SSM - Population Health 22 (2023) 101352

Contents lists available at ScienceDirect

SSM - Population Health

journal homepage: www.elsevier.com/locate/ssmph

A quantitative assessment of the frequency and magnitude of heterogeneous treatment effects in studies of the health effects of social policies

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

[a] Center for Health and Community, University of California, San Francisco, 3333 California St., Suite 465, Campus Box 0844, San Francisco, CA, 94143, USA
[b] Department of Epidemiology and Biostatistics, University of California, San Francisco, 550 16th Street, 2nd Floor, Campus Box 0560, San Francisco, CA, 94143, USA
[c] Yale School of Nursing at Yale University, 400 West Campus Drive, Room 32306, Orange, CT, 06477, USA
[d] Center for Opioid Epidemiology and Policy, Division of Epidemiology, Department of Population Health, New York University School of Medicine, 180 Madison Ave, New York, NY, 10016, USA

ARTICLE INFO

Keywords:
Population heterogeneity
Effect modifiers
Social policy
Health equity

ABSTRACT

Substantial heterogeneity in effects of social policies on health across subgroups may be common, but has not been systematically characterized. Using a sample of 55 contemporary studies on health effects of social policies, we recorded how often heterogeneous treatment effects (HTEs) were assessed, for what subgroups (e.g., male, female), and the subgroup-specific effect estimates expressed as Standardized Mean Differences (SMDs). For each study, outcome, and dimension (e.g., gender), we fit a random-effects meta-analysis. We characterized the magnitude of heterogeneity in policy effects using the standard deviation of the subgroup-specific effect estimates (τ). Among the 44% of studies reporting subgroup-specific estimates, policy effects were generally small (<0.1 SMDs) with mixed impacts on health (67% beneficial) and disparities (50% implied narrowing of disparities). Across study-outcome-dimensions, 54% indicated any heterogeneity in effects, and 20% had $\tau > 0.1$ SMDs. For 26% of study-outcome-dimensions, the magnitude of τ indicated that effects of opposite signs were plausible across subgroups. Heterogeneity was more common in policy effects not specified *a priori*. Our findings suggest social policies commonly have heterogeneous effects on health of different populations; these HTEs may substantially impact disparities. Studies of social policies and health should routinely evaluate HTEs.

1. Introduction

Social policies may have substantial impacts on a broad range of population health outcomes, and a growing body of health research seeks to quantify their causal effects (Matthay & Glymour, 2022). However, less research has evaluated differences in the effects of social policies across population subgroups. Many social policies could plausibly benefit some members of the community while harming others, or have larger or smaller benefits across population subgroups. For example, racist social policies widen disparities between racial groups (e.g., in health, housing, education, or policing), whereas anti-racist social policies narrow racial disparities by dismantling the racism embedded in social, economic, and political institutions (Boykin et al., 2020; Kendi, 2019).

Assessing heterogeneous treatment effects (HTEs) of social policies is critical to understand the implications of these policies for health inequities, and epidemiology researchers have increasingly called for HTE assessments for this reason (Matthay & Glymour, 2022). The health effects of a policy on the population overall need not be in the same direction as the effects on inequities: policies that improve average health may exacerbate inequities, or conversely, policies that harm health on average may nonetheless narrow inequities. Social policies that primarily benefit those with better health at baseline are likely to widen inequities whereas social policies that primarily benefit those with the poorest health may reduce inequities. For example, the Korean War GI Bill, which provided socioeconomic benefits to veterans, was associated

* Corresponding author. Center for Opioid Epidemiology and Policy, Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, NY, 10016, USA.
E-mail address: ellicott.matthay@nyulangone.org (E.C. Matthay).

https://doi.org/10.1016/j.ssmph.2023.101352
Received 9 November 2022; Received in revised form 25 January 2023; Accepted 26 January 2023
Available online 4 February 2023

with fewer subsequent depressive symptoms for veterans from backgrounds with low childhood socioeconomic status (SES) but not those from high childhood SES backgrounds (Vable et al., 2016); if veterans from low SES backgrounds had more depressive symptoms at baseline, then this policy may have reduced inequities in depressive symptoms. Well-controlled assessments of HTEs across subgroups contribute evidence on whether social policy effects differ across social categories associated with health disparities such as race, gender, or socioeconomic status. Understanding HTEs is also necessary to understand how to adapt policies to new populations because the characteristics of the new population may modify the effects of the social policy (Matthay, 2020).

