Huo, Shutong; Feng, Derek; Gill, Thomas M.; Chen, Xi

Working Paper

Childhood Circumstances and Health of American and Chinese Older Adults: A Machine Learning Evaluation of Inequality of Opportunity in Health

GLO Discussion Paper, No. 1384

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Huo, Shutong; Feng, Derek; Gill, Thomas M.; Chen, Xi (2024) : Childhood Circumstances and Health of American and Chinese Older Adults: A Machine Learning Evaluation of Inequality of Opportunity in Health, GLO Discussion Paper, No. 1384, Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/281669

Childhood Circumstances and Health of American and Chinese Older Adults: A Machine Learning Evaluation of Inequality of Opportunity in Health

Shutong Huo
University of California, Irvine

Derek Feng
Yale University

Thomas M. Gill
Yale University

Xi Chen*
Yale University and GLO

Abstract

Childhood circumstances may impact senior health, prompting this study to introduce novel machine learning methods to assess their individual and collective contributions to health inequality in old age. Using the US Health and Retirement Study (HRS) and the China Health and Retirement Longitudinal Study (CHARLS), we analyzed health outcomes of American and Chinese participants aged 60 and above. Conditional inference trees and forest were employed to estimate the influence of childhood circumstances on self-rated health (SRH), comparing with the conventional parametric Roemer method. The conventional parametric Roemer method estimated higher IOP in health (China: 0.039, 22.67% of the total Gini coefficient 0.172; US: 0.067, 35.08% of the total Gini coefficient 0.191) than conditional inference tree (China: 0.022, 12.79% of 0.172; US: 0.044, 23.04% of 0.191) and forest (China: 0.035, 20.35% of 0.172; US: 0.054, 28.27% of 0.191). Key determinants of health in old age were identified, including childhood health, family financial status, and regional differences. The conditional inference forest consistently outperformed other methods in predictive accuracy as measured by out-of-sample mean squared error (MSE). The findings demonstrate the importance of early-life circumstances in shaping later health outcomes and stress the early-life interventions for health equity in aging societies. Our methods highlight the utility of machine learning in public health to identify determinants of health inequality.

Keywords: Life Course, Inequality of Opportunity, Childhood Circumstances, Machine Learning, Conditional Inference Tree, Random Forest

JEL Codes: I14, J13, J14, O57, C53

*Corresponding author: Xi Chen, Yale University and GLO. 60 College St, New Haven, CT 06520, USA. xi.chen@yale.edu

1. Introduction

A global trend of rapid population aging alongside an increase in health burden among older adults necessitates a better understanding of the lasting imprint of early life stages on the aging process (1). Prior economics and epidemiology research converges on the assertion that various childhood circumstances significantly influence health later in life, thus indicating childhood as a critical window for interventions to narrow health disparities (2). These circumstances encompass diverse factors related to parents (3), family socioeconomic status (SES) (4), and community or higher-level factors such as rural/urban status (5) and natural environments (6).

While both early-life and later-life factors contribute to health outcomes in older age, childhood circumstances, especially those beyond an individual's control, are considered the most unacceptable and illegitimate sources of health inequality in older ages (7,8). The form of inequality attributable to childhood circumstances is often named inequality of opportunity (IOP). The prioritization of reducing IOP emerges from a broad political and social dialogue that seeks to level the playing field in early stages of life and address the unfair health inequalities identified by the WHO Commission on Social Determinants of Health (9).

Despite extensive research on the influence of childhood circumstances on health outcomes, methodological challenges persist, such as arbitrary selection of childhood circumstances and potential biases in estimating health inequality among older adults (10,11). Our study addressed these issues by employing machine learning methods to select the appropriate set of childhood circumstances. This approach allowed the data to guide the understanding of unequal childhood circumstances, minimizing the imposition of researcher bias on the model specification (10,12). Additionally, we compared the outcomes of our approach with those of the conventional parametric Roemer method to highlight our substantial improvements in measuring inequality over the life course.

2. Methods

Our study leveraged the Health and Retirement Study (HRS) and the China Health and Retirement Longitudinal Study (CHARLS), analyzing the matched 2020-2021 wave HRS in the US and the 2020 wave CHARLS in China with respective life history survey. The final analytic sample comprised 2,434 Americans and 5,612 Chinese, aged 60 and above. We used self-rated health (SRH) as the health outcome measure, ascertained similarly across both surveys on a scale from excellent(=1) to poor(=5). Data on 43 childhood circumstances from

HRS and 36 from CHARLS, both categorized into seven domains, such as birth environment, family SES, and relationships in childhood, were included in the analysis (Appendix A). Despite slight variations, the domains essentially contained the same core measures for both countries. We used R 4.3.1 for the analysis.

Appendix B details the complete conceptual and analytic framework for this study. Firstly, we used the conventional parametric Roemer method (with Shapley value decomposition) to estimate to what extent childhood circumstances individually and collectively contribute to health inequality in later life, setting the stage for policy intervention evaluation. A counterfactual health outcome distribution can be derived by partitioning the population into non-overlapping homogeneous groups using observable circumstances. For example, consider two binary childhood circumstances, i.e., parental education (high v. low) and financial hardship (yes v. no), which can classify all samples into four non-overlapping groups. Health inequality across the four groups can solely be attributed to differences in childhood circumstances, defining the IOP. In this study, we measured the extent to which childhood circumstances contribute to health inequality by Gini coefficient (8,11). We also divide this absolute health inequality measure by the overall health inequality to obtain IOP, i.e. the fraction of health inequality explained by childhood circumstances. Despite not being causal, it offered insight into statistical importance of childhood circumstances (13).

Conditional inference trees, with their sequential hypothesis tests, are especially useful in the context of the IOP analysis, offering a graphical illustration for comparing childhood circumstances. Each test probes IOP within a subsample. The deeper the tree, the more varied childhood circumstances within a society. Furthermore, these trees alleviate issues of arbitrary variable and model selection that afflict the IOP literature, retaining a complete set of observed variables qualifying as childhood circumstances. Specifically, we used these childhood circumstances to divide the population into non-overlapping groups, i.e., terminal nodes in the regression tree context. The predicted value for the outcome of an individual observation was then calculated as the mean outcome of the group to which the individual was assigned, with a number of observations in that group. We also used K-fold cross-validation to tune the model parameters to perform optimally. We used 5 folds in this paper. Our results are robust to the choice of K.

Conditional inference trees, while providing non-arbitrary population segmentation, have limitations. They use limited data, struggle with highly correlated childhood circumstances, and show high prediction variance, making them sensitive to sample changes. However, random forest mitigates these issues by decorrelating trees, reducing prediction variance. They form forest of decision trees from bootstrapped samples, using a random selection of predictors at each split, making the model more reliable. This paper used 200 trees, a number determined based on computational cost-efficiency and prediction accuracy, to predict outcomes (Appendix C). Half of the observations were randomly selected in each tree, following a 4-step method using random data subsamples and random subsets of circumstances to determine the optimal parameters via out-of-bag error minimization. Although the collection of bagged trees was much more difficult to interpret than a single tree, we gauged predictor importance of each childhood circumstance using the residual sum of squares (RSS).

To assess potentials of both downward and upward biases of IOP in health that may affect out-of-sample performance, we followed the standard practice to split sample into a training set (2/3*N) and a test set (1/3*N). We fitted our model on the training set and compared the performance on the test set for the conventional parametric Roemer method, conditional inference trees, and conditional inference forest, respectively.

