

UC Davis

UC Davis Previously Published Works

Title

Applying Machine Learning Approach to Explore Childhood Circumstances and Self-Rated Health in Old Age - China and the US, 2020-2021.

Permalink

<https://escholarship.org/uc/item/98j703xc>

Journal

China CDC Weekly, 6(11)

Authors

Huo, Shutong

Feng, Derek

Gill, Thomas

et al.

Publication Date

2024-03-15

DOI

10.46234/ccdcw2024.043

Peer reviewed

Methods and Applications

Applying Machine Learning Approach to Explore Childhood Circumstances and Self-Rated Health in Old Age — China and the US, 2020–2021

Shutong Huo¹; Derek Feng²; Thomas M. Gill³; Xi Chen^{4,5,#}

ABSTRACT

Introduction: Childhood circumstances impact senior health, prompting the introduction of machine learning methods to assess their individual and collective contributions to senior health.

Methods: Using health and retirement study (HRS) and China Health and Retirement Longitudinal Study (CHARLS), we analyzed 2,434 American and 5,612 Chinese participants aged 60 and above. Conditional inference trees and forests were employed to estimate the influence of childhood circumstances on self-rated health (SRH).

Results: The conventional method estimated higher inequality of opportunity (IOP) values in both China (0.039, accounting for 22.67% of the total Gini coefficient 0.172) and the US (0.067, accounting for 35.08% of the total Gini coefficient 0.191). In contrast, the conditional inference tree yielded lower estimates (China: 0.022, accounting for 12.79% of 0.172; US: 0.044, accounting for 23.04% of 0.191), as did the forest (China: 0.035, accounting for 20.35% of 0.172; US: 0.054, accounting for 28.27% of 0.191). Childhood health, financial status, and regional differences were key determinants of senior health. The conditional inference forest consistently outperformed others in predictive accuracy, as demonstrated by lower out-of-sample mean squared error (MSE).

Discussion: The findings emphasize the need for early-life interventions to promote health equity in aging populations. Machine learning showcases the potential in identifying contributing factors.

INTRODUCTION

The global phenomenon of rapid population aging, coupled with the growing health burden among older adults, highlights the importance of investigating the long-term effects of early life stages on the aging process (1). Previous research in the fields of economics

and epidemiology has consistently shown that childhood circumstances have a significant impact on later-life health outcomes. This suggests that childhood is a crucial period for implementing interventions aimed at reducing health disparities (2). These circumstances encompass a wide range of factors, including parental influences (3), family socioeconomic status (SES) (4), as well as community and environmental factors such as rural/urban status (5) and natural surroundings (6).

Both early-life and later-life factors contribute to health outcomes in older age. However, childhood circumstances, particularly those that are beyond an individual's control, are considered to be the most unacceptable and illegitimate sources of health inequality in older age (7–8). This type of inequality, attributed to childhood circumstances, is commonly referred to as inequality of opportunity (IOP). The focus on reducing IOP arises from a wide-ranging political and social discussion aimed at creating equal opportunities during the early stages of life and addressing the unfair health inequalities identified by the World Health Organization Commission on Social Determinants of Health (9).

Despite the considerable amount of research conducted on the impact of childhood circumstances on health outcomes, there are still methodological challenges that need to be addressed. These challenges include the arbitrary selection of childhood circumstances and potential biases in estimating health inequality among older adults (10–11). In our study, we aimed to overcome these challenges by utilizing machine learning techniques to identify the most relevant set of childhood circumstances. By adopting this approach, we allowed the data to inform our understanding of unequal childhood circumstances, thus minimizing the influence of researcher bias on the model specification (10–12). Furthermore, we compared our findings to those obtained using the conventional parametric Roemer method in order to

highlight the significant improvements our approach offers in measuring inequality throughout an individual's life.

METHODS

Our study utilized data from the health and retirement study (HRS) in the US and the China Health and Retirement Longitudinal Study (CHARLS) in China. We analyzed 2020–2021 wave of HRS and the 2020 wave of CHARLS, both of which matched with life history surveys. The final sample consisted of 2,434 Americans and 5,612 Chinese individuals aged 60 and above. Self-rated health (SRH) was used as the health outcome measure, assessed on a scale from excellent (=1) to poor (=5) in both surveys. The analysis included data on 43 childhood circumstances from HRS and 36 from CHARLS, categorized into seven domains such as birth environment, family SES, and childhood relationships (Supplementary Tables S1 and S2, available at <https://weekly.chinacdc.cn/>). While there were slight variations, the domains predominantly included the same core measures for both countries. The analysis was conducted using R (version 4.3.1; R Core Team, Vienna, Austria).