Government and funders are interested in determining what interventions works best to improve health and for whom. Understanding HTEs of social policies is at the heart of these questions. HTE evaluations add cost and complexity to a study, so it is important to be able to identify the types of policies and population subgroups for which large heterogeneities are most likely. Despite their potential value, HTE evaluations are not routine in research on the health effects of social policies (Cintron et al., 2022; Fernandez y Garcia et al., 2010; Gabler et al., 2009; Glymour et al., 2013; Rojas-Saunero et al., 2022; Thomson et al., 2022). A prior review of social policy studies found that only 44% evaluated any HTEs, and of these, the population dimensions (e.g., race) examined varied widely (Cintron et al., 2022). Given limited resources and the potential for increased chance findings, additional guidance is needed on when and for which dimensions HTEs should be assessed (Breck & Wakar, 2021). Priority setting therefore requires answers to questions such as: How often does treatment effect heterogeneity happen? How often is heterogeneity trivial in magnitude? How often is it substantial in magnitude? Does the magnitude of heterogeneity and frequency of substantial heterogeneity vary by population dimension or policy type? If effects differ somewhat but are at least the same sign for everyone in the population, it may not be as important to precisely quantify heterogeneity. But if an intervention may harm some people while helping others, it is essential to understand this. Although a handful of systematic reviews explore HTEs in randomized trials of biomedical interventions (Fan et al., 2019; Fernandez y Garcia et al., 2010; Gabler et al., 2009; Kasenda et al., 2014; Starks et al., 2019; Sun et al., 2012), little work has examined HTEs of social policies and no work has examined the magnitude and distribution of HTEs in any research domain. This study takes a first step towards answering these questions.

Given the lack of systematic reviews and empirical evaluations of HTEs in practice, it is not clear if or when large HTEs of social policies are common. We address this gap by characterizing the extent of heterogeneity in estimated policy effects across population subgroups in a sample of 55 studies of the health effects of social policies. This study builds on prior work that found that less than half of the 55 studies evaluated heterogeneity in estimated policy effects across any population subgrouping dimension (Cintron et al., 2022). Here, we extend this work to characterize the findings of the studies that did evaluate HTEs. Specifically, we use meta-analyses to examine how frequently studies found heterogeneity in estimated policy effects across population subgroups, and to characterize the magnitude and distribution of heterogeneity overall and by population subgrouping dimension, policy domain, and whether the authors specified their HTE evaluations *a priori*. We also quantify how often researchers should expect subgroup effects on the opposite side of the null from the overall population effect, highlighting the potential consequences of failure to assess HTEs.

2. Materials and methods

2.1. Identification of social policy studies

We used a previously reported sample of 55 contemporary studies evaluating the health effects of social policies (Cintron et al., 2022; Matthay et al., 2022a, 2022b). The sample included all studies evaluating the health effects of social policies that were published in 2019 in a multidisciplinary set of high-impact journals: *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*. We confirmed the comprehensiveness and relevance of this set of journals using a convenience sample of 66 researchers from diverse disciplines who were asked to rank the most relevant high-impact journals publishing research on the health effects of social policies. Additional details on the sample and survey can be found elsewhere (Cintron et al., 2022; Matthay et al., 2022a, 2022b). This sample provides a snapshot of HTE evaluations in social policy research across diverse policy domains with an emphasis on high-profile publications with high methodological rigor that may influence public policy.

2.2. Data extraction and measures

We re-abstracted the studies in the original sample using a structured data extraction form (Web Table A1) to collect information on estimated policy effects across population subgroups. We classified studies as evaluating HTEs if they reported effects of the social policy on the health-related outcome(s) for subgroups of the study population defined by any subgrouping (e.g. age, gender, race/ethnicity, geography, health status). For clarity, we refer to the population characteristics along which subgroupings are evaluated (e.g., gender) as "dimensions" and the specific subgroups (e.g., men, women) as "subgroups." Dimensions defined intersectionally (e.g., race by gender) were treated as unique dimensions.