3. Results

First, inequality in self-rated health, as measured by the Gini coefficient, was higher in the US than in China. Next, we measured IOP using the Gini coefficients in the counterfactual distribution. Figure 1 demonstrated that the conventional parametric Roemer method produced the highest IOP estimates, followed by the conditional inference forest method and then the conditional inference tree method. Specifically, in China IOP explained 22.67% (0.039 of 0.172 total Gini coefficient) of inequality in self-rated health, and in the US it accounted for 35.08% (0.067 of 0.191 total Gini coefficient). In contrast, the conditional inference tree method accounted for 12.79% in China (0.022 of 0.172 total Gini coefficient) and 23.04% in the US (0.044 of 0.191 total Gini coefficient), while the forest method represented 20.35% in China (0.035 of 0.172 total Gini coefficient) and 28.27% in the US (0.054 of 0.191 total Gini coefficient).

Figure 2A portrayed the IOP structure for self-rated health in China, using a tree with five terminal nodes. Factors such as childhood health, birth region, and childhood family financial

status formed a tree. The most advantaged type (terminal node 5) included people with good childhood health and family financial status, and born in Eastern China. In contrast, the group in the worst self-rated health (terminal node 6) typically had poorer child health. In the US, as shown in Figure 2B, those of poor childhood health fell into disadvantaged circumstance type (terminal nodes 7). In contrast, individuals of certain favorable conditions, such as having more books at home, being healthy in childhood, and being White, generally reported better health in old age (terminal node 6).

Figure 3A revealed that in China, using conditional inference forest, the key factors impacting self-rated health were childhood health and being born in Eastern China, corroborating findings from the conditional inference trees (Figure 2A). In addition, parents' health status (staying long time in bed) and relationship with parents also highly impacted self-rated health in older ages. Likewise, Figure 3B demonstrated that in the US, childhood health, number of books at home at age 10, and race/ethnicity emerged as significant factors, largely aligning with results obtained through conditional inference trees (Figure 2B).

As illustrated before, all models we tested aim to minimize the mean squared error (MSE). 95% confidence intervals are derived based on 200 bootstrapped re-sampling of the test data. The MSE for random forest was standardized to 1 to facilitate comparison in prediction performance across models, such that a MSE larger than 1 represented a worse out-of-sample fit. Both conditional inference tree and parametric Roemer methods performed worse than conditional inference forest in self-rated health (Figure 4A and Figure 4B). On average, conditional inference trees demonstrated smaller test error rates than the conventional parametric Roemer method.

4. Discussion

This study introduced two machine learning methods, i.e., the conditional inference tree and forest, to examine how a comprehensive set of childhood circumstances may contribute to health inequality among older adults in China and US, respectively. We identified several leading predictors of health conditions in older adults, such as childhood health, socioeconomic status, number of books at home for Americans, as well as birth region for Chinese. These methods addressed concerns over the arbitrary selection of childhood circumstances, while balancing the potential biases in IOP estimates. Our findings underscore the importance of addressing health inequality stemming from childhood circumstances, indicating policy and

intervention strategies for health equity in both China and US. Preventive measures from childhood can reduce the economic burden of diseases and enhance quality of life and longevity, especially when advances in medicine still lack effective treatments to slow or reverse the progression of chronic diseases like Alzheimer's, hypertension, and diabetes.

The superiority of conditional inference forest in out-of-sample performance, rendering the most accurate estimates of childhood circumstances on health inequality in old age, aligns with previous studies in other fields (14,15). While conditional inference trees offered a less complex model and a convenient visual illustration of childhood circumstances structure, forest utilized information on childhood circumstances more efficiently, offering results consistent with trees regarding IOP estimates and importance assigned to specific circumstances. These machine learning methods employed explicit algorithms to interpret health outcomes, making no strong assumptions about which childhood circumstances significantly influence health outcomes. Using statistical methods like K-fold cross-validation and bootstrap, our modeling became more transparent and generalizable.

This study has limitations. First, our life course approach focused on current older adults and may not necessarily reflect the realities of younger cohorts. Future research should monitor younger cohorts. Second, the associations identified in the current study should not be interpreted as causal. For example, it is possible that unobservable childhood circumstances may bias our estimates. Additional research is needed to identify causal channels. Third, this analysis utilized the most recently released CHARLS (2020) and HRS (2020-2021) surveys, their overlap with the COVID-19 pandemic may bias self-rated health. That said, our robustness checks using CHARLS/HRS pre-pandemic waves provided reassuringly consistent results.

In conclusion, our research used a life course approach and machine learning methods to identify critical determinants of health in old age, applying this to the world's two largest economies and aging societies. Our results reinforce the necessity of adopting a life course perspective in public health research and policy direction.

Conflicts of interest: The authors declare no conflict of interest.

Funding: XC acknowledges research funding from the U.S. National Institute on Aging (R01AG077529; K01AG053408; P30AG021342). The views expressed are those of the authors and not necessarily those of the funders. The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. The GLO Discussion Paper Series serves as a preprint server to deposit latest research for feedback.

References

1. Moffitt TE, Belsky DW, Danese A, Poulton R, Caspi A. The Longitudinal Study of Aging in Human Young Adults: Knowledge Gaps and Research Agenda. J Gerontol A Biol Sci Med Sci. 2017 Feb;72(2):210–5.

2. Bor J, Cohen GH, Galea S. Population health in an era of rising income inequality: USA, 1980–2015. The Lancet. 2017 Apr 8;389(10077):1475–90.

3. Carrieri V, Jones AM. Inequality of opportunity in health: A decomposition-based approach. Health Econ. 2018;27(12):1981–95.

4. Moody-Ayers S, Lindquist K, Sen S, Covinsky KE. Childhood Social and Economic Well-Being and Health in Older Age. Am J Epidemiol. 2007 Nov 1;166(9):1059–67.

5. Strauss J, Witoelar F, Meng Q, Chen X, Zhao Y, Sikoki B, et al. Cognition and SES Relationships Among the Mid-Aged and Elderly: A Comparison of China and Indonesia [Internet]. National Bureau of Economic Research; 2018 [cited 2023 Aug 10]. (Working Paper Series). Available from: https://www.nber.org/papers/w24583

6. Isen A, Rossin-Slater M, Walker WR. Every Breath You Take—Every Dollar You'll Make: The Long-Term Consequences of the Clean Air Act of 1970. J Polit Econ. 2017 Jun;125(3):848–902.

7. ROEMER JE. Equality of Opportunity. Harvard University Press; 1998. 130 p.

8. Roemer J, Trannoy A. Equality of Opportunity: Theory and Measurement. J Econ Lit. 2016 Dec 1;54:1288–332.

9. Marmot M, Friel S, Bell R, Houweling TAJ, Taylor S, Commission on Social Determinants of Health. Closing the gap in a generation: health equity through action on the social determinants of health. Lancet Lond Engl. 2008 Nov 8;372(9650):1661–9.

10. Brunori P, Hufe P, Mahler D. The roots of inequality: estimating inequality of opportunity from regression trees and forests*. Scand J Econ [Internet]. 2023 Feb 20 [cited 2023 Aug 17];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/sjoe.12530

11. Ferreira FHG, Gignoux J. The Measurement of Inequality of Opportunity: Theory and an Application to Latin America. Rev Income Wealth. 2011;57(4):622–57.