Supplementary Material (available at <https://weekly.chinacdc.cn/>) provides a comprehensive conceptual and analytic framework for this study. Initially, we used the Roemer method with Shapley value decomposition to estimate the individual and collective impact of childhood circumstances on health inequality in later life. This framework serves as a foundation for evaluating policy interventions. By partitioning the population into distinct, non-overlapping groups based on observable circumstances, such as parental education (high *vs.* low) and financial hardship (yes *vs.* no), we can derive a counterfactual distribution of health outcomes. The disparity in health across these groups can be solely attributed to differences in childhood circumstances, which we refer to as the IOP. In our study, we quantified the contribution of childhood circumstances to health inequality using the Gini coefficient (8,11). We also calculated the IOP by dividing this measure of absolute health inequality by the overall health inequality, representing the proportion of health inequality explained by childhood circumstances. While not establishing causality, this analysis provides valuable insights into the statistical significance of childhood circumstances (13).

Conditional inference trees are particularly advantageous for analyzing the impact of childhood circumstances on IOP. They allow for sequential hypothesis tests and provide a visual representation for comparing different childhood circumstances. Each test examines IOP within a specific subset of the population, and the depth of the tree reflects the diversity of childhood circumstances within a society. Additionally, these trees address the issue of arbitrary variable and model selection that often arises in the IOP literature. They consider a comprehensive set of observed variables that qualify as childhood circumstances. In our study, we used these childhood circumstances to divide the population into distinct groups (terminal nodes) in the context of regression trees. We calculated the predicted outcome value for an individual observation as the average outcome of the group to which the individual was assigned, taking into account the number of observations in that group. Furthermore, we used 5-fold cross-validation to optimize the model parameters. We found that our results are consistent regardless of the choice of K .

Conditional inference trees have advantages in providing non-arbitrary population segmentation. However, they have limitations such as using limited data, struggling with highly correlated childhood circumstances, and exhibiting high prediction variance, making them sensitive to sample changes. To address these limitations, random forest is employed to mitigate these issues. Random forest forms a forest of decision trees from bootstrapped samples, utilizing a random selection of predictors at each split to reduce prediction variance, resulting in a more reliable model. In this study, 200 trees were used based on considerations of computational cost-efficiency and prediction accuracy to predict outcomes (Supplementary Figure S1, available at <https://weekly.chinacdc.cn/>). A 4-step method was applied, involving the random selection of half the observations in each tree, along with random data subsampling and subsets of circumstances, to determine optimal parameters through out-of-bag error minimization. Predictor importance for each childhood circumstance was evaluated using the residual sum of squares (RSS).

To evaluate the potential biases in measuring IOP in healthy individuals that could impact the accuracy of predictions, we divided the dataset into a training set representing 2/3 of the total sample size (N) and a test set representing the remaining 1/3. The training set was used to train our model, while the test set was used to assess the performance of three different methods:

the conventional parametric Roemer method, conditional inference trees, and conditional inference forest.

RESULTS

First, the Gini coefficient indicated that there was a higher level of inequality in self-rated health in the US compared to China. We then used the Gini coefficients to measure the IOP in the counterfactual distribution. Figure 1 illustrates that the conventional parametric Roemer method yielded the highest estimates of IOP, followed by the conditional inference forest method and the conditional inference tree method. Specifically, in China, IOP accounted for 22.67% (0.039 out of 0.172 total Gini coefficient) of the inequality in self-rated health, while in the US it accounted for 35.08% (0.067 out of 0.191 total Gini coefficient). In contrast, the conditional inference tree method accounted for 12.79% in China (0.022 out of 0.172 total Gini coefficient) and 23.04% in the US (0.044 out of 0.191 total Gini coefficient), while the forest method represented 20.35% in China (0.035

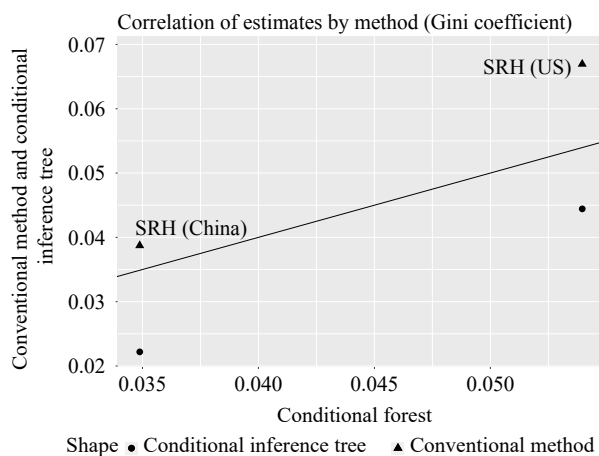


FIGURE 1. Correlation of estimates by method.

