Title

Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences.

Authors

Cintron, Dakota W
Adler, Nancy E
Gottlieb, Laura M
et al.

Publication Date

2022-06-01

DOI

10.1016/j.annepidem.2022.04.009

Peer reviewed

HHS Public Access

Author manuscript

Ann Epidemiol. Author manuscript; available in PMC 2022 November 26.

Published in final edited form as:

Ann Epidemiol. 2022 June ; 70: 79–88. doi:10.1016/j.annepidem.2022.04.009.

Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences

Dakota W. Cintron[a,b], **Nancy E. Adler**[a], **Laura M. Gottlieb**[a], **Erin Hagan**[a], **May Lynn Tan**[a], **David Vlahov**[c], **Madellena Maria Glymour**[a,b], **Ellicott C. Matthay**[b,d,e,*]

[a]Center for Health and Community, School of Medicine, University of California, San Francisco, CA

[b]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

[c]Yale School of Nursing, Yale University, New Haven, CT

[d]Department of Population Health, New York University Grossman School of Medicine, New York, NY

[e]Center for Opioid Epidemiology and Policy, Division of Epidemiology, Department of Population Health, New York University Grossman School of Medicine, 180 Madison Ave, New York, NY 10016.

Abstract

Purpose: Social policies are important determinants of population health but may have varying effects on subgroups of people. Evaluating heterogeneous treatment effects (HTEs) of social policies is critical to determine how social policies will affect health inequities. Methods for evaluating HTEs are not standardized. Little is known about how often and by what methods HTEs are assessed in social policy and health research.

Methods: A sample of 55 articles from 2019 on the health effects of social policies were evaluated for frequency of reporting HTEs; for what subgroupings HTEs were reported; frequency of *a priori* specification of intent to assess HTEs; and methods used for assessing HTEs.

Results: A total of 24 (44%) studies described some form of HTE assessment, including by age, gender, education, race/ethnicity, and/or geography. Among studies assessing HTEs, 63% specified HTE assessment *a priori*, and most (71%) used descriptive methods such as stratification;

*Corresponding author. Department of Epidemiology and Biostatistics, University of California, San Francisco, 550 16th Street, 2nd Floor, Campus Box 0560, San Francisco, CA, 94143. ellicott.matthay@nyulangone.org (E.C. Matthay).

Author contributions

Dakota W. Cintron: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Writing – Review & Editing. Nancy E. Adler.: Writing – Review & Editing. Laura M. Gottlieb: Writing – Review & Editing. Erin Hagan: Writing – Review & Editing. May Lynn Tan: Writing – Review & Editing. David Vlahov: Writing – Review & Editing. M. Maria Glymour: Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision. *Ellicott C. Matthay*: Conceptualization, Methodology, Writing – Original Draft, Writing – Review & Editing, Supervision.

Conflicts of interest statement: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

21% used statistical tests (e.g., interaction terms in a regression); and no studies used data-driven algorithms.

Conclusions: Although understanding HTEs could enhance policy and practice-based efforts to reduce inequities, it is not routine research practice. Increased evaluation of HTEs across relevant subgroups is needed.

Keywords

Heterogeneous treatment effects; Effect modification; Subgroup analysis; Social policy; Health equity

Introduction

Social policies are appealing as tools for improving population health because they have the potential to affect everyone in a population. However, policies may have larger health effects for some groups and smaller effects for others. Some policies may even have qualitatively different effects across subgroups, benefitting some while harming others [1,2]. Differences in the effect of a policy or intervention for different people are known as heterogeneous treatment effects (HTEs). HTEs are also known by terms such as effect modification, interaction, or subgroup effects [3].

Knowing the HTEs of social policies can help to predict the likely impact of a policy on health disparities: if an intervention disproportionately benefits individuals with the worst health, it will reduce disparities; if an intervention primarily benefits those who are already healthiest, it will exacerbate disparities [4–7]. Understanding heterogeneous effects of policies also can help policy-makers to predict whether results from a social policy in one community will generalize to new settings with different population compositions [8].

Despite the importance of assessing HTEs for translating policy research into population health gains, there is no consistent guidance on best practices for evaluating HTEs in research on social policies and health [9]. Methods for evaluating HTEs vary across disciplines but, generally, they can be assessed based on *a priori* theoretical considerations via stratified analyses, with interaction terms in a regression model, or using data-driven algorithms that are agnostic regarding pre-specification of subgroups [9]. Data-driven or machine-learning algorithms (e.g., recursive partitioning to identify unique subgroups for whom the magnitude of the policy effect is different [10] or Bayesian modeling averaging for subgroup selection [11]) may allow users to identify complex, novel interactions that were not anticipated beforehand and routinely involve cross-validation to reduce chance findings. [9, 12] The choice of method to assess HTEs constrains the inferences that can be drawn. For example, interaction terms directly quantify the heterogeneity in effects across groups and the interaction *P*-value indicates whether this heterogeneity is within a range expected by chance. Stratified estimates do not support such inferences without additional calculations.

We distinguish between two main types of HTE research: (1) research evaluating whether there is *any* heterogeneity in response to the treatment, across as-yet unidentified

characteristics, and (2) research evaluating whether a prespecified characteristic (e.g., age, race/ethnicity, education) defines groups who on average respond differently to treatment. Analyses involving stratification and interaction terms fall in the latter category, whereas many data-driven algorithms fall in the former. Table 1 summarizes the main distinctions between these methods.

Several studies have systematically reviewed evaluation of HTEs for non-social policy determinants of health, such as medical interventions [16–21]. For instance, Starks et al. [20] evaluated the prevalence of HTE analyses and statistical methods used in cluster-randomized trials focused on treating cardiovascular disease, cancer, or chronic lower respiratory disease. Only 28.1% of the trials assessed HTEs and when HTE analyses were performed, HTE assessment by demographic subgroups was uncommon (6.3%), despite National Institute of Health and Food and Drug Administration policies requiring that investigators examine demographic HTEs for trials randomized at the individual level [20,22,23]. To our knowledge, no prior studies have reviewed either the frequency of HTE evaluation or methods used for HTE evaluation in research on the health effects of social policies.