12. Hufe P, Peichl A, Roemer J, Ungerer M. Inequality of income acquisition: the role of childhood circumstances. Soc Choice Welf. 2017 Dec 1;49(3):499–544.

13. Ferreira FHG, Gignoux J. The Measurement of Educational Inequality: Achievement and Opportunity. World Bank Econ Rev. 2014 May 27;28(2):210–46.

14. Qi Y. Random Forest for Bioinformatics. In: Zhang C, Ma Y, editors. Ensemble Machine Learning [Internet]. New York, NY: Springer New York; 2012 [cited 2023 Aug 17]. p. 307–23. Available from: https://link.springer.com/10.1007/978-1-4419-9326-7_11

15. Schneider J, Hapfelmeier A, Thöres S, Obermeier A, Schulz C, Pförringer D, et al. Mortality Risk for Acute Cholangitis (MAC): a risk prediction model for in-hospital mortality in patients with acute cholangitis. BMC Gastroenterol. 2016 Feb 9;16(1):15.

Figure 1 Correlation of Estimates by Method

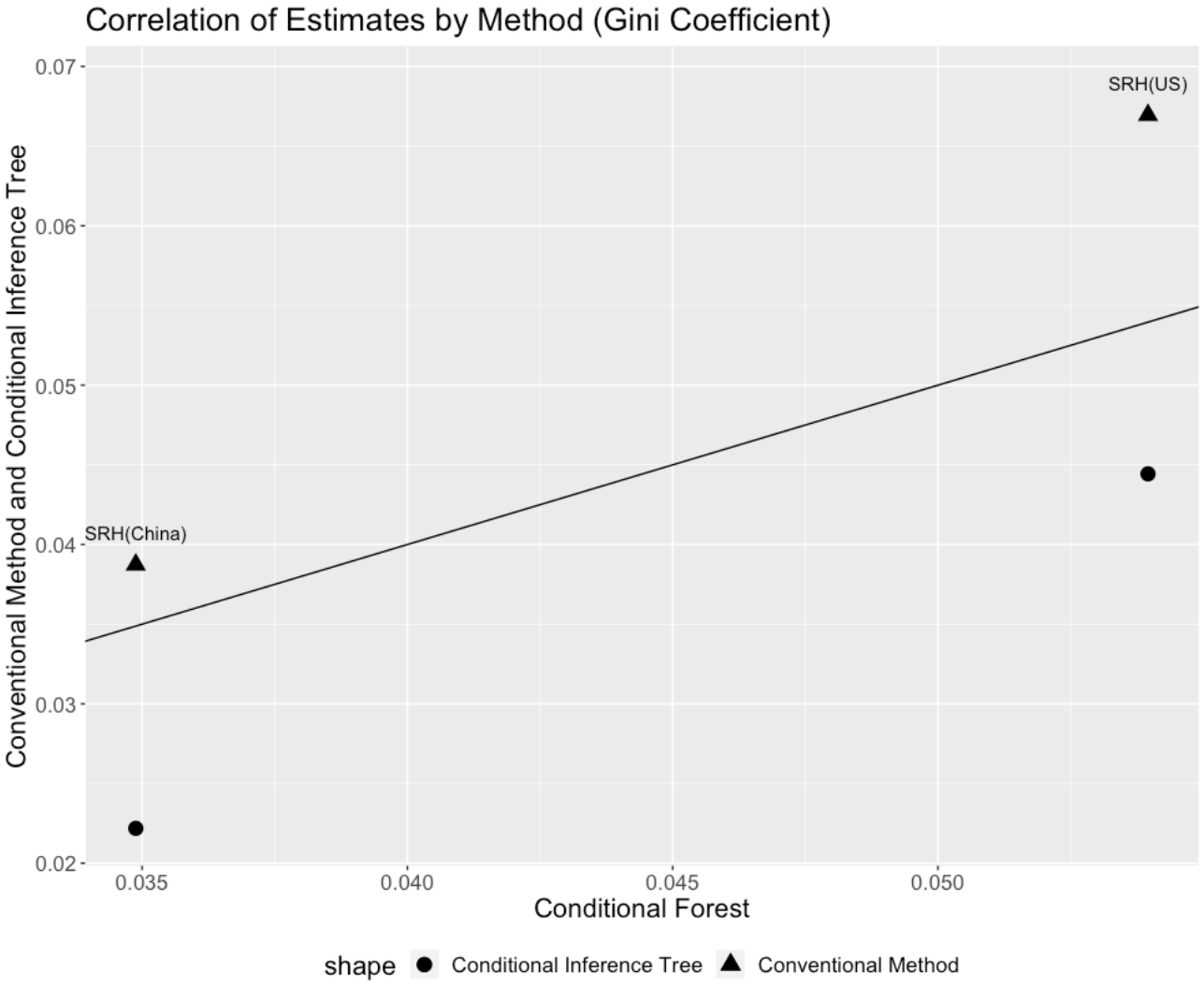


Figure 2. Conditional Inference Tree for Self-Rated Health, China (A) and the US (B)