Note: The plot shows the estimates using each method (i.e., the conventional parametric Roemer method and the conditional inference trees) against the estimates from conditional inference forest. The x-axis represents the scale of Gini coefficients for the forest method. The Gini coefficients range between 0 and 1. The larger the more unequal. The y-axis represents the scale of Gini coefficients for the Roemer method and tree methods. The black diagonal indicates the 45-degree line, on which all data points should align if the different methods were perfectly congruent. This plot confirms that the conventional parametric Roemer method delivers higher estimates than forest, while tree estimates are lower than those based on forest.

Abbreviation: SRH=self-rated health.

out of 0.172 total Gini coefficient) and 28.27% in the US (0.054 out of 0.191 total Gini coefficient).

Figure 2A shows the structure of the IOP for self-rated health in China using a tree with five terminal nodes. The tree is formed by factors such as childhood health, birth region, and childhood family financial status. The most advantaged type (terminal node 5) includes people with good childhood health, good family financial status, and born in Eastern China. On the other hand, the group with the worst self-rated health (terminal node 6) typically had poorer child health. In the US, as depicted in Figure 2B, individuals with poor childhood health fell into the disadvantaged circumstance type (terminal nodes 7). In contrast, individuals with 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 reveals that in China, using conditional inference forest, the key factors impacting self-rated health are childhood health and being born in the eastern China, which corroborates findings from the conditional inference trees (Figure 2A). Additionally, parents' health status (staying in bed for a long time) and relationship with parents also have a high impact on self-rated health in older ages. Similarly, Figure 3B demonstrates that in the US, childhood health, number of books at home at age 10, and race/ethnicity are significant factors, which largely align with results obtained through conditional inference trees (Figure 2B).

As previously mentioned, all tested models were designed to minimize the mean squared error (MSE). We derived 95% confidence intervals using 200 bootstrap re-samplings of the test data. The MSE for the random forest model was standardized to a value of 1 to facilitate comparison of prediction performance across models. Therefore, an MSE greater than 1 indicated a poorer out-of-sample fit. In terms of self-rated health, both the conditional inference tree and parametric Roemer methods performed worse than the conditional inference forest, as shown in Figure 4A–B. On average, the conditional inference trees demonstrated lower test error rates compared to the conventional parametric Roemer method.

DISCUSSION

This study utilized two machine learning methods, namely the conditional inference tree and forest, to investigate the effects of various childhood