For each outcome in each study, we extracted the overall policy effect estimate and the effect estimates for all available subgroups along all available subgrouping dimensions. Studies frequently performed HTE evaluations along multiple independent dimensions. For each subgroup effect estimate, we also extracted the sample size on which the estimate is based, whether the estimate corresponds to a beneficial or harmful effect, the measure of association (e.g., incidence rate ratio, risk difference), and any quantities required to standardize the measures of association for comparability across studies. Because HTEs or estimates of effect measure modification can differ meaningfully depending on whether effects are reported on the additive versus multiplicative scale (Rothman et al., 2008), we transformed all estimated effects to both the multiplicative scale (lnOR) and additive scale (standardized mean difference, SMD) and conducted the statistical analyses on both scales (see Web Appendix B for details). Lastly, for each study, outcome, and HTE dimension, we also recorded whether the estimated policy effects indicated a narrowing or widening of disparities across subgroups as a result of the social policy (see Web Appendix B for details). Descriptive results for other data abstracted from the 55 studies can be found elsewhere (Cintron et al., 2022; Matthay et al., 2022a, 2022b).

We categorized the population characteristics used to define the HTE evaluation dimensions into demographic characteristics (e.g., age, gender, race/ethnicity), geographic location (e.g., states, cities), health characteristics (e.g., depression scores, substance use, body mass index), and socioeconomic characteristics (e.g., education, income, socioeconomic status) (Web Appendix Table A2). We categorized the social policies into the following domains: firearm (e.g., right-to-carry), immigration (e.g., deferred action for childhood arrivals), macroeconomic (e.g., austerity), employment and income (e.g., minimum wage, cash transfers), family benefits (e.g., paid family leave), population parity (e.g., 1-child/2-child), alcohol and substance use (e.g., blood alcohol concentration limits for drivers), and education policies (e.g., education system stratification) (Web AppendixTable A3). Finally, because we expected that some HTE assessments might be conducted *post-hoc* and therefore lack the study planning for sample sizes needed to ensure sufficient statistical precision, we also recorded whether authors specified their HTE evaluations *a priori* (i.e., they made their intent to

evaluate HTEs along particular dimensions known prior to reporting them).

2.3. Statistical analysis

The outcome variable in all statistical analyses was the study-outcome-subgroup-specific estimate of the effect of the social policy. For each study-outcome-dimension, we fit a random-effects meta-analysis model to characterize the heterogeneity in policy effect estimates across population subgroups defined by the corresponding dimension (Viechtbauer, 2010). For example, if a study evaluated HTEs by both race and gender for two outcomes, we fit four random-effects meta--analysis models for that study: one model for each outcome with the race dimension and one model for each outcome with the gender dimension. Since we recorded effect estimates on both the additive (SMD) and multiplicative (lnOR) scales, we fit two random-effects meta-analysis models for each unique study-outcome-dimension–one on each scale.

All models were fit using restricted maximum-likelihood estimation using the *rma* function of the *metafor* package in R version 4.1.3 (Viechtbauer, 2010). Random-effects meta-analysis models assume that the true effect estimates vary across studies (in this case, the "studies" are study-outcome-dimensions (Borenstein et al., 2021)), as opposed to fixed-effects meta-analyses which would assume that the true underlying policy effect is the same for all subgroups (Viechtbauer, 2010). The variation in observed effect estimates across subgroups may be due to real differences in policy effects across subgroups or sampling variability (i.e., chance). Random-effects meta-analysis models decompose the variation in effect estimates into these two components using the following model. For each study-outcome-dimension (Viechtbauer, 2010):

$$y_i = \theta_i + e_i \quad (1)$$

$$\theta_i = \mu + u_i \quad (2)$$

where y_i denotes the observed effect in subgroup i, θ_i corresponds to the true effect, μ is the overall average true effect across subgroups, e_i represents the sampling error of the effect estimate and is distributed $e_i \sim N(0, v_i)$, and most importantly, u_i represents the true subgroup-specific deviation from the true overall effect μ and is distributed $u_i \sim N(0, \tau^2)$. The v_i represent the variances in the sampling errors for each subgroup (i.e., the variances of the estimated effects for each subgroup) and, because many studies did not report these variances, were approximated using $v_i = \frac{1}{\sqrt{(var(E)*(n_i-1))}}$ here E is the policy exposure variable (binary or continuous) and n_i is the corresponding subgroup analytic sample size (Viechtbauer, 2010) (see Web Appendix B for details and justification).