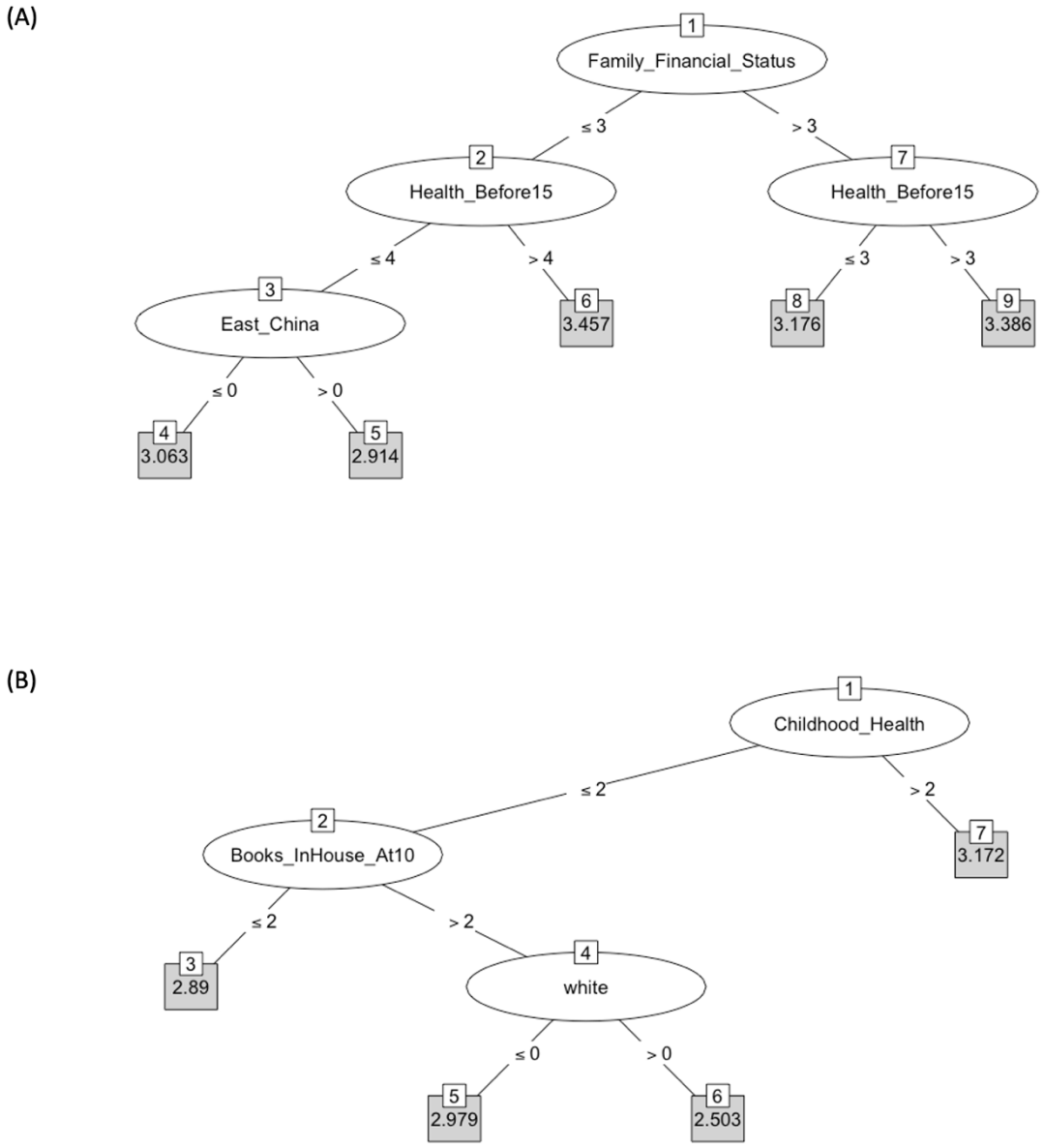


Figure 3. Importance of Childhood Circumstances to Self-Rated Health using Conditional Inference Forest, China (A) and the US (B)

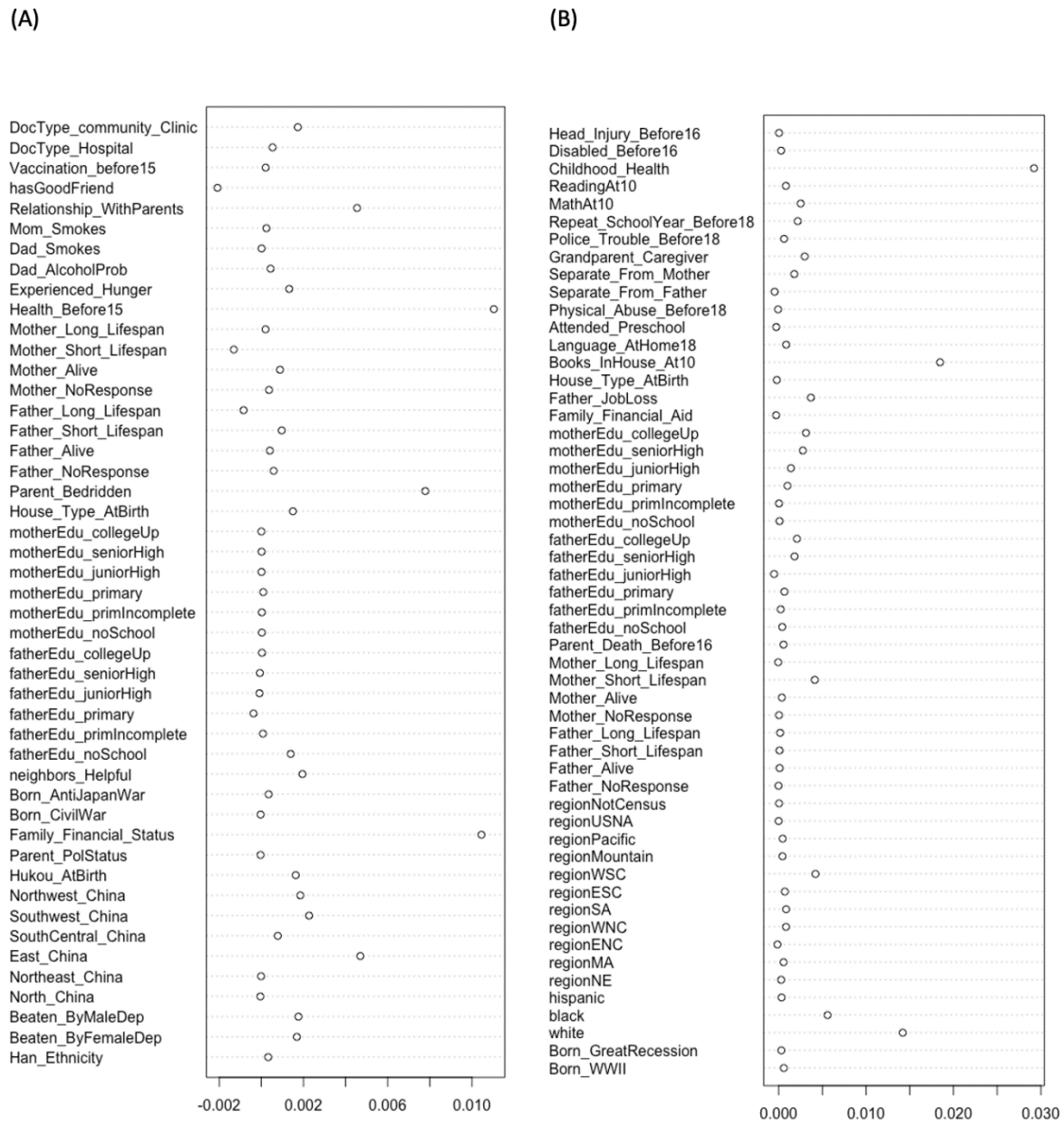
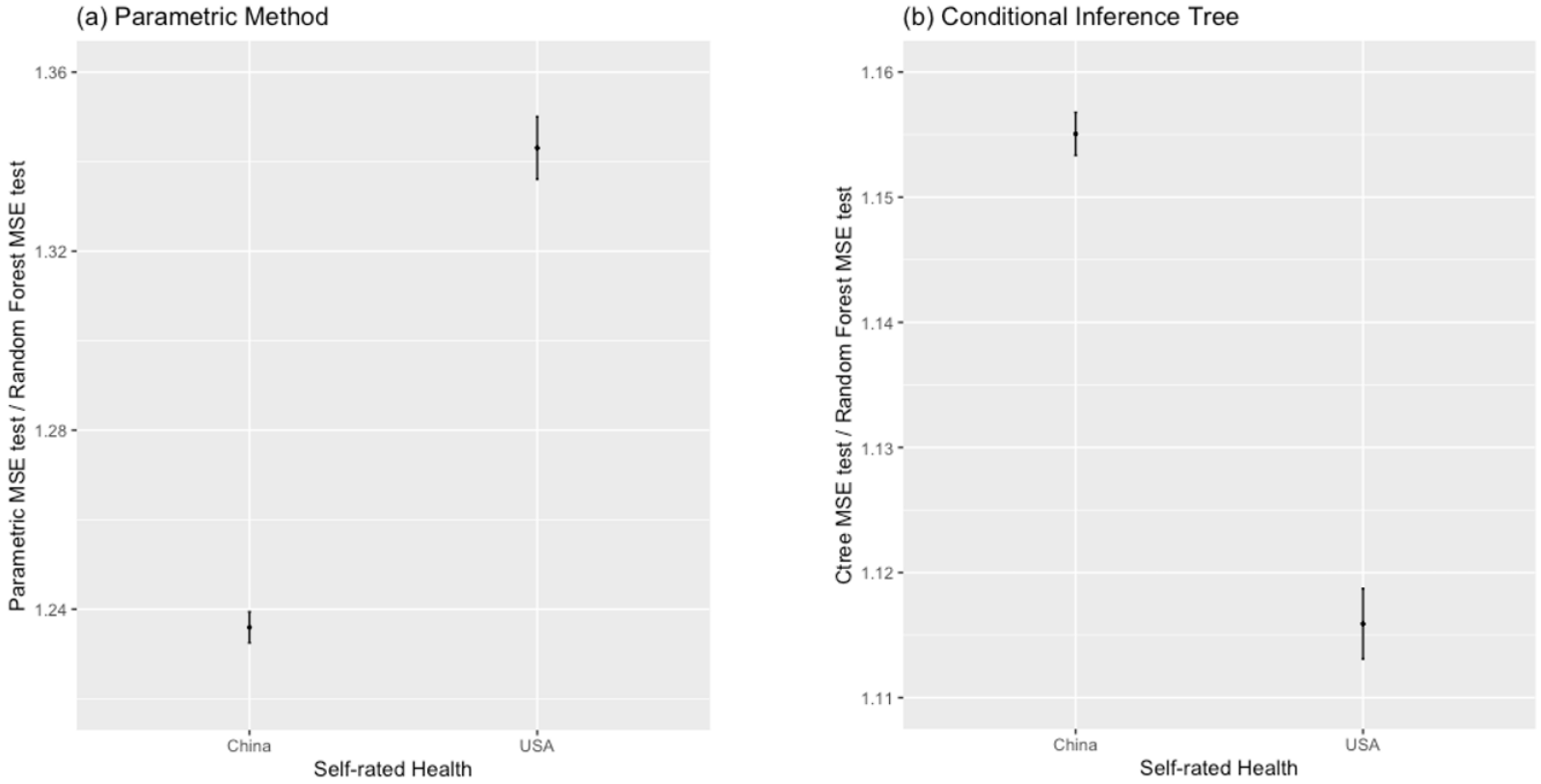