We leverage an existing sample of studies on the health effects of social policies published in leading health and social science journals in 2019 [24]. We review these studies to determine how frequently HTEs are evaluated in contemporary social policy research, to characterize the distribution of design decisions (e.g., whether the dimensions for HTE evaluation were specified *a priori*) and methods for HTE evaluation, and to describe the population subgroups for which HTEs were reported. Findings will help to identify gaps in HTE assessment in research on the health effects of social policies and inform methodological standards for future research.

Materials and methods

Identification of social policy studies

We used a sample of social policy studies developed in the course of a previously published systematic review to evaluate the reporting of HTEs in research on the health effects of social policies [24]. Briefly, the sample included studies published in 2019 evaluating the health effects of a social policy in a multidisciplinary set of high-impact journals publishing research on the health effects of social policies: *American Journal of Public Health, American Journal of Epidemiology, Journal of the American Medical Association, New England Journal of Medicine, The Lancet, American Journal of Preventive Medicine, Social Science and Medicine, Health Affairs, Demography*, and *American Economic Review*. The relevance of these journals was subsequently evaluated with a convenience sample of 66 researchers from diverse disciplines who were asked to rank the most relevant journals. The results of the survey confirmed our view that these journals represent a common perception of the most appropriate outlets to conduct research on the health effects of social policies (detailed results are presented in the Web Appendix of [24])

To construct the sample, the authors screened all 6794 articles published in these journals in 2019 and identified all empirical studies evaluating effects of one or more social policies on health-related outcomes ($N = 55$). Inclusion criteria included evaluation of a health-related

outcome (broadly defined to include morbidity, mortality, health conditions, and factors such as smoking, homelessness, and sales of unhealthy products) and evaluation of a non-medical policy that was adopted at a community or higher level and that was hypothesized to affect health or health inequities via changes in social or behavioral determinants. Additional details can be found elsewhere [24].

Data extraction and analysis

We re-abstracted the studies in the original sample using a structured data extraction form (Appendix Table A.1) to collect information on both the policy domain and the study design or causal identification strategy (e.g., instrumental variable or difference-in-differences). The primary outcome was whether the study reported an HTE analysis. We classified studies as reporting an HTE analysis if they reported subgroup differences in the causal effect of the social policy on the health-related outcome(s) (details in Appendix Fig. A.1).

For studies that reported HTEs, we captured the methods used for this evaluation, and subsequently categorized these methods as: stratified analysis between subgroups; regression with an interaction term(s); or data-driven algorithms (e.g., machine learning to identify which subgroups benefit most from treatment). Because we expected that many HTE assessments might be conducted post hoc and therefore lack statistical precision, we also recorded whether authors prespecified subgroups for HTE evaluation. Finally, we tabulated the population dimensions along which HTEs were assessed (e.g., age, gender, race/ethnicity). For each field, we calculated descriptive frequencies.

Results

Study characteristics

The sample of social policy studies included several different countries of origin (e.g., United States, Canada, Mexico, Malawi), sample sizes (range: 15–5 million), and policy levels (i.e., country, state, local). For further details on the sample of social policy studies, see Appendix Table A.2 and [24,25]. Of the 55 studies, we excluded one that reported only simulated policy effects. Across the remaining 54 studies, causal effect estimates were evaluated for a range of social policy domains including family, maternal, and child health (n = 12), income and employment (n = 10), food and beverage (n = 6), firearms (n = 5), immigration (n = 4), alcohol (n = 4), education (n = 3), tobacco (n = 3), austerity/economic (n = 2), housing (n = 1), cannabis (n = 1), road traffic safety (n = 1), same-sex marriage (n = 1), and voting (n = 1). The study designs employed were: difference-in-differences (n = 12), before-after (n = 9), regression (n = 7), panel-fixed effects (n = 6), instrumental-variable (n = 3), propensity score matching (n = 1), randomized stepped wedge (n = 1), synthetic control (n = 1), and comparative interrupted time series (n = 1).

HTE analyses: planning, reporting, methods, and subgroups

HTEs were reported in 44% (n = 24) of the studies, and of those, 63% (n = 15) specified their intent to assess HTEs *a priori* (Table 2). Most (n = 17, 71%) studies evaluated HTEs with stratification but not statistical tests; 5 studies (21%) assessed HTEs with statistical tests by including an interaction term in a regression framework; and 2 studies (8%) used

both approaches. None used data-driven algorithms. HTEs were evaluated for numerous individual characteristics—most frequently geographic location, gender, education, and age (Fig. 1). HTEs were also evaluated across several characteristics unique to specific studies such as immigration status, hospital delivery level (e.g., secondary vs. tertiary delivering hospital) and body mass index.

Discussion

About half of the contemporary studies on the health effects of social policies reported examining HTEs. Geographic location, age, gender, and education defined the subgroups most frequently assessed for HTEs, but even these were evaluated in fewer than one in five studies. Most studies that reported HTEs prespecified the dimensions of heterogeneity tested; no studies used machine learning to evaluate heterogeneity.

Theory or evidence suggest that HTEs are likely for many social interventions. For example, compared to those with high childhood socioeconomic status, individuals with low childhood socioeconomic status benefit more from additional education [26]. HTEs by demographic subgroups have been documented for a variety of social exposures including education [27], exposure to community violence [28], and poverty programs [29]. An important insight from research on structural racism is that policies which, at a cursory level, seem race-neutral have differential consequences by race and exacerbate racial inequities [30,31]. Given this, it is striking that HTEs are not evaluated more consistently, especially HTEs by race/ethnicity.

Our findings indicate a lack of broadly adopted standards on what studies should evaluate HTEs, for what groups HTEs should be evaluated, and how to evaluate them [32,33]. Several recent articles offer suggestions for potential approaches to estimating and reporting HTEs, although this work is almost entirely focused on clinical trials and studies of medical interventions [9,18,33–46].