The primary parameter of interest was τ, the standard deviation of the true subgroup-specific effects about the true overall effect, which quantifies the degree of heterogeneity in the effect of the social policy across subgroups after accounting for variability due to chance. τ was on the same scale as the policy effect estimates (SMD or lnOR), and $\tau = 0$ indicated that there were no differences in the true effects of the policy across population subgroups.

We used frequency statistics and histograms to characterize the distribution of τ estimates overall and by dimension, social policy domain, and *a priori* specification of HTE analyses. We did not calculate inferential statistics to evaluate whether the distributions of τ differed by dimension or study characteristics because of the small number of effect estimates in each category. Given that the effects of social policies on health outcomes are likely to be small (i.e., <0.2 SMD) (Matthay et al., 2021), we considered $\tau \geq 0.1$ SMDs (or equivalently $\tau \geq 0.18$ on the lnOR scale) (Cohen, 2013) to be "large" heterogeneity.

Special ethical considerations arise in the context of qualitative interaction, i.e., when a policy benefits some subgroups but harms others. To assess whether such qualitative interactions were likely, we also used the estimated parameters from the meta-analysis models to quantify the proportion of the time we would expect a subgroup effect on the opposite of the null from the overall population effect. Specifically, for each study-outcome-dimension, we computed the area under the curve distributed $N(\mu, \tau^2)$ that was on the opposite side of the null from the overall population effect. This phenomenon indicates when failure to evaluate HTEs could lead to policy decisions that inadvertently harm some subgroups.

Lastly, for each meta-analysis model, we examined I^2: the estimated percentage of variability in the effect estimates due to real between-subgroup heterogeneity rather than chance. $I^2 = 0\%$ indicated that variation in effect estimates was entirely due to chance.

A complete overview of the steps in this study is presented in Fig. 1.

3. Results

3.1. Characteristics of the study sample

Of the 55 studies, 24 evaluated some form of HTEs. Studies assessed a range of health outcomes (e.g., infant mortality, self-rated health, firearm suicides) and social policies. After data extraction, the database included 557 subgroup effect estimates from 159 unique study-outcome-dimensions. Of these, 24 estimates (4%) were excluded because the quantities needed to transform the measure of association to the lnOR or SMD were not reported and could not be approximated. Another 16 estimates (2%) were excluded due to missing information needed to compute the variance of the subgroup-specific effect estimate. Finally, 7 estimates (1%) were excluded because they were not reported in the original study due to model non-convergence. We treated these estimates as missing completely at random. The final analytic database included 510 subgroup effect estimates from 136 unique study-outcome-dimensions (Table 1).

3.2. Benefits and harms of social policies and implications for disparities

Standardized policy effect estimates for all studies, outcomes, and subgroups are presented in Fig. 2. Social policy effect estimates were generally small (<0.1 SMDs). Effects on health were mixed: 342 (67%) estimates implied health benefits and 168 (33%) implied health harms. Across the 136 study-outcome-dimensions, the estimated policy effects for 68 (50%) implied a widening of disparities in the outcome between subgroups. Cross-tabulating this information with the direction of effect (harmful versus beneficial), 14% of effect estimates corresponded to harmful effects on average that nonetheless reduced the magnitude of disparities in the outcome across subgroups, 36% corresponded to beneficial effects on average that reduced disparities, 18% corresponded to harmful effects on average that widened disparities, and 31% corresponded to beneficial effects on average that widened disparities.

3.3. Overall distribution of heterogeneity in social policy effect estimates

The effects of social policies frequently varied by population subgroup. Fig. 3 presents the distribution of estimated τ values. Across study-outcome-dimensions, the median τ (degree of heterogeneity) on the SMD scale was 0.03 (range: 0.0–0.9), 54% of τ estimates had 95% confidence intervals that excluded 0, indicating statistically significant evidence of heterogeneity, and 20% of τ estimates were greater than our benchmark for large heterogeneity of 0.1 (Table 1). On the multiplicative scale, the distribution of τ estimates was similar in pattern (Appendix Figures A4,Table A5), but larger in magnitude compared to the additive scale. 47 (35%) study-outcome-dimensions had estimated I^2 values of 0, indicating that for these dimensions, the magnitude of any apparent variation in effects across subgroups was within that expected due to chance in finite samples (Appendix Figures A8-9).