Figure 4 Comparison of Models' Test Error, (a) Parametric Method vs. Random Forest and (b) Conditional Inference Trees vs. Random Forest



Note: All models aim to minimize the mean squared error (MSE). MSE from Random Forest is used as the reference group. Ratios larger than 1 means the corresponding methods and outcome measures generate larger MSE than using Random Forest. 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data.

Supplementary Materials

Appendix A: Summary statistics of health outcome and childhood circumstances in the US and China

Appendix B: Conceptual and Analytic Framework

Appendix C: Optimal Number of Trees

Appendix A: Summary statistics of health outcome and childhood circumstances in the US and China

Table A1 Summary statistics of self-rated health in the US and China

Variable	Country	Obs	Mean	Std.Dev.	Min	Max	Variable Description	CV
Self-Rated health	US	2,434	2.835	0.994	1	5	The value of self-rated health in 2020-2021 (Would you say your health is excellent, very good, good, fair, or poor? 1. excellent, 2.very good, 3.good, 4.fair, 5.poor.)	0.351
	CHN	5,612	3.879	0.772	1	5	The value of self-rated health in 2020 (Would you say your health is excellent, very good, good, fair, or poor? 1. excellent, 2.very good, 3.good, 4.fair, 5.poor.)	0.199

Table A2 Summary statistics of childhood circumstances in the US and China

Domain	Country	Obs	Mean	SD	Min	Max	Variable Description
War or economic crisis	US (2)	2,434	0.077	0.267	0	1	Born in the Great Recession during 1929-1933 (1: Yes; 0: No)
		2,434	0.190	0.392	0	1	Born in the World War II during 1941-1945 (1: Yes; 0: No)
	CHN (2)	5,612	0.295	0.456	0	1	Born in the Anti-Japan War during 1937-1945 (1: Yes; 0: No)
		5,612	0.274	0.446	0	1	Born in the Civil War during 1946-1949 (1: Yes; 0: No)
Regional and urban/ rural status	US (11)	2,434	0.051	0.220	0	1	Northeast region: new England division (me, nh, vt, ma, ri, ct) (1:Yes; 0: No)
		2,434	0.147	0.354	0	1	Northeast region: middle Atlantic division (ny, nj, pa) (1:Yes; 0: No)
		2,434	0.199	0.399	0	1	Midwest region: east north central division (oh, in, il, mi, wi) (1:Yes; 0: No)
		2,434	0.115	0.319	0	1	Midwest region: west north central division (mn, ia, mo, nd, sd, ne, ks) (1:Yes; 0: No)
		2,434	0.154	0.361	0	1	South region: south Atlantic division (de, md, dc, va, wv, nc, sc, ga, fl) (1:Yes; 0: No)
		2,434	0.082	0.274	0	1	South region: east south central division (ky, tn, al, ms) (1:Yes; 0: No)
		2,434	0.091	0.287	0	1	South region: west south central division (ar, la, ok, tx) (1:Yes; 0: No)
		2,434	0.032	0.175	0	1	West region: mountain division (mt, id, wy, co, nm, az, ut, nv) (1:Yes; 0: No)
		2,434	0.063	0.244	0	1	West region: pacific division (wa, or, ca, ak, hi) (1:Yes; 0: No)
		2,434	0.008	0.091	0	1	U.S., na state (1:Yes; 0: No)
		2,434	0.058	0.234	0	1	Foreign country: not in a census division (includes U.S territories) (1:Yes; 0: No)

	CHN (7)	5,612	0.099	0.299	0	1	Rural or urban status at birth (0: rural; 1: urban)
		5,612	0.106	0.308	0	1	Northern China (1:Yes; 0: No)
		5,612	0.074	0.262	0	1	Northeastern China (1:Yes; 0: No)
		5,612	0.328	0.469	0	1	Eastern China (1:Yes; 0: No)
		5,612	0.241	0.427	0	1	South Central China (1:Yes; 0: No)
		5,612	0.181	0.385	0	1	Southwestern China (1:Yes; 0: No)
		5,612	0.070	0.255	0	1	Northwestern China (1:Yes; 0: No)
Family socioeconomic status	US (10)	2,434	0.020	0.140	0	1	Father: No schooling (1:Yes; 0: No)
		2,434	0.776	0.008	0	1	Ethnicity: white (1: Yes; 0: No)
		2,434	0.149	0.006	0	1	Ethnicity: black (1: Yes; 0: No)
		2,434	0.049	0.004	0	1	Ethnicity: Hispanic (1: Yes; 0: No)
		2,434	0.062	0.242	0	1	Father: educated without completing primary school(1:Yes; 0: No)
		2,434	0.136	0.342	0	1	Father: Graduated from primary school(1:Yes; 0: No)
		2,434	0.300	0.458	0	1	Father: Graduated from junior high school(1:Yes; 0: No)
		2,434	0.325	0.468	0	1	Father: Graduated from senior high school(1:Yes; 0: No)
		2,434	0.157	0.364	0	1	Father: Graduated from college or above(1:Yes; 0: No)
		2,434	0.018	0.134	0	1	Mother: No schooling (1:Yes; 0: No)
		2,434	0.035	0.183	0	1	Mother: educated without completing primary school(1:Yes; 0: No)
		2,434	0.108	0.311	0	1	Mother: Graduated from primary school(1:Yes; 0: No)
		2,434	0.271	0.445	0	1	Mother: Graduated from junior high school(1:Yes; 0: No)
		2,434	0.430	0.495	0	1	Mother: Graduated from senior high school(1:Yes; 0: No)