Core epidemiology methods texts, for example, provide guidance on testing statistical interaction but no guidance on when such a test would be desirable [47]. Are heterogeneities in the health effects of social policies frequently large enough to alter recommendations for policy adoption? Currently, we do not know. Collecting the sample sizes needed to evaluate policy effects disaggregated by subgroup is expensive and time-consuming. However, it is essential to understand how frequently large differences in treatment effects occur. Evidence of HTEs is important for understanding for whom a policy is effective and the likely implications of the policy for health inequities.

One possible explanation for infrequent evaluation of HTEs is that papers on HTEs may have been published separately. The journals included in our systematic sample may not be representative of the field-wide norms, although they represent an interdisciplinary swath of high-impact journals [25]. Further, many studies are likely underpowered to detect HTEs, which typically require larger sample sizes to detect than overall effects. Given the pressure to suppress results that do not meet statistical significance thresholds [48], analyses finding no statistically significant evidence of HTEs may not be published. This type of publication

bias will lead to suppression of both underpowered studies and studies with adequate power to detect important effect heterogeneity, where no such heterogeneity was found. It is, therefore, possible that HTEs are assessed more frequently than the published literature reflects. [49,50]

For many small subgroups, no single study is likely to include a large enough sample to identify differential effects in that subgroup; publishing findings with appropriate information on uncertainty may allow for later meta-analysis. When small sample sizes hamper a researcher's ability to evaluate HTEs, mixed methods or qualitative approaches may be helpful, for example, to identify important modifiers and distinct participant or subgroup experiences that were unexpected. [51,52] Resource limitations or lack of study sample diversity may also be barriers to evaluation of HTEs. Creating standards for routine reporting of certain HTE tests might reduce the risk of such suppressed results. Given these issues, in this brief report, we did not focus on the statistical significance of the HTEs reported but rather focus on the frequency of reporting of HTEs across studies. In future research, we plan to provide more detail on the distribution and magnitude of HTEs across the social policy studies herein.

Just as there is no consensus on when to evaluate HTEs, there is also no consensus on how to evaluate HTEs. Researchers must balance the importance of identifying meaningful variation against the increasing possibility of chance findings as more subgroups are evaluated. Pre-specification of the intent to evaluate HTEs and rationale for subgroup selection only partially ameliorates this problem. To avoid publication bias, null results should be routinely reported for all *a priori* specified groups as well as for exploratory analyses that were not prespecified [9,53,54]. For small subgroups, substantively important differences in effects may be imprecise. If these results are reported, they can be incorporated into future evidence reviews or meta-analyses. Qualitative research can help identify sources of effect heterogeneity and are especially important in early evaluations of a policy when there is little prior evidence to guide hypothesis generation. Visualization of subgroup effect sizes has also been flagged as a critical aid in the identification of subgroup differences and communication of HTE results to wider audiences [55]. Additionally, although no studies in our sample utilized data-driven algorithms, these methods may be well-suited to addressing challenges associated with multiple testing. Future research should investigate how often data-driven methods discover relevant, theoretically surprising, and reproducible differences in treatment effects of social policies. It would be valuable to identify which methods are most robust when sample sizes are limited.

Conclusions

There is significant opportunity for improvement in the design, reporting, and interpretation of analyses used for the identification of HTEs in research on the health effects of social policies. Evaluating the extent to which policies differentially affect people of different race/ethnicities, genders, socioeconomic status, or geographic region is a foundational step for identifying effective strategies to promote health equity. Despite the relevance of HTEs for social inequities, evaluating HTEs is not standard, and the methods adopted for evaluating HTEs vary. Additional guidance is needed on what dimensions of HTEs should be evaluated

and when, and how such evaluations should be conducted and reported. To the extent that promoting health equity is a goal of social interventions and understanding HTEs is a priority for research on social policies, addressing barriers to evaluating HTEs should be prioritized.

Acknowledgments

This work was supported by the Evidence for Action program of the Robert Wood Johnson Foundation (RWJF).

Appendices

Table A.1

Data extraction checklist.

Feature extracted	Description
Article Extraction Checklist	
Identification number	Identification number for study
Journal name	The journal that the study was published in
Journal impact factor	The journal impact factor
Substantive policy domain	The substantive domain of the policy
Study design or causal identification strategy	The analytical study design for causal identification (e.g., instrumental variable or difference-in-differences)
Years of social policy intervention	The relevant years for the policy intervention
Sample size	The sample size in the study
Heterogeneous treatment effect (HTE) Extraction Checklist	
Outcome	The relevant outcome for HTE estimate
Outcome variable type	The type of measure of the outcome (e.g., binary or continuous)
HTE assessed	A dummy indicator of whether a study assessed HTEs
Intent to assess HTE specified *a priori*	A dummy indicator of whether a study specified the HTE analysis *a priori*
HTE supported by theory	A dummy indicator of whether the choice of HTE analyses supported by theory
HTE method	Specification of method for assessing HTE, if applicable
Subgroups for HTE estimate	What groups were specified to test HTE
HTE estimates	Estimate of HTE by group
Subgroup sample sizes	Sample size of subgroup in HTE analysis
Standard error HTE estimates	The standard error of HTE estimate
Confidence interval (CI) HTE estimates	The CI for HTE estimates
Statistical significance of HTE estimates	A dummy indicator for whether HTE estimate statistically significant
p-value	The *p*-value for HTE estimate
Effect measure (measure of association)	The effect measure or measure of association of HTE estimate (e.g., relative risk, odds ratio, or standardized mean difference)

Table A.2

Study details.

