	4561 (11.5%)
pacific	971 (2.4%)
asian	2124 (5.4%)
Missing	525 (1.3%)
Disability Status	
Mean (SD)	0.223 (0.416)
Median [Min, Max]	0 [0, 1.00]
Missing	745 (1.9%)
Log Hours with Children	
Mean (SD)	1.18 (1.61)
Median [Min, Max]	0.0341 [0, 5.13]
Missing	1242 (3.1%)
Log Hours Commuting	
Mean (SD)	1.50 (0.832)
Median [Min, Max]	1.61 [0, 4.40]
Missing	1242 (3.1%)
Log Hours Exercising	
Mean (SD)	1.55 (0.846)
Median [Min, Max]	1.61 [0, 4.40]
Missing	1242 (3.1%)
Log Hours on Housework	
Mean (SD)	2.14 (0.782)
Median [Min, Max]	2.20 [0, 5.13]
Missing	1242 (3.1%)
Log Household Income	
Mean (SD)	11.4 (0.765)
Median [Min, Max]	11.5 [0.685, 14.9]
Missing	3067 (7.7%)
Male	
0	24766 (62.5%)
1	14767 (37.3%)
Missing	102 (0.3%)
Neuroticism	
Mean (SD)	3.49 (1.15)
Median [Min, Max]	3.48 [1.00, 7.00]
Missing	10 (0.0%)
Non-heterosexual	
0	35100 (88.6%)
1	2562 (6.5%)
Missing	1973 (5.0%)
NZ Deprivation Index	
Mean (SD)	4.77 (2.73)
Median [Min, Max]	4.05 [1.00, 10.0]
Missing	255 (0.6%)
Occupational Prestige Index	
Mean (SD)	54.1 (16.5)
Median [Min, Max]	54.0 [10.0, 90.0]
Missing	536 (1.4%)
Openness	

	2018
Mean (SD)	4.96 (1.12)
Median [Min, Max]	5.00 [1.00, 7.00]
Missing	3 (0.0%)
Parent	
0	11539 (29.1%)
1	27776 (70.1%)
Missing	320 (0.8%)
Has Partner	
Mean (SD)	0.752 (0.432)
Median [Min, Max]	1.00 [0, 1.00]
Missing	1244 (3.1%)
Political Conservatism	
Mean (SD)	3.59 (1.38)
Median [Min, Max]	3.97 [1.00, 7.00]
Missing	2682 (6.8%)
Major Religions	
not_rel	24886 (62.8%)
anglican	2087 (5.3%)
buddist	332 (0.8%)
catholic	3123 (7.9%)
christian_nfd	4534 (11.4%)
christian_others	1738 (4.4%)
hindu	206 (0.5%)
jewish	80 (0.2%)
muslim	90 (0.2%)
presby_cong_reform	875 (2.2%)
the_others	1068 (2.7%)
Missing	616 (1.6%)
Religious Identification	
Mean (SD)	2.36 (2.18)
Median [Min, Max]	1.00 [1.00, 7.00]
Missing	1050 (2.6%)
Rural Classification	
High Urban Accessibility	24406 (61.6%)
Medium Urban Accessibility	7431 (18.7%)
Low Urban Accessibility	4818 (12.2%)
Remote	2241 (5.7%)
Very Remote	486 (1.2%)
Missing	253 (0.6%)
Sample Frame Opt-In	
0	38485 (97.1%)
1	1150 (2.9%)
Short Form Health	
Mean (SD)	5.05 (1.17)
Median [Min, Max]	5.04 [1.00, 7.00]
Missing	6 (0.0%)
Smoker	
0	35771 (90.3%)
1	2880 (7.3%)

	2018
Missing	984 (2.5%)

Sample Statistics: Exposure Variable

Table 5: Demographic statistics for New Zealand Attitudes and Values Cohort waves 2018.