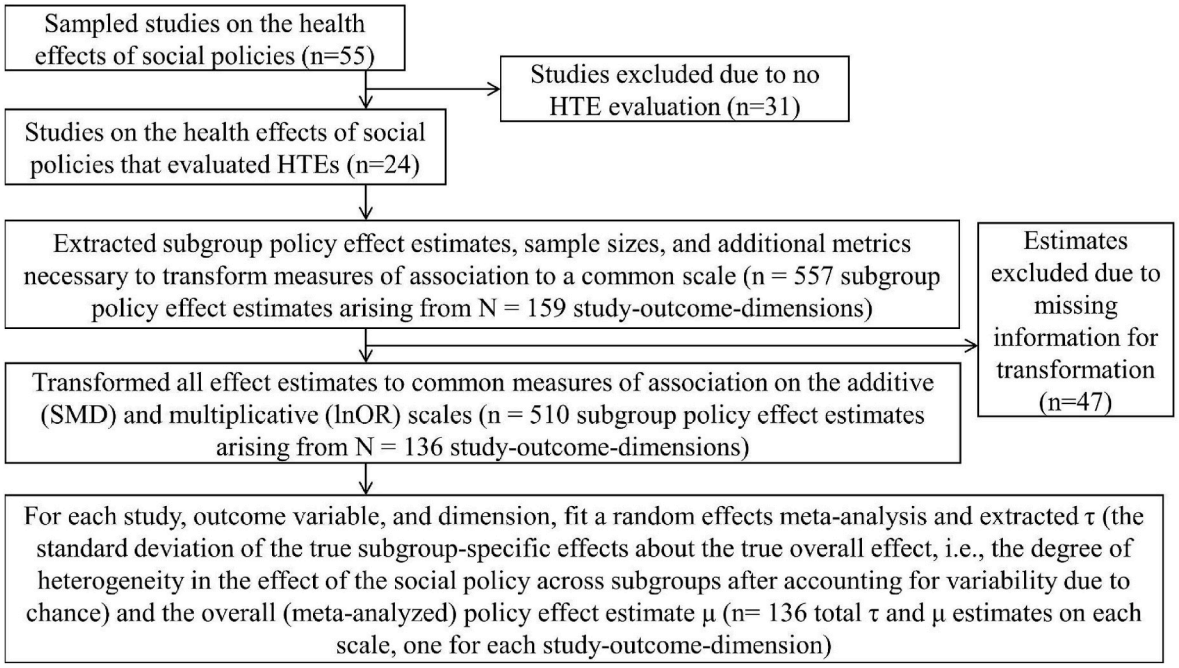

Fig. 1. Data extraction and analysis flow chart.

Table 1
Sample characteristics and estimated heterogeneity of effects within study-outcome-dimensions for studies on the health effects of social policies.

	# Study-Outcome-Dimensions	# of estimates	Median τ (range)	%τ > 0.1	% τ CIs excluding null
Overall	136	510	0.03 (0.0–0.9)	20	54
Social policy domain					
Firearm	11	76	0.00 (0.0–0.0)	0	0
Immigration	2	4	0.09 (0.0–0.2)	50	50
Macroeconomic	2	4	0.00 (0.0–0.0)	0	0
Employment and income	41	114	0.05 (0.0–0.4)	29	71
Family benefits	52	221	0.03 (0.0–0.9)	13	62
Population parity	8	16	0.06 (0.0–0.2)	25	75
Alcohol and substance use	18	62	0.03 (0.0–0.3)	28	28
Education	2	13	0.00 (0.0–0.0)	0	0
Population characteristic					
Demographic characteristics	71	210	0.02 (0.0–0.9)	18	48
Geographic location	22	162	0.02 (0.0–0.4)	18	41
Health characteristics	6	32	0.04 (0.0–0.4)	33	50
Socioeconomic characteristics	37	106	0.04 (0.0–0.8)	22	73
A priori specification					
Yes	99	422	0.02 (0.0–0.4)	11	51
No	37	88	0.08 (0.0–0.9)	43	54

Note. Tau (τ) is the standard deviation of the effect estimates across the given study-outcome-dimension after accounting for sampling variability. CIs - 95% Confidence intervals.