Study title	HTE Evaluation	Sample Size	Country of Origin	Impact Level of Social Policy
1. The Violent Legacy of Conflict: Evidence on Asylum Seekers, Crime, and Public Policy in Switzerland	No	23223	Switzerland	Local
2. Associations Between Gun Laws and Suicides	Yes	416391	United States	State
3. The Impact of the Revised WIC Food Package on Maternal Nutrition During Pregnancy and Postpartum	No	1454	United States	Country and State
4. Restrictive Immigration Law and Birth Outcomes of Immigrant Women	Yes	5352146	United States	State
5. Mortality in Spain in the Context of the Economic Crisis and Austerity Policies	Yes	16	Spain	Country
6. The Rates and Medical Necessity of Cesarean Delivery in the Era of the Two-Child Policy in Hubei and Gansu Provinces, China	Yes	121722	China	Country
7. Alcohol Availability Across Neighborhoods in Ontario Following Alcohol Sales Deregulation, 2013–2017	No	19964	Canada	Province
8. Post-Legalization Opening of Retail Cannabis Stores and Adult Cannabis Use in Washington State, 2009–2016	No	85135	United States	State
9. Impact of the Food-Labeling and Advertising Law Banning Competitive Food and Beverages in Chilean Public Schools, 2014–2016	Yes	21	Chile	Country
10. Texting-While-Driving Bans and Motor Vehicle Crash–Related Emergency Department Visits in 16 US States: 2007–2014	Yes	1344	United States	State
11. Sugar-Sweetened Beverage Consumption 3 Years After the Berkeley, California, Sugar-Sweetened Beverage Tax	No	5225	United States	City
12. Paid Family Leave Effects on Breastfeeding: A Quasi-Experimental Study of US Policies	Yes	306266	United States	State
13. Impact of a Municipal Policy Restricting Trans Fatty Acid Use in New York City Restaurants on Serum Trans Fatty Acid Levels in Adults	No	459	United States	City
14. Austerity Policies and Mortality Rates in European Countries, 2011–2015	No	75	European Countries	Country
15. The Mental Health of Hispanic/Latino Americans Following National Immigration Policy Changes: United States, 2014–2018	No	118883	United States	Country
16. Housing and Urban Development–Veterans Affairs Supportive Housing Vouchers and Veterans' Homelessness, 2007–2017	No	3850	United States	Country
17. The Effects of SNAP Work Requirements in Reducing Participation and Benefits From 2013 to 2017	No	24100	United States	Country and State
18. Right-to-Carry Laws and Firearm Workplace Homicides: A Longitudinal Analysis (1992–2017)	Yes	1300	United States	State
19. The Effect of Large-Capacity Magazine Bans on High-Fatality Mass Shootings, 1990–2017	No	1428	United States	State

Study title	HTE Evaluation	Sample Size	Country of Origin	Impact Level of Social Policy
20. Firearm and Nonfirearm Violence After Operation Peacemaker Fellowship in Richmond, California, 1996–2016	No	2649	United States	City
21. Dietary Guidance and New School Meal Standards: Schoolchildren's Whole Grain Consumption Over 1994–2014	No	17016	United States	Country
22. Supermarket Purchases Over the Supplemental Nutrition Assistance Program Benefit Month: A Comparison Between Participants and Nonparticipants	No	950	United States	Country and State
23. Short-Term Impact of a Flavored Tobacco Restriction: Changes in Youth Tobacco Use in a Massachusetts Community	No	158	United States	Local
24. Association Between State Minimum Wages and Suicide Rates in the U.S.	Yes	550	United States	State
25. SNAP, Young Children's Health, and Family Food Security and Healthcare Access	No	28782	United States	Country and State
26. Association of State Firearm Legislation With Female Intimate Partner Homicide	No	1693	United States	State
27. Alcohol Policies and Alcohol Involvement in Intimate Partner Homicide in the U.S.	Yes	2729	United States	State
28. Smoke-Free Policies and 30-Day Readmission Rates for Chronic Obstructive Pulmonary Disease	No	1788	United States	State and Local
29. The Minnesota SimSmoke Tobacco Control Policy Model of Smokeless Tobacco and Cigarette Use	No	-[a]	United States	State
30. State-Level Beer Excise Tax and Firearm Homicide in Adolescents and Young Adults	No	12	United States	State
31. Legalizing Same-Sex Marriage Matters for the Subjective Well-being of Individuals in Same-Sex Unions	No	476411	Cross-national	Country
32. The Effect of the Earned Income Tax Credit on Housing and Living Arrangements	No	853012	United States	State
33. Moving Upstream: The Effect of Tobacco Clean Air Restrictions on Educational Inequalities in Smoking Among Young Adults	Yes	42132	United States	Local
34. Reexamining the Influence of Conditional Cash Transfers on Migration From a Gendered Lens	Yes	21803	Mexico	Country
35. Uncertainty About DACA May Undermine Its Positive Impact On Health For Recipients And Their Children	Yes	16697	United States	Country
36. Evaluating A USDA Program That Gives SNAP Participants Financial Incentives To Buy Fresh Produce In Supermarkets	No	32	United States	Country and State
37. The Effect Of The Supplemental Nutrition Assistance Program On Mortality	No	970137	United States	State
38. Loss Of SNAP Is Associated With Food Insecurity And Poor Health In Working Families With Young Children	No	8569	United States	Country and State
39. Association of a Beverage Tax on Sugar-Sweetened and Artificially Sweetened Beverages With Changes in Beverage Prices and Sales at Chain Retailers in a Large Urban Setting	No	291	United States	City
40. An evaluation of the effects of lowering blood alcohol concentration limits for drivers on the rates	Yes	561646	Cross-national	Country