	2018	2019
	(N=39635)	(N=39635)
Extraversion		
Mean (SD)	3.91 (1.20)	3.86 (1.19)
Median [Min, Max]	3.96 [1.00, 7.00]	3.79 [1.00, 7.00]
Missing	0 (0%)	11117 (28.0%)
Extraversion (binary)		
[1.0,4.0]	21138 (53.3%)	15637 (39.5%)
(4.0,7.0]	18497 (46.7%)	12881 (32.5%)
Missing	0 (0%)	11117 (28.0%)

Sample Statistics: Outcome Variables

Table 6: Outcome variables measured at

	2018 (N=39635)	2020 (N=39635)	Overall (N=79270)
Social Belonging			
Mean (SD)	5.14 (1.07)	5.06 (1.09)	5.11 (1.08)
Median [Min, Max]	5.31 [1.00, 7.00]	5.05 [1.00, 7.00]	5.30 [1.00, 7.00]
Missing	7 (0.0%)	13278 (33.5%)	13285 (16.8%)
Anxiety			
Mean (SD)	1.21 (0.774)	1.17 (0.756)	1.19 (0.767)
Median [Min, Max]	1.00 [0, 4.00]	1.00 [0, 4.00]	1.00 [0, 4.00]
Missing	51 (0.1%)	13275 (33.5%)	13326 (16.8%)
Depression			
Mean (SD)	0.584 (0.751)	0.550 (0.723)	0.571 (0.740)
Median [Min, Max]	0.333 [0, 4.00]	0.333 [0, 4.00]	0.333 [0, 4.00]
Missing	54 (0.1%)	13273 (33.5%)	13327 (16.8%)
Life Satisfaction			
Mean (SD)	5.30 (1.20)	5.25 (1.23)	5.28 (1.21)
Median [Min, Max]	5.50 [1.00, 7.00]	5.50 [1.00, 7.00]	5.50 [1.00, 7.00]
Missing	260 (0.7%)	13560 (34.2%)	13820 (17.4%)
Hours of Exercise (log)			
Mean (SD)	1.55 (0.846)	1.63 (0.839)	1.58 (0.844)
Median [Min, Max]	1.61 [0, 4.40]	1.78 [0, 4.40]	1.61 [0, 4.40]
Missing	1242 (3.1%)	13770 (34.7%)	15012 (18.9%)
Meaning: Purpose			
Mean (SD)	5.20 (1.41)	5.15 (1.44)	5.18 (1.42)
Median [Min, Max]	5.05 [1.00, 7.00]	5.04 [1.00, 7.00]	5.04 [1.00, 7.00]
Missing	1010 (2.5%)	13650 (34.4%)	14660 (18.5%)
Meaning: Sense			
Mean (SD)	5.71 (1.22)	5.71 (1.19)	5.71 (1.20)
Median [Min, Max]	5.99 [1.00, 7.00]	5.99 [1.00, 7.00]	5.99 [1.00, 7.00]
Missing	128 (0.3%)	13162 (33.2%)	13290 (16.8%)
Neighbourhood Community			
Mean (SD)	4.19 (1.66)	4.38 (1.57)	4.27 (1.63)
Median [Min, Max]	4.03 [1.00, 7.00]	4.95 [1.00, 7.00]	4.04 [1.00, 7.00]
Missing	212 (0.5%)	13202 (33.3%)	13414 (16.9%)
Personal Well-being Index			
Mean (SD)	7.09 (1.66)	7.18 (1.63)	7.12 (1.65)
Median [Min, Max]	7.29 [0, 10.0]	7.47 [0, 10.0]	7.46 [0, 10.0]
Missing	41 (0.1%)	13120 (33.1%)	13161 (16.6%)
Rumination			
Mean (SD)	0.853 (1.00)	0.797 (0.959)	0.831 (0.987)
Median [Min, Max]	0.955 [0, 4.00]	0.0495 [0, 4.00]	0.953 [0, 4.00]
Missing	135 (0.3%)	13335 (33.6%)	13470 (17.0%)
Self Esteem			
Mean (SD)	5.14 (1.28)	5.13 (1.27)	5.14 (1.28)
Median [Min, Max]	5.34 [1.00, 7.00]	5.34 [1.00, 7.00]	5.34 [1.00, 7.00]
Missing	11 (0.0%)	13280 (33.5%)	13291 (16.8%)