3.4. Distribution of heterogeneity social policy effect estimates by social policy domain

Most HTE evaluations involved employment and income ($n = 41$) or family benefits ($n = 52$) policies (Table 1). Heterogeneity in estimated policy effects was evident for most types of social policies (Table 1, Appendix Figure A1, AppendixTable A5, Appendix Figure A5). Specifically, τ was greater than the 0.1 benchmark for 29% of employment and income, 28% of alcohol and substance use, 25% of population parity, 13% of family benefits, and 50% of immigration study-outcome-dimensions, but 0% of firearm, macroeconomic, or education study-outcome-dimensions.

3.5. Distribution of heterogeneity in social policy effect estimates by population characteristics

Demographic characteristics ($n = 71$), geographic location ($n = 22$), and socioeconomic characteristics ($n = 37$) were the most common population characteristics for which HTEs were evaluated (Table 1). Heterogeneity in estimated policy effects was evident for all population characteristics (Table 1, Appendix Figure A2, AppendixTable A5, Appendix Figures A6). Specifically, τ was greater than the 0.1 benchmark for 18% of demographic characteristics, 18% of geographic location, 33% of individual health characteristics, and 22% of socioeconomic characteristics analyses.

3.6. Distribution of heterogeneity in social policy effect estimates by a priori specification of HTEs

HTE evaluations were specified *a priori* for 73% of study-outcome-dimensions. Heterogeneity in estimated effects was less common in studies with *a priori* HTE specification (11% with $\tau > 0.1$) than in studies that did not specify *a priori* a plan to evaluate HTEs (43% with $\tau > 0.1$) (Table 1, Appendix Figure A3, AppendixTable A5; Appendix Figures A7).

3.7. Frequency of qualitative interaction

Given the estimated variance of effect sizes across subgroups, effects

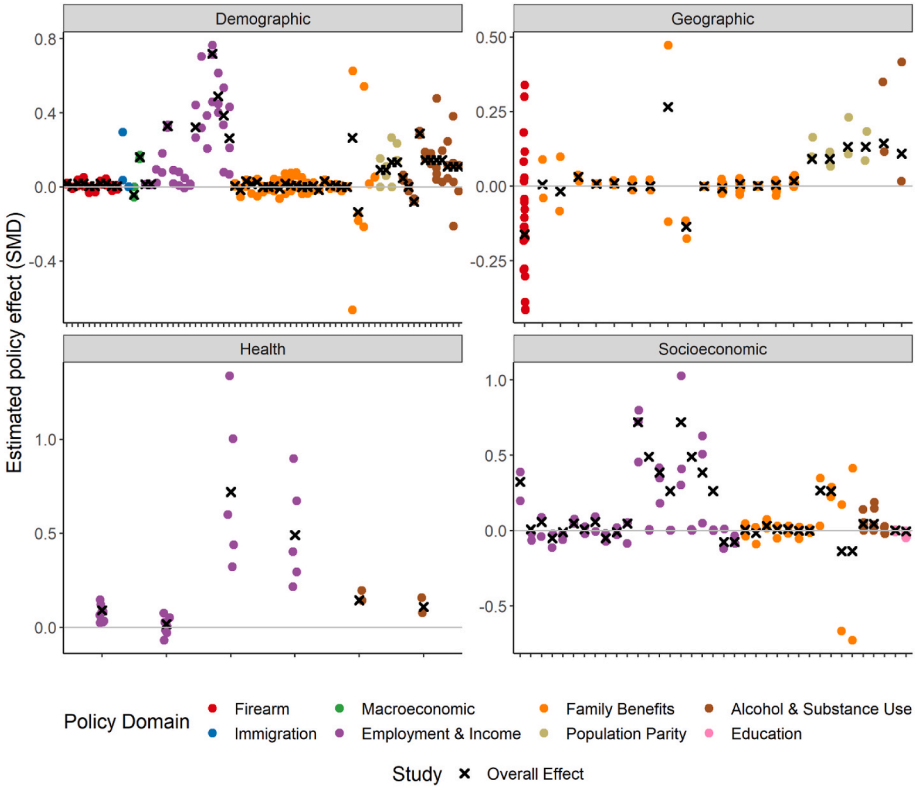

Fig. 2. Distribution of estimated subgroup social policy effects by population characteristic and social policy domain