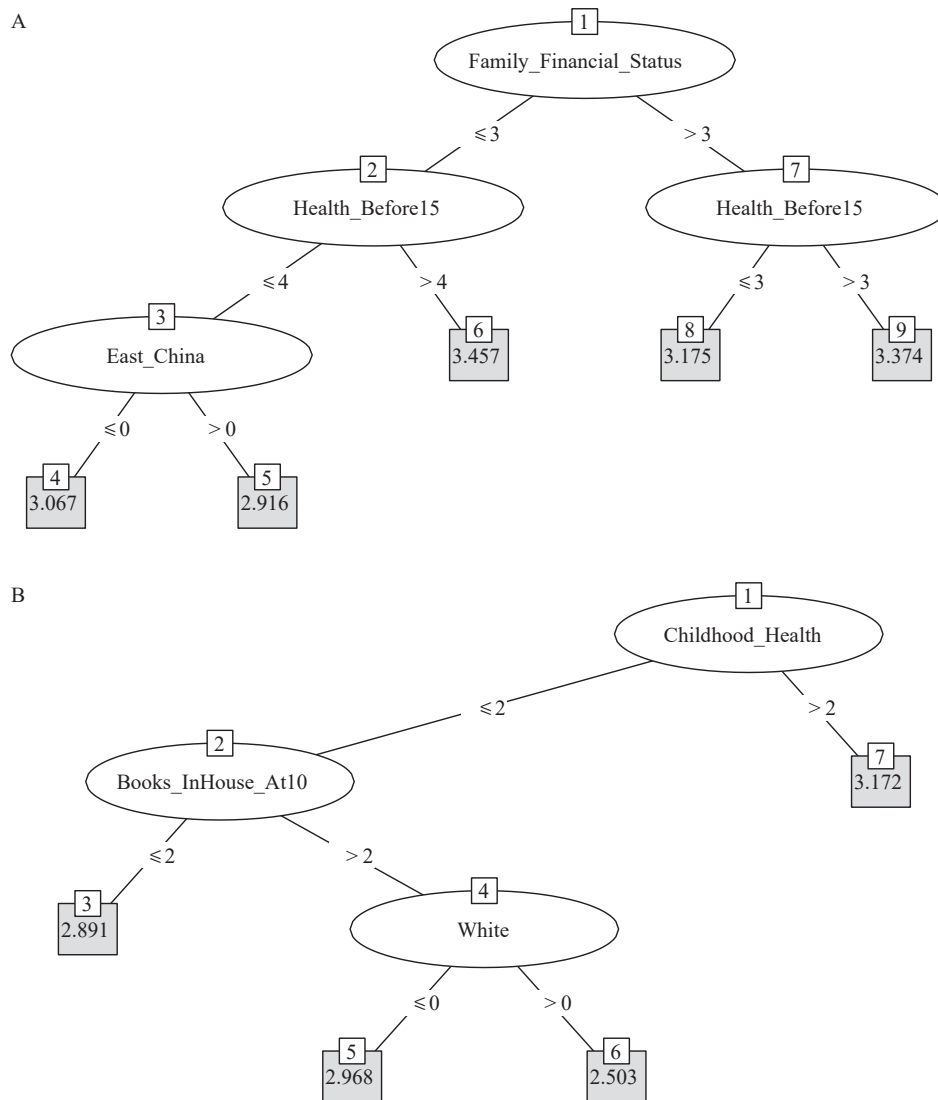


FIGURE 2. Conditional inference tree for self-rated health. (A) China; (B) the US.

circumstances on health disparity among older adults in China and the US. We identified several key predictors of health conditions in older adults, including childhood health, socioeconomic status, number of books at home (in the case of Americans), and birth region (in the case of Chinese). By employing these methods, we aimed to address concerns regarding the arbitrary selection of childhood circumstances and mitigate potential biases in our estimates of the impact of childhood circumstances on health. Our findings emphasize the importance of mitigating health disparities stemming from childhood circumstances, and suggest the need for policy and intervention strategies to promote health equity in both China and the US. Implementing preventive measures during childhood can alleviate the economic burden of diseases, enhance quality of life, and improve

longevity, particularly in the absence of effective treatments for chronic diseases like Alzheimer's, hypertension, and diabetes.

The conditional inference forest (CIF) demonstrates superior out-of-sample performance compared to other methods, resulting in the most accurate estimates of childhood circumstances on health inequality in old age. This finding is in line with previous studies in various fields (14–15). While conditional inference trees provide a simpler model and a visually accessible representation of childhood circumstances, the CIF leverages information on childhood circumstances more effectively, yielding results consistent with the trees in terms of importance and estimates of influence on health outcomes. These machine learning methods employ explicit algorithms to interpret health outcomes and do not rely on strong assumptions