	2018	2020	Overall
Social Support			
Mean (SD)	5.95 (1.12)	5.94 (1.12)	5.95 (1.12)
Median [Min, Max]	6.30 [1.00, 7.00]	6.29 [1.00, 7.00]	6.30 [1.00, 7.00]
Missing	30 (0.1%)	13112 (33.1%)	13142 (16.6%)

Appendix C: Transition Matrix to Check The Positivity Assumption

Table 7: Transition Matrix Showing Change

From / To	State 0	State 1	Total
State 0	17572	2271	19843
State 1	2400	6275	8675

These transition matrices capture shifts in states between consecutive waves. Each cell shows the count of individuals transitioning from one state to another. Rows are the initial state (From), columns the subsequent state (To). **Diagonal entries (in bold)** mark those who stayed in the same state.

Appendix D: RATE AUTOC and RATE Qini

Rate Test

The RATE metric shows how much extra gain (or avoided loss) we achieve by **targeting** instead of treating everyone identically.

Technical note: In code we always set `policy = "treat_best"`; for harmful exposures this is interpreted as '*treat-those-most-sensitive*' (i.e., prioritise protection or withholding).

- **Beneficial exposure:** we rank by positive CATEs and deliver the exposure to those predicted to **benefit most**.
- **Detrimental exposure:** we rank by increasingly **positive** CATEs (more predicted harm) and identify those who should be protected or withheld from the exposure.

Either way, a larger **absolute** RATE shows that a CATE-based targeting rule 'outperforms' a one-size-fits-all policy—by boosting outcomes for beneficial exposures or – in the case where we are explore sensitivity to harm – evaluating increasing harms for detrimental ones.

Recall we flipped Anxiety, Depression, Rumination so '**higher**' always tracks the analysis goal: **higher = more benefit for beneficial exposures, higher = more harm for detrimental exposures**.

Because we test several outcomes, RATE *p*-values are adjusted with Benjamini–Hochberg false-discovery-rate adjustment ($q = 0.1$) before we decide whether heterogeneity is actionable.

Comparison of targeting operating characteristic (TOC) by rank average treatment effect (RATE): AUTOC vs QINI

We applied two TOC by RATE methods to the same causal-forest $\tau(x)$ estimates:

- **AUTOC** intensifies focus on top responders via logarithmic weighting.
- **QINI** balances effect size and prevalence via linear weighting.

Exploratory RATE analysis; controlled FDR at $q=0.20$ over 12 outcomes.

Both methods yield positive RATE estimates for: **Hours of Exercise (log)**.

This concordance indicates robust heterogeneity evidence.

When methods disagree (only QINI yields positive RATE for Anxiety (reversed)), choose **QINI** for overall benefit or **AUTOC** to focus top responders.

Refer to Appendix D for details.

RATE AUTOC RESULTS

Evidence for heterogeneous treatment effects (policy = treat best responders) using AUTOC

AUTOC uses logarithmic weighting to focus treatment on top responders.

Note: The following outcomes were inverted during preprocessing because higher values of the exposure correspond to worse outcomes: Anxiety, Depression, Rumination.

Positive RATE estimates for: **Hours of Exercise (log)**.

Estimates (**Hours of Exercise (log)**): 0.084 (95% CI 0.035, 0.133)) show robust heterogeneity.

Negative RATE estimates for: Neighbourhood Community.

Estimates (Neighbourhood Community: -0.076 (95% CI -0.137, -0.015)) caution against CATE prioritisation.

For outcomes with adjusted p-values not meeting the FDR threshold of $q = 0.20$ (Meaning: Sense, Anxiety (reversed), Rumination (reversed), Self Esteem, Social Support, Life Satisfaction, Meaning: Purpose, Depression (reversed), Social Belonging, Personal Well-being Index), evidence is inconclusive.