	2,434	0.138	0.345	0	1	Mother: Graduated from college or above(1:Yes; 0: No)
	2,434	0.147	0.355	0	1	Family received financial help (1: yes; 0: no)
	2,434	0.443	0.016	0	3	Father lost job (1:yes, no job for several month or longer; 2: yes, never worked/always disable; 3: yes, never lived with father/ father was not alive in childhood; 0: No)
	2,434	0.225	0.008	0	1	Before age 16, one or both parents died (1: Yes; 0: No)
	2,434	0.875	0.330	0	1	Type of house at birth (1: single-family house;0 apartment/townhouse/condo or: mobile home)
	2,434	2.153	1.132	1	5	When you were age 10, approximately how many books were in the place you lived? (1: <=10; 2: 11-27; 3: 27-100; 4:101-200; 5: >200)
	2,434	0.940	0.238	0	1	Was English the language that you usually spoke at home when you were growing up, before you were age 18?
	2,434	0.131	0.337	0	1	Did you attend any organized pre-school program (1: yes; 0: no)
CHN (5)	8585	0.075	0.263	0	1	parents' political status (1:either father or mother is party member; 0: None of them are)
	7795	0.654	0.476	0	1	Father: No schooling (1:Yes; 0: No)
	7795	0.212	0.409	0	1	Father: educated without completing primary school(1:Yes; 0: No)
	7795	0.082	0.276	0	1	Father: Graduated from primary school(1:Yes; 0: No)
	7795	0.027	0.163	0	1	Father: Graduated from junior high school(1:Yes; 0: No)
	7795	0.015	0.121	0	1	Father: Graduated from senior high school(1:Yes; 0: No)

		7795	0.009	0.095	0	1	Father: Graduated from college or above(1:Yes; 0: No)
		8156	0.945	0.228	0	1	Mother: No schooling (1:Yes; 0: No)
		8156	0.032	0.177	0	1	Mother: educated without completing primary school(1:Yes; 0: No)
		8156	0.015	0.123	0	1	Mother: Graduated from primary school(1:Yes; 0: No)
		8156	0.004	0.062	0	1	Mother: Graduated from junior high school(1:Yes; 0: No)
		8156	0.003	0.053	0	1	Mother: Graduated from senior high school(1:Yes; 0: No)
		8156	0.001	0.022	0	1	Mother: Graduated from college or above(1:Yes; 0: No)
		8484	3.559	0.996	1	5	Family financial status (1:a lot better; 2: somewhat better; 3: same as; 4: somewhat worse; 5: a lot worse)
		8552	2.168	0.621	1	3	Type of house at birth (1: concrete; 2 adobe; 3 wood or others)
Parents' health status and health behaviors	US (8)	2,434	0.011	0.103	0	1	Non-response (1: yes; 0: no)
		2,434	0.047	0.211	0	1	Alive (1: yes; 0:no)
		2,434	0.422	0.494	0	1	Short longevity (1: yes; 0: no) fathers who died younger or same age relative to the median life expectancy in sample
		2,434	0.521	0.500	0	1	High longevity (1: yes; 0: no) fathers who died older than the median life expectancy
		2,434	0.018	0.133	0	1	Non-response (1: yes; 0: no)
		2,434	0.127	0.333	0	1	Alive (1: yes; 0:no)
		2,434	0.355	0.478	0	1	Short longevity (1: yes; 0: no) mothers who died younger or same age relative to the median life expectancy
		2,434	0.500	0.500	0	1	High longevity (1: yes; 0: no) mothers who died older than the median life expectancy

		5,612	0.171	0.376	0	1	parents' health condition (1: anyone spent long time in bed; 0: None)
		5,612	0.062	0.241	0	1	Father has drinking problem (1: alcoholism; 0: None)
		5,612	0.099	0.298	0	1	Mother smokes (1: Yes; 0: None)
		5,612	0.444	0.497	0	1	Father smokes (1: Yes; 0: None)
		5,612	0.203	0.403	0	1	Non-response of father (1: yes; 0: no)
		5,612	0.035	0.184	0	1	Alive father (1: yes; 0:no)
	CHN (12)	5,612	0.367	0.482	0	1	Short longevity (1: yes; 0: no) fathers who died younger or same age relative to the median life expectancy
		5,612	0.394	0.489	0	1	High longevity (1: yes; 0: no) fathers who died older than the median life expectancy
		5,612	0.174	0.379	0	1	Non-response of mother (1: yes; 0: no)
		5,612	0.095	0.293	0	1	Alive mother (1: yes; 0:no)
		5,612	0.389	0.488	0	1	Short longevity (1: yes; 0: no) mothers who died younger or same age relative to the median life expectancy
		5,612	0.177	0.382	0	1	High longevity (1: yes; 0: no) mothers who died older than the median life expectancy
Health and nutrition conditions in Childhood	US (5)	2,434	1.685	0.941	1	5	Would you say that your health during that time was (1: excellent, 2: very good, 3: good, 4: fair, 5: poor
		2,434	0.040	0.196	0	1	Before you were 16 years old, were you ever disabled for six months or more because of a health problem? That is, were you unable to do the usual activities of classmates or other children your age?
		2,434	0.104	0.305	0	1	Before you were 16 years old, did you have a blow to the head, a head injury or head trauma that was severe enough to require medical

							attention, to cause loss of consciousness or memory loss for a period of time?
		2,434	2.583	0.895	1	5	when you were 10 how well did you do in math compared to other children in your class (1: much better, 2: better, 3: about the same, 4: worse, 5: much worse)
		2,434	2.400	0.928	1	5	when you were 10 how well did you do in reading and writing compared to other children in your class? (1: much better, 2: better, 3: about the same, 4: worse, 5: much worse)
	CHN (5)	5,612	2.684	0.995	1	5	Self-rated health status before age 15 (1: much healthier; 2: somewhat healthier; 3: about average; 4: some less healthy; 5: much less healthy)
		5,612	1.071	0.733	0	2	Did you ever experience hunger (0: No; 1:yes after age 5; 2: yes before age 5)
		5,612	0.787	0.410	0	1	Have you received any vaccinations before 15 years old?(1: Yes; 0: No)
		5,612	0.275	0.446	0	1	The type of doctor you visited for the first time was in general hospital specialized hospital or township health clinics? (1:Yes; 0: No)
		5,612	0.274	0.446	0	1	The type of doctor you visited for the first time was in community (or village) health centers or private clinics? (1:Yes; 0: No)
Relationship with parents	US (5)	2,434	0.064	0.244	0	1	Before you were 18 years old, were you ever physically abused by either of your parents? 0 also for missing data
		2,434	0.131	0.337	0	1	before age 16 did you ever seperated from you mother for 6 months or longer?
		2,434	0.239	0.427	0	1	before age 16 did you ever seperated from you father for 6 months or longer?