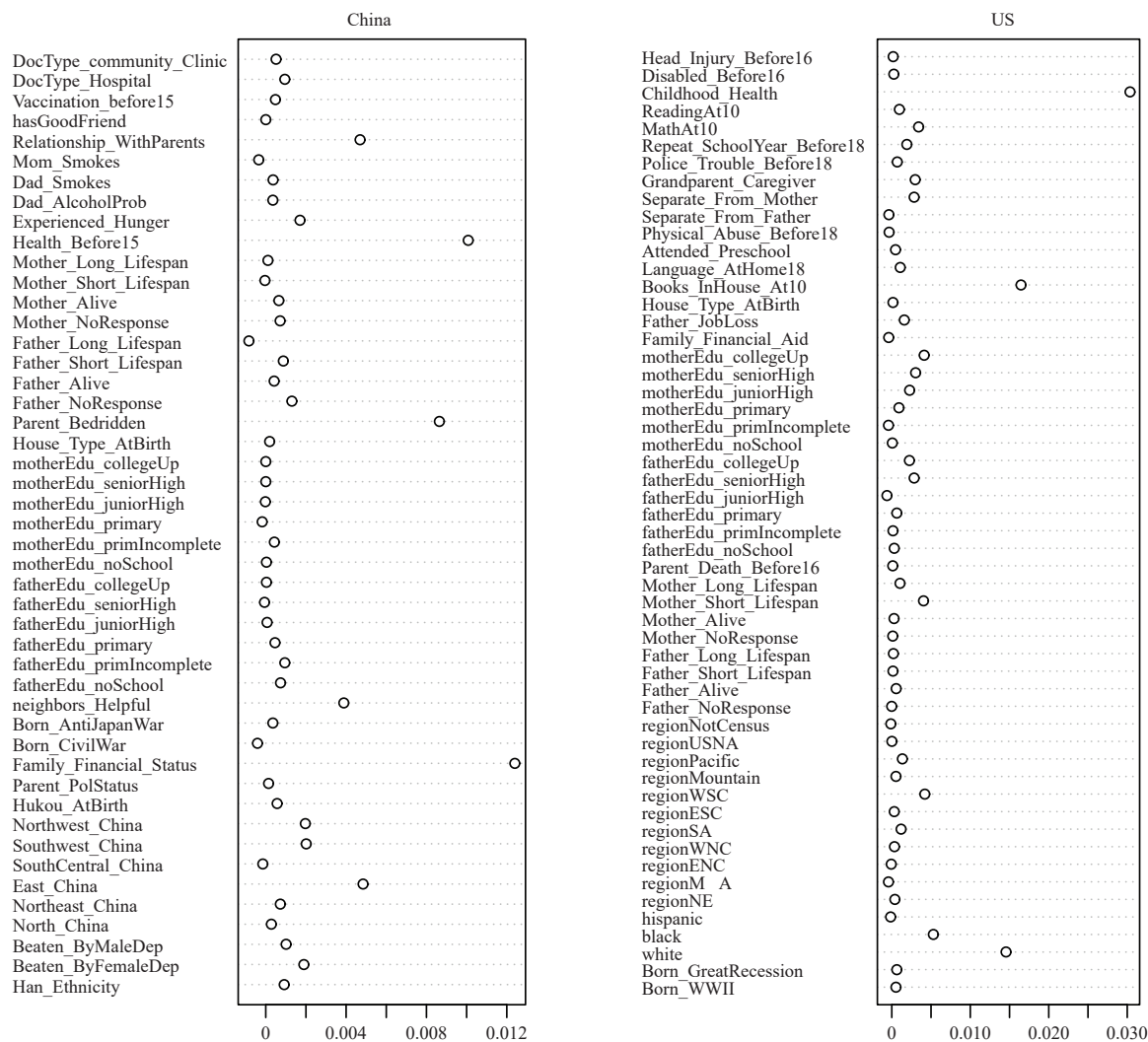


FIGURE 3. Importance of childhood circumstances to self-rated health using conditional inference forest. (A) China; (B) the US.

regarding the significance of specific childhood circumstances. By utilizing statistical techniques such as K-fold cross-validation and bootstrap, our modeling approach becomes more transparent and generalizable.

There are several limitations to this study. First, the life course approach used in this study only focuses on current older adults, which may not accurately reflect the experiences of younger cohorts. Therefore, future research should also consider monitoring younger cohorts. Second, it is important to note that the associations identified in this study should not be interpreted as causal. It is possible that unobservable childhood circumstances may introduce bias to our estimates. Therefore, further research is needed to identify the causal mechanisms at play. Lastly, the data used in this analysis are from the most recently released CHARLS (2020) and HRS (2020–2021) surveys,

which overlap with the coronavirus disease 2019 (COVID-19) pandemic. This may introduce bias to self-rated health measures. However, our robustness checks using CHARLS/HRS pre-pandemic waves have yielded consistent results, providing reassurance.

In conclusion, our study utilized a life course approach and machine learning techniques to identify key factors influencing health in older adults. We applied this approach to the two largest economies and aging societies in the world. Our findings underscore the importance of incorporating a life course perspective in public health research and policy development.

Conflicts of interest: No conflicts of interest.

Funding: Supported by the U.S. National Institute on Aging (R01AG077529; P30AG021342; R01AG037031).