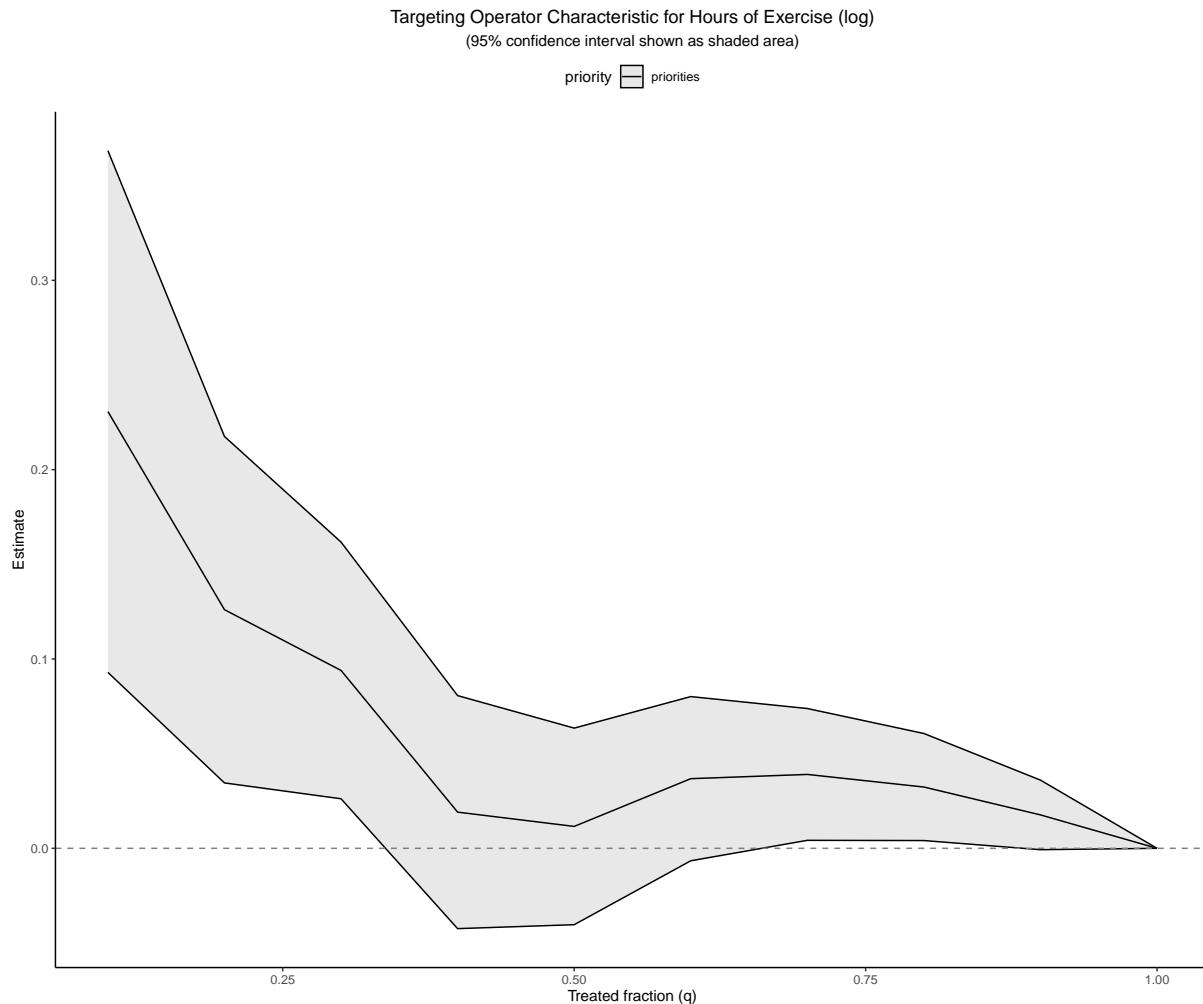


Figure 12: RATE AUTOC Graphs

Figure 12 presents the RATE AUTOC curve for **Hours of Exercise (log)**

Appendix E. Estimating and Interpreting Heterogeneous Treatment Effects with GRF

Appendix F: Strengths and Limitations of Causal Forests

Strengths and Limitations of Our Approach

We used causal forests (Tibshirani et al., 2024) to estimate how treatment effects may differ for individuals with different characteristics. This method is powerful, however it also depends on measuring all major variables influencing both treatment selection and outcomes. If such variables are missed or mismeasured, results can be biased. Additionally, interpreting subgroup effects can be tricky when many characteristics are involved and statistically significant differences may not always translate into meaningful real-world gains.