		2,434	0.072	0.258	0	1	were your grandparents ever your primary caregiver?
	CHN (3)	5,612	2.435	1.164	1	5	Relationship with parents (1: excellent; 2: very good; 3: good; 4:fair; 5: poor)
		5,612	0.141	0.348	0	1	Did Male Dependents ever beat you (1: often or somewhat; 0: rarely or never)
		5,612	0.218	0.413	0	1	Did Female Dependents ever beat you (1: often or somewhat; 0: rarely or never)
Friendship in childhood	US (2)	2,434	0.141	0.348	0	1	Before you were 18 years old, did you have to do a year of school over again?
		2,434	0.055	0.228	0	1	Before you were 18 years old, were you ever in trouble with the police?
	CHN (2)	5,612	0.878	0.081	0	1	The average value of neighbors willing to help others at community level, the answers at individual level is 1: very or somewhat, 0: not at all
		5,612	0.438	0.496	0	1	Did you have a good friend (1: yes; 0: no)

Appendix B: Conceptual and Analytic Framework

1 Conventional Parametric Roemer Method

The IOP analysis enables us to identify the individual and collective contributions of childhood environments and their domains to health inequality in later life, which lays the foundation of evaluating policy interventions in childhood (Andreoli et al., 2019). We begin with a simple illustrating example. Suppose we have two binary childhood circumstances in total, i.e. parental education (high/low) and financial hardship (no/yes). Therefore, there are four types, i.e. (high, no), (high, yes), (low, no), (low, yes). All individuals are partitioned into these four groups. Let us assume for now all individuals within each type have the same health status in old age, which means that people with the same values of childhood circumstances have the same health. The variation across the four types in health can only be due to differences in childhood circumstances. This variation as a proportion of the overall health variation across all individuals is the definition of IOP, i.e. the proportion of health inequality that can be explained by observable childhood circumstances.

More generally, existing studies often adopt the following linear parametric model

$$Y_i = \alpha C_i + \varepsilon_i \quad (1)$$

where C is a vector of childhood circumstances beyond the control of the individual, Y is a vector of health outcomes in old age, and i represents individual i. In practice, we do not observe the full set of circumstances C. Instead we only observe a subset $\check{C} \subseteq C$ from which we further choose a subset $\hat{C} \subseteq \check{C} \subseteq C$. Furthermore, we have to consider limited degrees of freedom and choose P circumstances $C^p \in \hat{C}$. Each circumstance C^p is characterized by a total of X^p possible realizations, where each realization is denoted as x^p. Based on the realization x^p we can partition the population into a set of non-overlapping groups (i.e. types), $G = \{g_1,\ldots,g_m,\ldots,g_M\}$, where each group g_m is homogeneous in the expression of each input variable.

We obtain the counterfactual distribution of Y by estimating equation (1). The counterfactual distribution can be constructed from the predicted values of equation (1). IOP is then computed using a common inequality measure $I(.)$. Following Ferreira and Gignoux (2011), we measure the extent to which childhood circumstances contribute to health inequality $I(.)$ by Gini coefficient. We also divide this absolute inequality measure by the same metric applied to the actual outcome to obtain IOP, i.e. the fraction of variation explained by childhood circumstances:

$$\theta_r = \frac{I(\hat{Y})}{I(Y)} \quad (2)$$

To estimate the relative importance of each childhood circumstance, we decompose IOP based on the idea of the Shapley value. To compute the Shapley value decomposition, we first estimate the inequality measure for all possible permutations of the circumstance variables. In a second step, we compute the average marginal effect of each circumstance variable on the measure of IOP (Juarez and Soloaga, 2014). This decomposition method is order independent, meaning that the order of circumstances for decomposition does not affect results and that components of contributions can be added up to the total IOP value. Though the decomposition should not be seen as causal, it offers an idea of the relative importance of circumstances (Ferreira and Gignoux 2013).

2 Conditional inference trees

Conditional inference trees offer a particularly relevant structure in the context of IOP. Sequential hypothesis tests in tree-based methods can segment the population into types. Moreover, its simple graphical illustration is particularly instructive for comparisons of childhood circumstances structures. Each hypothesis test is essentially a test for whether equal childhood circumstances exist within a particular subsample. If the algorithm results in no splits at all, then we cannot reject the null hypothesis of equality of childhood circumstances. The deeper the tree grows, the more types are necessary to fully account for the inherent IOP in the society under consideration. Each split tells us that the resulting types have significantly different childhood circumstances under an ex-ante interpretation. In all of the resulting types (terminal nodes), we cannot reject the null of equal childhood circumstances.

In addition, tree-based methods address concerns over arbitrary circumstances variable selection and model selection that plague the IOP literature. Conventional estimation approaches often leave researchers to select circumstances C^P, restrict the number of realizations of each circumstance, and determine relevant interactions among these circumstances. However, considering all possible ways in which the population can be split into groups is a daunting task when applying the Reomer's theory if the set of input variables is large. The size of this choice set oftentimes leads to arbitrary model selection. Compared to arbitrarily selecting C^P from all observed childhood circumstances $\check{C}$ in the conventional regression-based modeling, we retain the full and unrestricted set of observed variables that may qualify as childhood circumstances to trees.

Specifically, we use the circumstances set $\hat{C}$ to partition the population into a set of non-overlapping groups, $G = \{g_1, \ldots, g_m, \ldots, g_M\}$, which are also called terminal nodes in the regression tree context. Then we calculate the predicted value for outcome *y* of observation *i,* which is the mean outcome μ_m of the group g_m to which the individual is assigned. *N* is the number of observations in *m* group.

$$\hat{y}_i = \mu_m = \frac{1}{N_m} \sum_{i \in g_m} y_i \,, \forall i \in g_m, \forall g_m \in G \tag{2}$$

The standard algorithm illustrated below chooses most relevant circumstances, their subpartition and the respective interactions. First, we use four steps to grow the conditional inference tree with a set of circumstances in childhood:

1. Test the null hypothesis of independence, $H_0^{C^p}: D(Y|C^p) = D(Y)$,[1] for each input variable$C^p \in \hat{C}$, and obtain a *p*-value associated with each test, p^{C^p}.
    - Adjust the p-values for multiple hypothesis testing, such that $p_{adj.}^{C^p} = 1 - (1 - p^{C^p})^P$ (Bonferroni Correction).
2. Select the variable, C^*, with the lowest adjusted *p*-value, i.e. $C^* = \{C^p: argmin\ p_{adj.}^{C^P}\}$.
    - If $p_{adj.}^{C^p} > \alpha$: Exit the algorithm.
    - If $p_{adj.}^{C^p} < \alpha$: Continue, and select C^* as the splitting variable.
3. Test the discrepancy between the subsamples for each possible binary partition, *s*, based on C^*, i.e. $Y_s = \{Y_i: C_i^* < x^p\}$ and $Y_{-s} = \{Y_i: C_i^* \geq x^p\}$, and obtain a *p*-value associated with each test, $p^{C_S^*}$.
    - Split the sample based on $C_{s^*}^*$, by choosing the split point *s* that yields the lowest *p*-value, i.e. $C_{s^*}^* = \{C_s^*: argmin\ p^{C_S^*}\}$
4. Repeat the algorithm for each of the resulting subsamples.

α is defined by using K-fold cross-validation to tune model parameters that performs optimally according to a pre-specified testing criterion[2]. It starts by splitting the sample into *K* subsamples (folds). Then, one implements the conditional inference algorithm on the union of *K*-1 folds for varying levels of α, while leaving out the *k*th as test set to calculate mean squared error (MSE) of prediction.