Note. The overall effect is the overall policy effect reported for a given study population. In certain studies, the overall policy effect is not presented because it was not reported in the original study or the subgroup analyses were the primary focus of the study. X-axis tick marks represent a specific study-outcome-dimension. A grey line is placed at a null effect (zero). Positive values indicate beneficial effects whereas negative values indicate harmful effects. Estimates for each study-outcome-dimension are jittered for clarity. SeeTable A2 and A3 for more information on population dimensions and social policy domains.

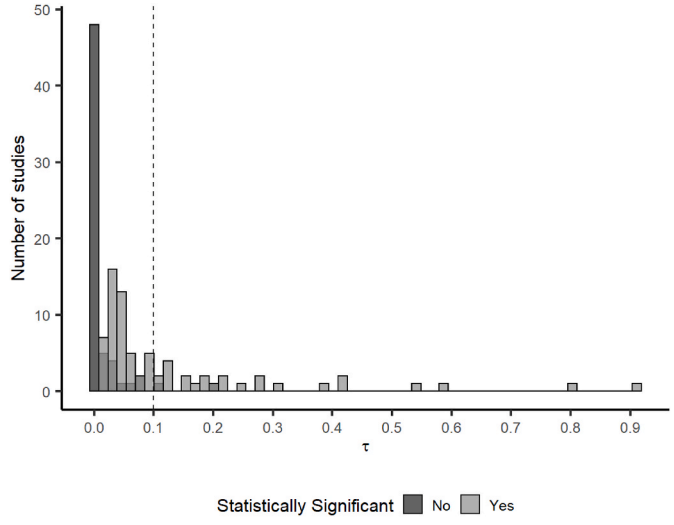

Fig. 3. Distribution of standard deviations (heterogeneity) in standardized mean difference estimates of social policy effects (τ) across study-outcome-dimensions

Note. Tau *(τ)* is the standard deviation of the effect estimates across study-outcome-dimensions after accounting for sampling variability. The vertical dashed line represents our benchmark for considerable heterogeneity (i.e., $\tau = 0.1$). Two sets of histograms are overlaid and shaded by the statistical significance of the τ's, where statistical significance refers to a 95% confidence interval for the estimated τ that excluded the null.

in opposite directions are likely to be common. Of the 136 study-outcome-dimensions, 104 (76%) had a non-zero area under the curve on the opposite side of the null from the overall population effect. Of these 104, 44% corresponded to an expected effect on the opposite side of the null from the overall population effect at least 25% of the time; 26% corresponded to an expected effect opposite to the overall population effect at least 50% of the time; and 16% corresponded to an expected effect to the overall population effect at least 75% of the time.

4. Discussion

We characterized the frequency, magnitude, and distribution of HTEs in a contemporary sample of studies on the health effects of social policies. Less than half of studies (44%) evaluated heterogeneity in estimated policy effects across any population subgrouping dimension. Across reported HTE evaluations (study-outcome-dimensions), 54% indicated statistically significant differences in policy effects across population subgroups and 20% reported large heterogeneities. With some variation in frequency and magnitude, HTEs were observed for most social policy domains, all types of population characteristics, and regardless of whether the HTE evaluation was specified *a priori*. These findings underscore the importance of evaluating HTEs of social policies. HTEs are important for both understanding the implications of the social policy for health disparities and anticipating how population-level social policy effects will differ in jurisdictions with different compositions from the one initially studied.

Evidence of considerable heterogeneity in social policy effects across population subgroups is consistent with social theory. For example, resource substitution theory hypothesizes that those who have been historically denied health promoting resources (e.g., education, income, and power) will benefit more from access to these resources compared to those who more readily receive them (Ross & Mirowsky, 2006). Numerous prior policy evaluations have reported important HTEs for at least some subgrouping dimensions (Leventhal & Brooks-Gunn, 2003; Nguyen et al., 2016; Vable et al., 2016), but to our knowledge, this is the first study to systematically assess the frequency, magnitude, and distribution of HTEs across the social policy literature. It is also the first to apply meta-analysis methodology as a tool for characterizing heterogeneity and enabling discussions of health equity impacts.