Despite these concerns, causal forests offer notable advantages. They allow for flexible, non-parametric modelling (Tibshirani et al., 2024), avoiding strict assumptions that might miss complex interactions. We used a robust evaluation method—training our model on half the data and testing it on the remaining half—to avoid overfitting. We then checked whether the predicted differences were genuine and estimated how much benefit we might gain by targeting treatment to those likely to benefit most (Tibshirani et al., 2024; Wager & Athey, 2018). Qini curves (Tibshirani et al., 2024) let us see the overall improvement from treating the top-ranked individuals first, and policy trees (Athey & Wager, 2021b, 2021a; Sverdrup et al., 2024) turn these findings into simple ‘if-then’ rules. Together, this approach provides a practical means of identifying and acting on genuine treatment effect differences.

Here we explain a heterogeneous-treatment-effect (HTE) analysis using causal forests (Tibshirani et al., 2024). In our workflow, we move from the average treatment effect (ATE) to individualised effects, quantify the practical value of targeting, and finish with interpretable decision rules.

1 Average Treatment Effect (ATE)

The ATE answers: ‘*What would happen, on average, if everyone received treatment versus no one?*’

$$\text{ATE} = E[Y(1) - Y(0)].$$

Using the grf package, we estimate the ATE doubly-robustly. Because we analyse several outcomes, we adjust ATE p -values with bonferroni ($\alpha = 0.05$) to control the family-wise error rate.

2 Do Effects Vary? Formal Test of Heterogeneity

Define the conditional average treatment effect (CATE)

$$\tau(x) = E[Y(1) - Y(0) | X = x].$$

If $\tau(x)$ is constant, effects are homogeneous; otherwise they vary. Classical interaction models impose strong forms; grf uses *causal forests* to discover complex, nonlinear heterogeneity (Wager & Athey, 2018). We assess heterogeneity with RATE p -values corrected via Benjamini–Hochberg false-discovery-rate adjustment ($q = 0.1$), controlling the false-discovery rate (Benjamini & Hochberg, 1995).

3 Causal Forests for Individualised Estimates

A causal forest is an ensemble of ‘honest’ causal trees that split on covariates to maximise treated-control contrasts. For each unit i we obtain

$$\hat{\tau}(x_i)$$

Strengths are flexibility, orthogonalisation, and per-person estimates.

4 Built-in Protection Against Over-fitting

Honesty (split half/estimate half) plus out-of-bag (OOB) predictions yield unbiased $\hat{\tau}(x)$ and standard errors without manual hyper-tuning.

5 Missing Data Handling

grf deploys ‘Missing Incorporated in Attributes’ (MIA): missingness is a valid split, so cases stay in the analysis – no ad-hoc imputation required.

6 Testing for Actionable Heterogeneity: the TOC & RATE Metrics

Ranking units by $\hat{\tau}$ defines a **Targeting Operator Characteristic** (TOC) curve: the cumulative gain from treating the top fraction q of predicted responders. Two scalar summaries:

- **RATE AUTOC** – area under the entire TOC; emphasises the very highest responders.
- **RATE Qini** – weighted area with weight q ; rewards sustained gains across larger coverage ([Yadlowsky et al., 2021](#)).

Under $H_0: \tau(x)$ constant, both equal 0. grf::rank_average_treatment_effect() supplies point estimates, standard errors, and t -tests.

Multiplicity control: We adjust AUTOC and Qini p -values with Benjamini–Hochberg false-discovery-rate adjustment ($q = 0.1$) before declaring actionable heterogeneity.

Here is an **interpretation tip**:

- AUTOC answers ‘*How sharply can we prioritise?*’
- Qini answers ‘*How valuable is targeting when budgets are modest but not tiny?*’

7 Visualising Policy Value: the Qini Curve

Plotting the Qini curve (cumulative gain vs q) reveals where returns plateau. Investigators (and policy audiences) can see at a glance whether benefits concentrate in, say, the top 20 % or persist up to 50 %.