Study title	HTE Evaluation	Sample Size	Country of Origin	Impact Level of Social Policy
of road traffic accidents and alcohol consumption: a natural experiment				
41. Sugar-based beverage taxes and beverage prices: Evidence from South Africa's Health Promotion Levy	No	71677	South Africa	Country
42. Effects of a voter initiative on disparities in punishment severity for drug offenses across California counties	No	451139	United States	State
43. Government of Malawi's unconditional cash transfer improves youth mental health	Yes	1366	Malawi	Country
44. Unconditional cash transfers and parental obesity	Yes	60682	Canada	Country
45. Do comprehensive school reforms impact the health of early school leavers? Results of a comparative difference-in-difference design	Yes	220408	European Countries	Country
46. Effects of tuition-free primary education on women's access to family planning and on health decision-making: A cross-national study	No	429001	Cross-national	Country
47. The impact of employment protection on health: Evidence from fixed-term contract workers in South Korea	Yes	2683	South Korea	Country
48. A conditional cash transfer and Women's empowerment: Does Bolsa Familia Influence intimate partner violence?	No[b]	12543	Brazil	Country and local
49. Impact of an employment guarantee scheme on utilisation of maternal healthcare services: Results from a natural experiment in India	Yes	127879	India	Country
50. Center-based childcare expansion and grandparents' employment and well-being	Yes	11598	China	Province
51. Does money relieve depression? Evidence from social pension expansions in China	Yes	8636	China	Country
52. The effect of unemployment benefits on health: A propensity score analysis	Yes	7558	Canada	Country
53. Changes in maternity leave coverage: Implications for fertility, labour force participation and child mortality	Yes	396	Africa and Asia	Country
54. SNAP benefits and childhood asthma	No	2477560	United States	Country and State
55. Education system stratification and health complaints among school-aged children	Yes	184160	Cross-national	Country

Note.

[a] - Study excluded because results based on simulation model;

[b] - This study reported that they performed a Heterogenous Treatment Effect (HTE) evaluation but did not report the results.

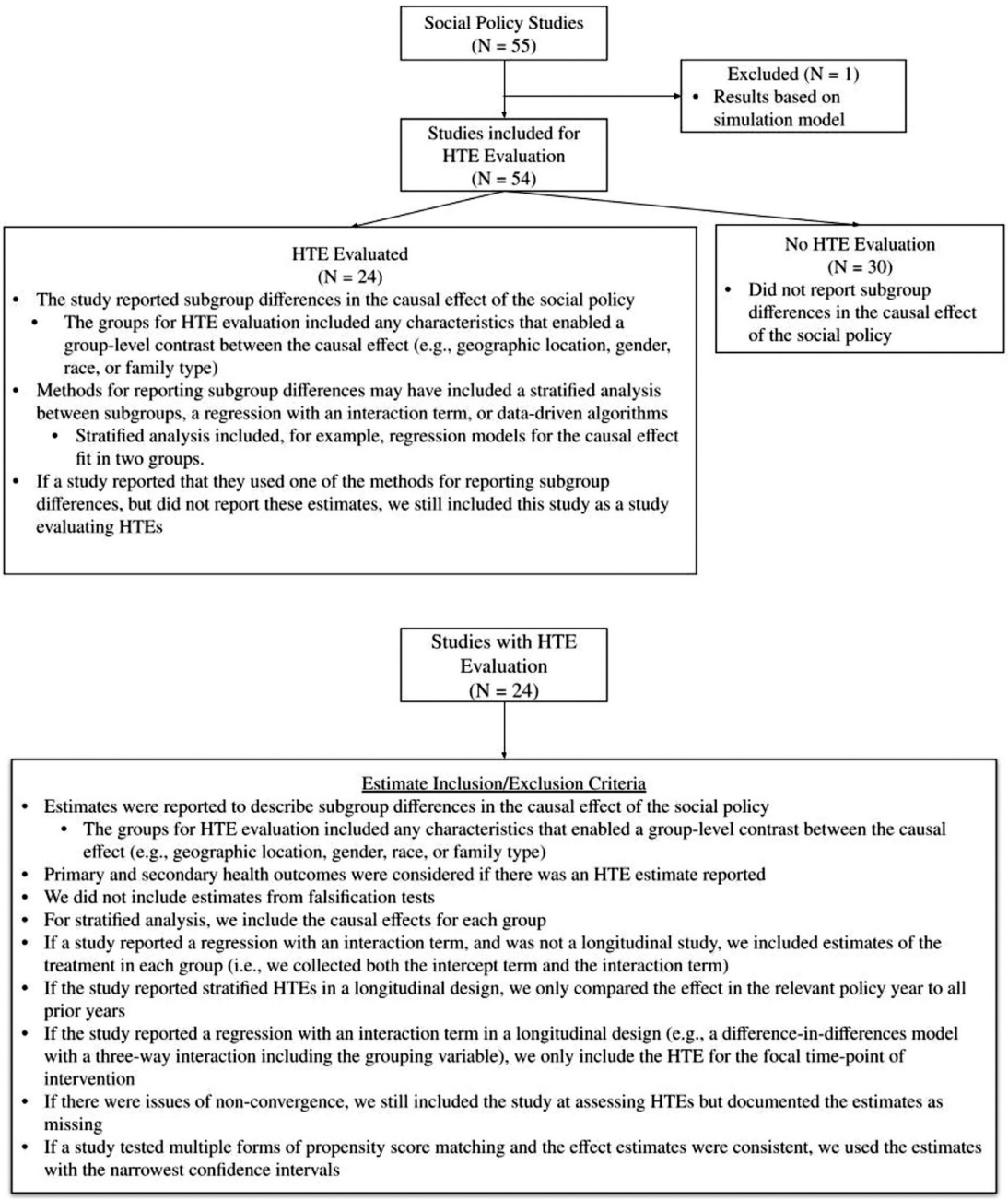

Fig. A.1.
Detailed inclusion/exclusion criteria used for identifying studies with HTE analyses.