[1] D donates the distribution of *Y*.
[2] K=5 in this paper. Our results are robust to the choice of K.

$$\mathrm{MSE}_k^{CV}(\alpha) = \sum_m \frac{N_m^k}{N^k} \sum_{i \in t_m} \frac{1}{N_m^k} (y_i^k - \mu_m(\alpha))^2 \quad (3)$$

Repeat the prediction for all *K* folds, and calculate $MSE^{CV}(\alpha) = \frac{1}{K}\sum_k MSE_k^{CV}(\alpha)$. Then we can choose the α∗ that delivers the lowest $MSE^{CV}(\alpha)$:

$$\alpha^* = \{\alpha \in A: argmin\ MSE^{CV}(\alpha)\} \quad (4)$$

3 Conditional inference forest

While conditional inference trees provide an easily mapped and non-arbitrary way to divide population into types, they are some limitations. First, trees only use limited information inherent in the set of observed circumstances, since not all $\check{C}$ are used for construction of each tree. However, omitted variables may possess rich information that can increase predictive power even if they are not significant at level α∗. This is particularly an issue if key circumstances are highly correlated. Once a split is made using either of them, the others will unlikely yield enough information to cause another split. Second, the predictions of trees have high variance. The structure of trees is sensitive to alternations in the respective samples, an issue if there are various circumstances that are close competitors in defining the first split (Friedman et al., 2001).

Random forest improves over trees via decorrelating the trees, the average of the resulting trees has lower variance of the predicted outcomes and hence is more reliable. We grow a large number of decision trees to form a forest on bootstrapped training samples. Each time a split in a tree is considered when growing these decision trees. A random sample of $\overline{\mathrm{P}}$ predictors is chosen as split candidates from the full set of *P* predictors, $\check{C}$. At each split the algorithm uses only one of those $\overline{\mathrm{P}}$ predictors.

This paper creates B number of trees and count all trees by weight in the prediction of $\hat{y}$. To reduce computational cost, we fix B^* at 200 at which the marginal gain of drawing an additional subsample in terms of out-of-sample prediction accuracy becomes negligible (Figure 9). In each tree, we randomly select half of the observations[3]. Trees are constructed according to the same 4-step procedure outlined in the previous subsection. Each tree is estimated on a random subsample *b* of the original data. A random subset of circumstances $\overline{\mathrm{P}}$

[3] Conventionally, researchers bootstrap to select sample for each tree in random forest. However, it has been shown that the bootstrapping can lead to biased variable selection (Strobl et al., 2007).

is used at each splitting point. Then we determine $\alpha^*$and $\bar{P}^*$by minimizing the out-of-bag error.

1. Run a random forest with *B* subsamples, where $\bar{P}$ circumstances are randomly chosen to be considered at each splitting point, and α is used as the cut-off p value for the hypothesis tests.
2. Calculate the average predicted value of observation *i* using each of the subsamples b_{-i} (so called *bags*) in which *i* does not enter: $\hat{y}_i^{OOB}(\alpha, \bar{P}) = \frac{1}{B_{-i}} \sum_{b_{-i}} \mu_m^b(\alpha, \bar{P})$.
3. Calculate the out-of-bag mean squared error: $MSE^{OOB}(\alpha, m) = \frac{1}{N} \sum_i [y_i - \hat{y}_i^{OOB}(\alpha, \bar{P})]^2$.
4. Choose $(\alpha^*, \bar{P}^*) = \{(\{\alpha \in A\}, \{\bar{P} \in \check{P}\}) : argmin\, MSE^{OOB}\}$.

The prediction of y is averaging over the B predictions, which cushions the variance of individual predictions μ_m.

$$\hat{y}_i(\alpha, \bar{P}, B) = \frac{1}{B} \sum_{b=1}^{B} \mu_m^b(\alpha, \bar{P}) \quad (5)$$

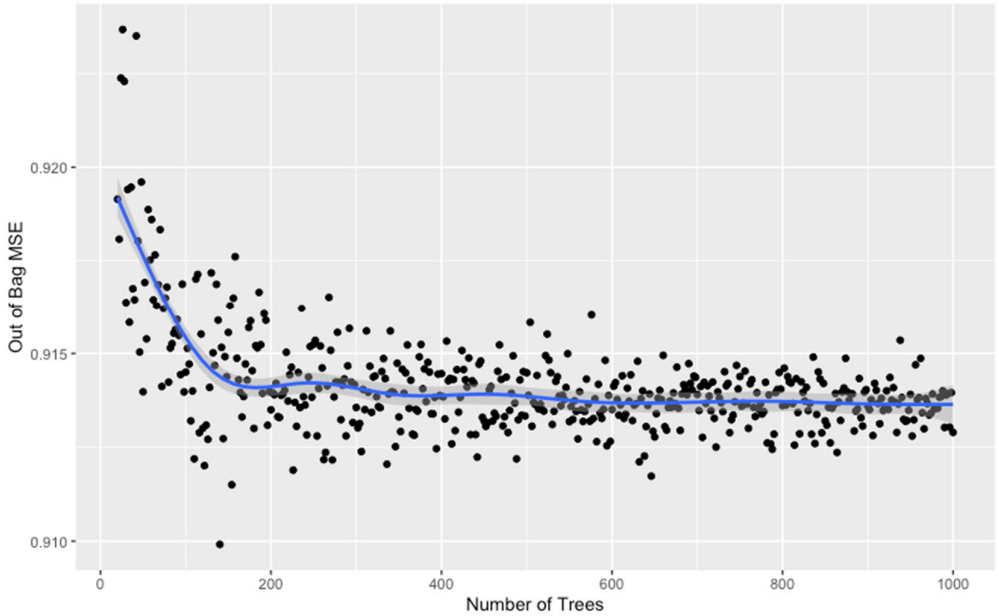
Although the collection of bagged trees is much more difficult to interpret than a single tree, we can obtain an overall summary of the importance of each predictor using the residual sum of squares (RSS).

4 Out-of-Sample Performance Test

To assess potentials of both downward and upward biases of IOP in health that may affect out-of-sample performance, we follow the standard practice to split sample into a training set (2/3*N) and a test set (1/3*N). We fit our model on the training set and compare the performance on the test set for the conventional parametric Roemer method, conditional inference trees, and conditional inference forest, respectively. Specifically, we follow the same procedure:

1) Run the chosen models on the training data.
2) Store the prediction functions $\hat{f}_{train}(\check{C})$.
3) Predict the outcomes of observations in the test set: $\hat{y}_{i_{test}} = \hat{f}_{train}(\check{C}_{i_{test}})$.
4) Calculate the out-of-sample error: $\text{MSE}^{test} = \frac{1}{N_{test}} \sum_{i_{test}} [y_{i_{test}} - \hat{y}_{i_{test}}]^2$.

Appendix C: Optimal Number of Trees



Note: The x-axis shows the parameter value for B, i.e. the number of trees per forest. The dots show the MSE^{OOB} obtained from estimating a random forest with the given number of trees for the self-rated health in the US. We allow for 7 circumstances to be considered at each splitting point. The blue line is a non-parametric fitted line of the MSEOOB estimates and the shaded area the 95% confidence interval of this line. Evidently, as the tree size approaches 200, on expectation, the MSEOOB stops improving much.