8 Valid Inference for RATE / Qini

Although OOB predictions are out-of-sample per tree, they inherit forest-level dependence. We use an explicit **sample split**:

1. **Train set:** fit the causal forest and compute $\hat{\tau}(x)$.
2. **Test set:** compute RATE AUTOC/Qini and run H_0 tests.

This second split yields honest policy evaluation and guards against optimistic bias ([Tibshirani et al., 2024](#)).

9 From Black Box to Simple Rules: Policy Trees

Stakeholders value transparent criteria. The **policytree** algorithm takes $\hat{\tau}(x)$ or doubly-robust scores and learns a shallow decision tree that maximises expected welfare ([Sverdrup et al., 2024](#)).

Advantages: interpretability, the possibility of fairness constraints, and easy communication (e.g., ‘*treat if age < 25 and baseline severity high*’).

Training mirrors the split above: learn the tree on one fold, evaluate welfare on another.

Caveat Splits identify predictors of *effect variation*, not causal levers. Changing a covariate in the tree does **not** guarantee an effect on $\tau(x)$.

10 Ethical and Practical Considerations

Statistical optimisation rarely aligns perfectly with equity or political feasibility. Decisions about who *should* receive treatment belong to democratic processes that weigh fairness, cost, and broader social values.

Putting it together

The sequence—ATE, causal-forest CATEs, RATE/Qini diagnostics, Qini curve, and finally a shallow policy tree—delivers both rigorous evidence and a defensible targeting rule. Researchers learn **how large** heterogeneity is, **where** targeting pays off under budget constraints, and **which** simple covariate splits capture most of the welfare gain, all while guarding against over-fitting and multiplicity.

References

- Athey, S., & Wager, S. (2021b). Policy Learning With Observational Data. *Econometrica*, 89(1), 133–161. <https://doi.org/10.3982/ECTA15732>
- Athey, S., & Wager, S. (2021a). Policy learning with observational data. *Econometrica*, 89(1), 133–161. <https://doi.org/https://doi.org/10.3982/ECTA15732>
- Atkinson, J., Salmond, C., & Crampton, P. (2019). *NZDep2018 index of deprivation, user's manual*.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), 289–300.
- Bulbulia, J. A. (2024a). Margot: MARGinal observational treatment-effects. <https://doi.org/10.5281/zenodo.10907724>
- Bulbulia, J. A. (2024b). Methods in causal inference part 3: Measurement error and external validity threats. *Evolutionary Human Sciences*, 6, e42. <https://doi.org/10.1017/ehs.2024.33>
- Cutrona, C. E., & Russell, D. W. (1987). The provisions of social relationships and adaptation to stress. *Advances in Personal Relationships*, 1, 37–67.
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, 49(1), 71–75.
- Fahy, K. M., Lee, A., & Milne, B. J. (2017). *New Zealand socio-economic index 2013*. Statistics New Zealand-Tatauranga Aotearoa.
- Fraser, G., Bulbulia, J., Greaves, L. M., Wilson, M. S., & Sibley, C. G. (2020). Coding responses to an open-ended gender measure in a New Zealand national sample. *The Journal of Sex Research*, 57(8), 979–986. <https://doi.org/10.1080/00224499.2019.1687640>
- Greaves, L. M., Barlow, F. K., Lee, C. H., Matika, C. M., Wang, W., Lindsay, C.-J., Case, C. J., Sengupta, N. K., Huang, Y., Cowie, L. J., et al. (2017). The diversity and prevalence of sexual orientation self-labels in a New Zealand national sample. *Archives of Sexual Behavior*, 46, 1325–1336.
- Hagerty, B. M. K., & Patusky, K. (1995). Developing a Measure Of Sense of Belonging: *Nursing Research*, 44(1), 9–13. <https://doi.org/10.1097/00006199-199501000-00003>
- Health, Ministry of. (2013). *The New Zealand Health Survey: Content guide 2012-2013*. Princeton University Press.
- Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8), 758–764.
- Instrument Ware Jr, J., & Sherbourne, C. (1992). The MOS 36-item short-form health survey (SF-36): I. Conceptual framework and item selection. *Medical Care*, 30(6), 473–483.
- Jost, J. T. (2006). The end of the end of ideology. *American Psychologist*, 61(7), 651–670. <https://doi.org/10.1037/0003-066X.61.7.651>
- Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, 32(6), 959–976. <https://doi.org/10.1017/S0033291702006074>
- Linden, A., Mathur, M. B., & VanderWeele, T. J. (2020). Conducting sensitivity analysis for unmeasured confounding in observational studies using e-values: The evalute package. *The Stata Journal*, 20(1), 162–175.
- Nolen-hoeksema, S., & Morrow, J. (1993). Effects of rumination and distraction on naturally occurring depressed mood. *Cognition and Emotion*, 7(6), 561–570. <https://doi.org/10.1080/02699939308409206>
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
- Sengupta, N. K., Luyten, N., Greaves, L. M., Osborne, D., Robertson, A., Brunton, C., Armstrong, G., & Sibley, C. G. (2013). Sense of Community in New Zealand Neighbourhoods: A Multi-Level Model Predicting Social Capital. *New Zealand Journal of Psychology*, 42(1), 36–45.
- Sibley, C. G. (2021). *Sampling procedure and sample details for the New Zealand Attitudes and Values Study*. <https://doi.org/10.31234/osf.io/wgqvy>
- Sibley, C. G., Luyten, N., Purnomo, M., Mobberley, A., Wootton, L. W., Hammond, M. D., Sengupta, N., Perry, R., West-Newman, T., Wilson, M. S., McLellan, L., Hoverd, W. J., & Robertson, A. (2011). The Mini-IPIP6:

- Validation and extension of a short measure of the Big-Six factors of personality in New Zealand. *New Zealand Journal of Psychology*, 40(3), 142–159.
- Statistics New Zealand. (2017). *Statistical standard for geographic areas 2018 (SSGA18)*. Statistics New Zealand. <https://www.stats.govt.nz/methods/statistical-standard-for-geographic-areas-2018/>
- Steger, M. F., Frazier, P., Oishi, S., & Kaler, M. (2006). The meaning in life questionnaire: Assessing the presence of and search for meaning in life. *Journal of Counseling Psychology*, 53(1), 80–93. <https://doi.org/10.1037/0022-0167.53.1.80>
- Sverdrup, E., Kanodia, A., Zhou, Z., Athey, S., & Wager, S. (2024). *Policytree: Policy learning via doubly robust empirical welfare maximization over trees*. <https://CRAN.R-project.org/package=policytree>
- Tibshirani, J., Athey, S., Sverdrup, E., & Wager, S. (2024). *Grf: Generalized random forests*. <https://github.com/grf-labs/grf>
- VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3), 211–219.
- VanderWeele, T. J., & Ding, P. (2017). Sensitivity analysis in observational research: Introducing the E-value. *Annals of Internal Medicine*, 167(4), 268–274. <https://doi.org/10.7326/M16-2607>
- VanderWeele, T. J., Mathur, M. B., & Chen, Y. (2020). Outcome-wide longitudinal designs for causal inference: A new template for empirical studies. *Statistical Science*, 35(3), 437–466.
- Verbrugge, L. M. (1997). A global disability indicator. *Journal of Aging Studies*, 11(4), 337–362. [https://doi.org/10.1016/S0890-4065\(97\)90026-8](https://doi.org/10.1016/S0890-4065(97)90026-8)
- Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
- Whitehead, J., Davie, G., Graaf, B. de, Crengle, S., Lawrenson, R., Miller, R., & Nixon, G. (2023). Unmasking hidden disparities: A comparative observational study examining the impact of different rurality classifications for health research in aotearoa new zealand. *BMJ Open*, 13(4), e067927.
- Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E., & Wager, S. (2021). Evaluating treatment prioritization rules via rank-weighted average treatment effects. *arXiv Preprint arXiv:2111.07966*. <https://doi.org/10.48550/arXiv.2111.07966>