References

[1]. Vable AM, Canning D, Glymour MM, Kawachi I, Jimenez MP, Subramanian SV. Can social policy influence socioeconomic disparities? Korean War GI Bill eligibility and markers of depression. Ann Epidemiol 2016;26(2):129–35 e3. [PubMed: 26778285]

[2]. Butler SM, Ashford JW, Snowdon DA. Age, education, and changes in the mini–mental state exam scores of older women: findings from the nun study. J Am Geriatr Soc 1996;44(6):675–81. [PubMed: 8642159]

[3]. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ 2020;192(32):E901–6. [PubMed: 32778601]

[4]. Matthay EC. Who benefits most? The importance of identifying whether social interventions have different effects for different people. Evidence for Action 2020 [accessed 08.06.21]. Available from: https://www.evidenceforaction.org/blog-posts/who-benefits-most-importance-identifying-whether-social-interventions-have-different

[5]. Matthay EC Why is there so much uncertainty about heterogeneous treatment effects? [Internet]. Evidence for Action. 2020 [accessed 08.06.21]. Available from: https://www.evidenceforaction.org/sites/default/files/2021-04/E4A-Methods-Note-HTEp2.pdf

[6]. Matthay EC Do social interventions have different health effects for different people? Why heterogeneous treatment effects are important in population health research [Internet]. 2020 [accessed 08.06.21]. Available from: https://www.evidenceforaction.org/sites/default/files/2021-04/E4A-Methods-Note-HTEp1.pdf

[7]. Ward JB, Gartner DR, Keyes KM, Fliss MD, McClure ES, Robinson WR. How do we assess a racial disparity in health? Distribution, interaction, and interpretation in epidemiological studies. Ann Epidemiol 2019;29:1–7. [PubMed: 30342887]

[8]. Raghavan S, Josey K, Bahn G, Reda D, Basu S, Berkowitz SA, et al. Generalizability of heterogeneous treatment effects based on causal forests applied to two randomized clinical trials of intensive glycemic control. Ann Epidemiol 2021. [accessed 27.07.21]Available from https://www.sciencedirect.com/science/article/pii/S104727972100212X.

[9]. Varadhan R, Seeger JD. Estimation and reporting of heterogeneity of treatment effects. Developing a protocol for observational comparative effectiveness research: a user's guide. Velentgas P, Dreyer NA, Nourjah P, Smith SR, Torchia MM, editors, Rockville, MD: Agency for Healthcare Research and Quality; 2013. [accessed 07.05.21]Available from https://www.ncbi.nlm.nih.gov/books/NBK126188.

[10]. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proc Natl Acad Sci USA 2016;113(27):7353–60. [PubMed: 27382149]

[11]. Bornkamp B, Ohlssen D, Magnusson BP, Schmidli H. Model averaging for treatment effect estimation in subgroups. Pharm Stat 2017;16(2):133–42. [PubMed: 27935199]

[12]. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113(523):1228–42.

[13]. Clogg CC, Petkova E, Haritou A. Statistical methods for comparing regression coefficients between models. Am J Sociol 1995;100(5):1261–93.

[14]. Paternoster R, Brame R, Mazerolle P, Piquero A. Using the correct statistical test for the equality of regression coefficients. Criminology 1998;36(4):859–66.

[15]. Loh WY, Cao L, Zhou P. Subgroup identification for precision medicine: a comparative review of 13 methods. WIREs Data Min Knowl Discov 2019;9(5):e1326.

[16]. Fan J, Song F, Bachmann MO. Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. J Clin Epidemiol 2019;108:17–25. [PubMed: 30557676]

[17]. Fernandez y Garcia E, Nguyen H, Duan N, Gabler NB, Kravitz RL. Assessing heterogeneity of treatment effects: are authors misinterpreting their results? Health Serv Res 2010;45(1):283–301. [PubMed: 19929962]

[18]. Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? Trials 2009;10(1):43. [PubMed: 19545379]

[19]. Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Blumle A, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. BMJ 2014;349(1) g4539–g4539. [PubMed: 25030633]

[20]. Starks MA, Sanders GD, Coeytaux RR, Riley IL, Jackson LR, Brooks AM, et al. Papageorgiou SN, editor Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: a systematic review. editor. PLoS One 2019;14(8):e0219894.

[21]. Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ 2012;344(1) e1553–e1553. [PubMed: 22422832]

[22]. Office of Extramural Research National Institutes of Health. NIH policy and guidelines on the inclusion of women and minorities as subjects in clinical research. [accessed 07.05.21] Available from: https://grants.nih.gov/grants/funding/women_min/women_min.htm. 2001.

[23]. U.S. Food and Drug Administration. Guidance for industry: collection of race and ethnicity data in clinical trials. [accessed 07.05.21] Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/collection-race-and-ethnicity-data-clinical-trials. 2016.

[24]. Matthay EC, Hagan E, Joshi S, Tan ML, Vlahov D, Adler N, et al. The revolution will be hard to evaluate: how co-occurring policy changes affect research on the health effects of social policies. Epidemiol Rev 2022;43(1):19–32. [PubMed: 34622277]

[25]. Matthay EC, Gottlieb LM, Rehkopf D, Tan ML, Vlahov D, Glymour MM. What to do when everything happens at once: analytic approaches to estimate the health effects of co-occurring social policies. Epidemiol Rev 2022;43(1):33–47. [PubMed: 34215873]

[26]. Vable AM, Cohen AK, Leonard SA, Glymour MM, Duarte C dP, Yen IH. Do the health benefits of education vary by sociodemographic subgroup? differential returns to education and implications for health inequities. Ann Epidemiol 2018;28(11):759–66 e5. [PubMed: 30309690]

[27]. Ross CE, Mirowsky J. Sex differences in the effect of education on depression: resource multiplication or resource substitution? Soc Sci Med 2006;63(5):1400–13. [PubMed: 16644077]

[28]. Matthay EC, Farkas K, Ahern J. Racial and ethnic differences in associations of community violence with self-harm: a population-based case-control study. Ann Epidemiol 2019;34:71–4. [PubMed: 31072682]

[29]. Komro KA, Markowitz S, Livingston MD, Wagenaar AC. Effects of state-level earned income tax credit laws on birth outcomes by race and ethnicity. Health Equity 2019;3(1):61–7. [PubMed: 30886942]

[30]. Krieger N. Public health, embodied history, and social justice: looking forward. Int J Health Serv 2015;45(4):587–600. [PubMed: 26182941]

[31]. Krieger N. Structural racism, health inequities, and the two-edged sword of data: structural problems require structural solutions. Front Public Health 2021. [accessed 27.07.21]Available from https://www.frontiersin.org/articles/10.3389/fpubh.2021.655447/full?utm_source=S-TWT&utm_medium=SNET&utm_campaign=ECO_FPUBH_XXXXXXXX_auto-dlvrit.