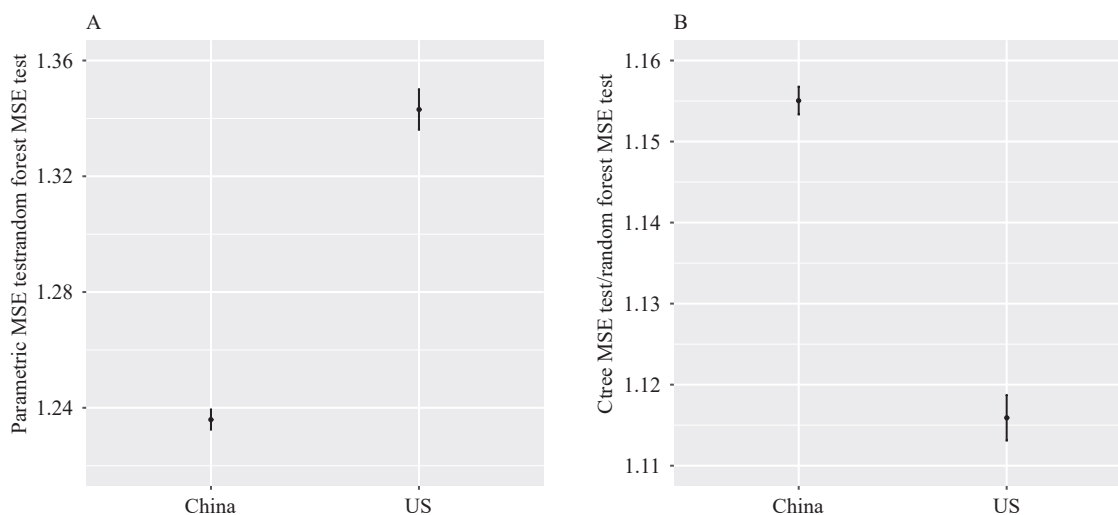


FIGURE 4. Comparison of models' test errors. (A) Parametric method vs. random forest; (B) Conditional inference trees vs. random forest.

Note: All models aim to minimize the 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. The 95% confidence intervals are derived based on 200 bootstrapped re-samples of the test data.

Abbreviation: MSE=mean squared error.

doi: 10.46234/ccdcw2024.043

Corresponding author: Xi Chen, xi.chen@yale.edu.

¹ Department of Health, Society & Behavior, Public Health, University of California, Irvine, CA, USA; ² Department of Statistics and Data Science, Yale University, New Haven, CT, US; ³ Department of Internal Medicine, Yale School of Medicine, New Haven, CT, US; ⁴ Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, US; ⁵ Department of Economics, Yale University, New Haven, CT, US.

Submitted: November 30, 2023; Accepted: January 29, 2024

REFERENCES

- 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;72(2):210 – 5. <https://doi.org/10.1093/gerona/glw191>.
- Bor J, Cohen GH, Galea S. Population health in an era of rising income inequality: USA, 1980-2015. *Lancet* 2017;389(10077):1475 – 90. [https://doi.org/10.1016/S0140-6736\(17\)30571-8](https://doi.org/10.1016/S0140-6736(17)30571-8).
- Carrieri V, Jones AM. Inequality of opportunity in health: a decomposition-based approach. *Health Econ* 2018;27(12):1981 – 95. <https://doi.org/10.1002/hec.3814>.
- Moody-Ayers S, Lindquist K, Sen S, Covinsky KE. Childhood social and economic well-being and health in older age. *Am J Epidemiol* 2007;166(9):1059 – 67. <https://doi.org/10.1093/aje/kwm185>.
- Strauss J, Witoelar F, Meng QQ, Chen XX, Zhao YH, Sikoki B, et al. Cognition and SES relationships among the mid-aged and elderly: a comparison of China and Indonesia. National Bureau of Economic Research; 2018 May Report No.: 24583. <https://www.nber.org/papers/w24583>.
- 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;125(3):848 – 902. <https://doi.org/10.1086/691465>.
- Roemer JE. Equality of opportunity. Cambridge: Harvard University Press. 1998; p. 130.
- Roemer JE, Trannoy A. Equality of opportunity: theory and measurement. *J Econ Lit* 2016;54(4):1288 – 332. <https://doi.org/10.1257/jel.20151206>.
- 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* 2008;372(9650):1661 – 9. [https://doi.org/10.1016/S0140-6736\(08\)61690-6](https://doi.org/10.1016/S0140-6736(08)61690-6).
- Brunori P, Hufe P, Mahler D. The roots of inequality: estimating inequality of opportunity from regression trees and forests. *Scand J Econ* 2023;125(4):900 – 32. <https://doi.org/10.1111/sjoe.12530>.
- 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. <https://doi.org/10.1111/j.1475-4991.2011.00467.x>.
- Hufe P, Peichl A, Roemer J, Ungerer M. Inequality of income acquisition: the role of childhood circumstances. *Soc Choice Welf* 2017;49(3):499 – 544. <https://doi.org/10.1007/s00355-017-1044-x>.
- Ferreira FHG, Gignoux J. The measurement of educational inequality: achievement and opportunity. *World Bank Econ Rev* 2014;28(2):210 – 46. <https://doi.org/10.1093/wber/lht004>.
- Qi YJ. Random forest for bioinformatics. In: Zhang C, Ma YQ, editors. Ensemble machine learning: methods and applications. New York: Springer. 2012; p. 307-23. http://dx.doi.org/10.1007/978-1-4419-9326-7_11.
- Schneider J, Hapfelmeier A, Thöres S, Obermeier A, Schulz C, Pflörringer 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;16(1):15. <https://doi.org/10.1186/s12876-016-0428-1>.