[32]. Thomas M, Bornkamp B. Comparing approaches to treatment effect estimation for subgroups in clinical trials. Stat Biopharm Res 2017;9(2):160–71.

[33]. Schandelmaier S, Chang Y, Devasenapathy N, Devji T, Kwong JSW, Colunga Lozano LE, et al. A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. J Clin Epidemiol 2019;113:159–67. [PubMed: 31132471]

[34]. Donegan S, Williams L, Dias S, Tudur-Smith C, Welton N. Exploring treatment by covariate interactions using subgroup analysis and meta-regression in cochrane reviews: a review of recent practice. PLoS One 2015;10(6):e0128804. [PubMed: 26029923]

[35]. Gil-Sierra MD, Fénix-Caballero S, Abdel kader-Martin L, Fraga-Fuentes MD, Sánchez-Hidalgo M, Alarcón de la Lastra-Romero C, et al. Checklist for clinical applicability of subgroup analysis. J Clin Pharm Ther Jun 2020;45(3):530–8. [PubMed: 31854128]

[36]. Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2018;100:22–31. [PubMed: 29654822]

[37]. Nasser M, van Weel C, van Binsbergen JJ, van de Laar FA. Generalizability of systematic reviews of the effectiveness of health care interventions to primary health care: concepts, methods and future research. Fam Pract 2012;29(Suppl. 1):i94–103. [PubMed: 22399564]

[38]. Petkovic J, Jull J, Yoganathan M, Dewidar O, Baird S, Grimshaw JM, et al. Reporting of health equity considerations in cluster and individually randomized trials. Trials 2020;21(1):308. [PubMed: 32245522]

[39]. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. CMAJ 2020;192(32):E901–6. [PubMed: 32778601]

[40]. Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010;340(3) c117–c117. [PubMed: 20354011]

[41]. Sun X, Ioannidis JPA, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014;311(4):405. [PubMed: 24449319]

[42]. van Hoorn R, Tummers M, Booth A, Gerhardus A, Rehfuess E, Hind D, et al. The development of CHAMP: a checklist for the appraisal of moderators and predictors. BMC Med Res Methodol 2017;17(1):173. [PubMed: 29268721]

[43]. Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol 2013;66(8):818–25. [PubMed: 23651763]

[44]. Welch VA, Akl EA, Pottie K, Ansari MT, Briel M, Christensen R, et al. GRADE equity guidelines 3: considering health equity in GRADE guideline development: rating the certainty of synthesized evidence. J Clin Epidemiol 2017;90:76–83. [PubMed: 28389397]

[45]. Whitlock EP, Eder M, Thompson JH, Jonas DE, Evans CV, Guirguis-Blake JM, et al. An approach to addressing subpopulation considerations in systematic reviews: the experience of reviewers supporting the U.S. preventive services task force. Syst Rev 2017;6(1):41 s13643–017-0437–3. [PubMed: 28253915]

[46]. Tipton E, Yeager DS, Iachan R, Schneider B. Designing probability samples to study treatment effect heterogeneity. In: Lavrakas P, Traugott M, Kennedy C, Holbrook A, de Leeuw E, West B, editors. Experimental methods in survey research [Internet]. Wiley; 2019. p. 435–56. [accessed 31.03.21]Available from.

[47]. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Lippincott Williams & Wilkins; 2008. 776 p.

[48]. Gelman A, Loken E. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. Am. Sci 2014;102(6):460.

[49]. Shields PG. Publication bias is a scientific problem with adverse ethical outcomes: the case for a section for null results. Cancer Epidemiol Biomark Prev 2000;9(8):771–2.

[50]. Marks-Anglin A, Chen Y. A historical review of publication bias. Res Synth Methods 2020;11(6):725–42. [PubMed: 32893970]

[51]. Davidson KM, Young JTN. Treatment engagement in a prison-based therapeutic community: a mixed-methods approach. J Subst Abuse Treat 2019 Aug 1;103:33–42. [PubMed: 31229190]

[52]. Bamberger M, Rao V, Woolcock M. Using mixed methods in monitoring and evaluation: experiences from international development. Washington, DC: World Bank Policy Research Working Paper; 2010.

[53]. Glymour MM, Osypuk TL, Rehkopf DH. Invited commentary: off-roading with social epidemiology—exploration, causation, translation. Am. J. Epidemiol 2013;178(6):858–63. [PubMed: 24008902]

[54]. Rehkopf DH, Glymour MM, Osypuk TL. The consistency assumption for causal inference in social epidemiology: when a rose is not a rose. Curr Epidemiol Rep 2016;3(1):63–71. [PubMed: 27326386]

[55]. Ballarini NM, Chiu YD, König F, Posch M, Jaki T. A critical review of graphics for subgroup analyses in clinical trials. Pharm Stat 2020;19(5):541–60. [PubMed: 32216035]