SUPPLEMENTARY MATERIAL

CONCEPTUAL AND ANALYTIC FRAMEWORK

Conventional Parametric Roemer Method

The analysis of individual and collective contributions of childhood environments to health inequality in later life can be conducted using the inequality of opportunity (IOP) method. This method allows us to evaluate policy interventions in childhood by identifying the specific impact of different childhood circumstances (Andreoli et al., 2019). To illustrate this, let's consider a simple example with two binary childhood circumstances: parental education (high/low) and financial hardship (no/yes). These circumstances create four distinct groups: (high education, no hardship), (high education, hardship), (low education, no hardship), and (low education, hardship). All individuals are grouped into these four categories. For the sake of simplicity, let's assume that individuals within each group have the same health status in old age. Therefore, any variation in health across the four groups can be attributed solely to differences in childhood circumstances. This variation, as a proportion of the overall health variation among all individuals, defines the IOP. In other words, the IOP represents the proportion of health inequality that can be explained by observable childhood circumstances

In general, existing studies frequently utilize 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 $\tilde{C} \subseteq C$ from which we further choose a subset $\tilde{C} \subseteq \tilde{C} \subseteq C$. Furthermore, we have to consider limited degrees of freedom and choose P circumstances $C^p \in \tilde{C}$. 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), $G = \{g_1, \dots, g_m, \dots, g_M\}$, where each group g_m is homogeneous in the expression of each input variable.

We estimated Equation 1 to obtain the counterfactual distribution of Y . The predicted values from Equation 1 were used to construct the counterfactual distribution. IOP was computed using a common inequality measure $I(\cdot)$. Following the approach of Ferreira and Gignoux (2011), we used the Gini coefficient to measure the contribution of childhood circumstances to health inequality, denoted as $I(\cdot)$. To obtain the fraction of variation explained by childhood circumstances, referred to as IOP, we divided this measure of absolute inequality by the same metric applied to the actual outcome.

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

We utilized the concept of the Shapley value to estimate the relative importance of each childhood circumstance in the decomposition of IOP. This decomposition method allows us to compute the average marginal effect of each circumstance variable on the measure of IOP, regardless of their order. It is worth noting that the order of circumstances for decomposition does not influence the results and that the components of contributions can be summed to obtain the total IOP value. It is important to clarify that while this decomposition provides insight into the relative importance of circumstances, it should not be interpreted as indicating causality (Juarez and Soloaga, 2014; Ferreira and Gignoux, 2013).

SUPPLEMENTARY TABLE S1. Summary statistics of self-rated health in the US and China.

Variable	Country	Obs	Mean	SD	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
	China	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

Abbreviation: Obs=number of observations; SD=standard deviation; CV=coefficient of variation.

Conditional Inference Trees

By performing sequential hypothesis tests, tree-based methods can divide the population into distinct groups. Each hypothesis test determines if equal childhood circumstances exist within a specific subset of the population. If the algorithm does not result in any splits, it suggests that the null hypothesis of equal childhood circumstances cannot be rejected. As the tree becomes more extensive, a greater number of groups are required to fully capture the inherent inequalities in the society of interest. Each split indicates that the resulting groups have significantly different childhood circumstances based on an ex-ante interpretation. It should be noted that within each resulting group (terminal node), the null hypothesis of equal childhood circumstances cannot be rejected.

In addition, tree-based methods provide a solution to the issues of arbitrary variable selection and model selection that are common in the IOP literature. Traditional estimation approaches often require 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 to divide the population into groups becomes overwhelming when there is a large set of input variables, particularly when using Reomer's theory. The sheer number of choices often leads to arbitrary model selection. Compared to arbitrarily selecting C^p from all observed childhood circumstances \tilde{C} in the conventional regression-based modeling, we retain the full and unrestricted set of observed variables that may qualify as childhood circumstances for trees.

Specifically, we use the circumstances set \tilde{C} to partition the population into a set of non-overlapping groups, $G = \{g_1, \dots, g_m, \dots, g_M\}$, 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.

$$\hat{y}_i = \mu_m = \frac{1}{N_m} \sum_{i \in g_m} y_i, \forall i \in g_m, \forall g_m \in G \quad (3)$$

Conditional Inference Forest

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 \bar{p} predictors is chosen as split candidates from the full set of P predictors, \tilde{C} . At each split the algorithm uses only one of those \bar{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 (Supplementary Figure S1). In each tree, we randomly select half of the observations*. 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 \bar{p} is used at each splitting point. Then we determine α^* and \bar{p}^* by minimizing the out-of-bag error.

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

$$\hat{y}_i(\alpha, \bar{P}, B) = \frac{1}{B} \sum_{b=1}^B \mu_m^b(\alpha, \bar{P}) \quad (4)$$

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).

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 ($2/3 \cdot N$) and a test set ($1/3 \cdot N$). We fit our model on the training set and compare the performance on the test set for the conventional parametric

* 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).

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: $MSE^{test} = \frac{1}{N_{test}} \sum_{i_{test}} [y_{i_{test}} - \hat{y}_{i_{test}}]^2$.

SUPPLEMENTARY TABLE S2. 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)
	China (2)	5,612	0.295	0.456	0	1	Born in the War of Against Japanese Aggression 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)
	China (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)		

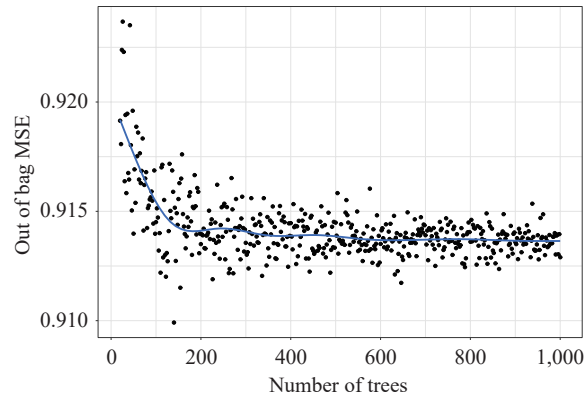
Continued

Domain	Country	Obs	Mean	SD	Min	Max	Variable description		
Family socioeconomic status	US (10)	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 months or longer; 2: yes, never worked/always disabled; 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?		
	China (5)	8,585	0.075	0.263	0	1	Did you attend any organized pre-school programs (1: yes; 0: no)		
		7,795	0.654	0.476	0	1	Parents' political status (1: either father or mother is party member; 0: none of them are)		
		7,795	0.212	0.409	0	1	Father: No schooling (1: yes; 0: no)		
		7,795	0.082	0.276	0	1	Father: educated without completing primary school (1: yes; 0: no)		
		7,795	0.027	0.163	0	1	Father: Graduated from primary school (1: yes; 0: no)		
		7,795	0.027	0.163	0	1	Father: Graduated from junior high school (1: yes; 0: no)		
		7,795	0.015	0.121	0	1	Father: Graduated from senior high school (1: yes; 0: no)		
		7,795	0.009	0.095	0	1	Father: Graduated from college or above (1: yes; 0: no)		
		8,156	0.945	0.228	0	1	Mother: No schooling (1: yes; 0: no)		
		8,156	0.032	0.177	0	1	Mother: educated without completing primary school (1: yes; 0: no)		
		8,156	0.015	0.123	0	1	Mother: Graduated from primary school (1: yes; 0: no)		
		8,156	0.004	0.062	0	1	Mother: Graduated from junior high school (1: yes; 0: no)		
		8,156	0.003	0.053	0	1	Mother: Graduated from senior high school (1: yes; 0: no)		
		8,156	0.001	0.022	0	1	Mother: Graduated from college or above (1: yes; 0: no)		
		8,484	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)		
		8,552	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 life expectancy (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		
China (12)	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 having drinking problems (1: alcoholism; 0: none)		
	5,612		0.099	0.298	0	1	Mother smoking (1: yes; 0: none)		
	5,612		0.444	0.497	0	1	Father smoking (1: yes; 0: none)		

Continued

Domain	Country	Obs	Mean	SD	Min	Max	Variable description
Parents' health status and health behaviors	China (12)	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)
		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)
	China (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	Have 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 separate from your mother for 6 months or longer?
		2,434	0.239	0.427	0	1	Before age 16 did you ever separate from your father for 6 months or longer?
		2,434	0.072	0.258	0	1	Were your grandparents ever your primary caregiver?
	China (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?
	China (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)

Abbreviation: Obs=number of observations; SD=standard deviation.



SUPPLEMENTARY FIGURE S1. Optimal number of trees in conditional random forest.

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 7 circumstances to be considered at each splitting point. The blue line is a non-parametric fitted line of the MSE^{OOB} estimates and the shaded area is the 95% confidence interval of this line. Evidently, as the tree size approaches 200, on expectation, the MSE^{OOB} stops improving much.

Abbreviation: MSE=mean square error; MSE^{OOB} =out-of-bag mean square error.