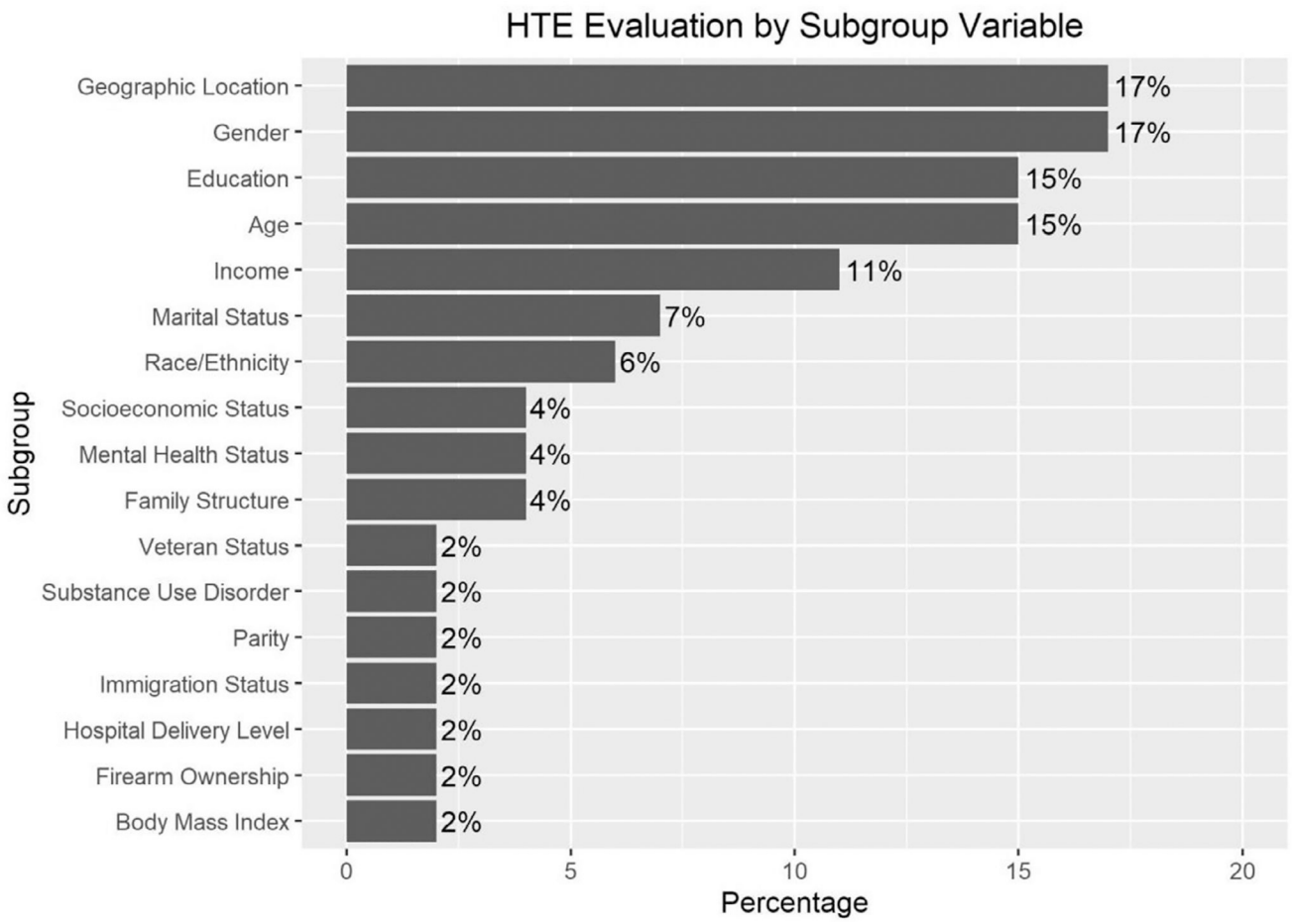

Fig. 1.

Subgroup variable representation in HTE analyses in this study sample ($N = 54$). The denominator for the percentages is 54: the total number of studies evaluated. Of the 54 studies, 30 did not report any HTEs.

Table 1

Evaluation of heterogeneous treatment effects in research on the health effects of social policies: research questions and methods.

Research Question	Method	Overview
What are the estimated effects of the policy within prespecified subgroups (e.g., defined by age, race/ethnicity, or education)?	Stratification	Separate effect estimates are derived by analyzing each population subgroup separately, with population subgroups defined by prespecified characteristics. Stratification is a descriptive approach and does not involve statistical testing for differences in treatment effects between population subgroups. Stratification analyses do not directly test for HTEs but provide results than can be used for formal tests of HTEs in future meta-analyses and systematic reviews. Statistical methods are available for comparing regression coefficients between models [13, 14], although no studies evaluated herein used these methods.
Do estimated effects of the policy differ between prespecified subgroups to a degree inconsistent with chance variation?	Test for interaction	Tests for interaction are one of the most commonly used approaches to examine HTEs. Traditionally, a test for interaction is performed by incorporating an interaction term between the treatment variable and a prespecified modifying variable in a regression model of the outcome. The modifying variable defines subgroups that are hypothesized to respond differently to the treatment. This analysis is used to determine if the treatment has substantively meaningful and statistically significant differences in effect for population subgroups defined by the modifying variable.
Is there evidence of HTEs across subgroupings defined by any measured variables, even if the subgroups are not prespecified?	Data-driven algorithms for HTE identification	Data driven algorithms can be used for exploratory purposes and do not require pre-specification of the subgroups of interest. However, these methods usually rely on prespecifying the candidate variables along which heterogeneity will be assessed (e.g., age, gender, race/ethnicity). These methods typically allow for complex interactions between the variables in defining subgroups (e.g., intersections of race, age, gender, region). The algorithms that will be used to assess HTEs must also be chosen *a priori*, although ensemble methods can be adopted to combine the best performing algorithms. See for example Loh et al. [15].

Table 2

Primary indicators and study findings in this quantitative assessment of HTE evaluation.

Question	Rationale for Assessment	Results
(1) How often were HTEs assessed?	Assessing HTEs is important for several reasons. For instance, for understanding potential inequities in social policies. HTEs may also provide an explanation for the magnitude of an effect size.	24 of 54 (44%) of studies assessed HTEs.
(2) How often was the intent to assess HTEs specified *a priori*?	*A priori* specification is evidence of some plan for assessing HTEs.	15 of 24 (63%) specified their intent to assess HTEs *a priori*.
(3) What methods for assessing HTEs were used?	The type of method used to assess an HTE has implications for what inferences can be made regarding an HTE (e.g., confirmatory, descriptive, or exploratory).	17 of 24 (71%) studies used descriptive methods based on stratification to evaluate HTEs. 5 of 24 (21%) studies used statistical inference methods to evaluate HTEs (e.g., an interaction term between the social policy and the subgroup for HTE). 2 of 24 (8%) studies used a combination of descriptive and inferential methods to evaluate HTEs.
(4) For what subgroups were HTEs assessed?	The subgroups provide a sense of what types of factors are being considered as potential sources of treatment effect heterogeneity.	Diverse subgroups were tested to explore HTEs. See Fig. 1 where 17 different subgroups were identified to evaluate HTEs.

Note. THE = heterogeneous treatment effect.