Only 35% of study-outcome-dimensions we evaluated had $I^2 = 0$, meaning that the differences in estimated policy effects across subgroups

could not generally be explained by chance alone. Yet, the estimates reported here likely provide a lower bound on the frequency of heterogeneity in social policy effects for two reasons. First, HTE evaluations are frequently underpowered. An apparent lack of heterogeneity may simply reflect insufficient sample size to derive precise effect estimates for each subgroup. We did not observe heterogeneity in the study-outcome-dimensions involving firearm, macroeconomic, or education policies, but this does not mean that these policies do not have heterogeneous effects; these findings may simply reflect insufficient sample sizes to derive precise effect estimates for each subgroup. Second, HTEs are not routinely assessed along all potentially relevant dimensions (Cintron et al., 2022). This study thus adds to accumulating evidence that substantial HTEs are common and should be routinely and systematically reported when evaluating the health effects of social policies (Cintron et al., 2022; Matthay & Glymour, 2022; Petticrew et al., 2012). It also highlights the importance of study planning to ensure evaluations of social policies are sufficiently precise to evaluate HTEs.

The importance of evaluating HTEs is especially evident given our finding of subgroup effects in the opposite direction from the overall population effect (e.g., 76% of the study-outcome-dimensions had a non-zero area under the curve on the opposite side of the null). Studies lacking HTE assessments may therefore lead to policy recommendations that inadvertently harm some groups, or conversely, lead to missed opportunities for some subgroups to benefit. This study provides a methodological framework for using meta-analysis to identify these patterns. Our findings suggest that prior studies that did not examine HTEs might well be revisited to check for differential effects across important subgroups. HTE evaluations also show how the directions of average health effects (benefit versus harm) intersect with implications for disparities: 36% of HTE evaluations in this study corresponded to beneficial average effects that also reduced the magnitude of disparities in the outcome across subgroups, but 14% reduced disparities despite on-average health-harming effects; and 31% benefitted health on-average but widened disparities. To enable informed policy discussions of these tradeoffs, adequate quantitative HTE assessments across all relevant subgroups are required.

Funders, policymakers, and researchers have limited resources, and powering studies to evaluate HTEs entails added cost and complexity. Thus, resources should be dedicated to evaluating heterogeneity only if meaningful heterogeneity is likely and relevant for the policy impact. Studies such as ours can help inform this priority setting, but a larger evidence base is needed to draw firm conclusions. For instance, in our sample substantial heterogeneity in effect estimates was common across subgroups defined by health characteristics and socioeconomic characteristics and somewhat less common for geographic and demographic characteristics. Ideally a comprehensive analysis and theoretical guidance could direct prioritization of both the dimensions of heterogeneity–including intersectionally defined groups–and policy domains most likely to have heterogeneous effects.

The variation in policy types, study contexts, study designs, and outcome measures across studies included in our analysis implies that there are many reasons that the magnitude of HTEs may differ across studies. Delving further into the reasons for differences in estimated HTEs for subsets of studies that are more homogeneous is an area for future research. We conducted the random effects meta-analyses at the level of the study-outcome-dimension, so our analytic approach makes no assumptions about the level of similarity or difference in the estimated subgroup treatment effects across studies. We view this analytic flexibility and the diversity of policy domains and study contexts as a strength because, to our knowledge, no research has quantified the magnitude and distribution of HTEs using our meta-analysis analytic approach across any social policy domains. Because of the paucity of research in this area, our study is an important first step towards quantifying the full range of HTEs across diverse domains, so that subsequent research can further investigate individual domains and reasons for differences across domains.

No consensus on when or how to evaluate HTEs in social policy and health research has been established. For example, how should researchers balance the importance of identifying meaningful heterogeneity against the increased risk of spurious findings as the number of subgroups grows (Heckman et al., 2010)? Among the many diverse methods for evaluating HTEs (Breck & Wakar, 2021), which approaches are most appropriate for social policies? Which perform best and under what conditions (e.g., for few versus many subgroups)? Which available methods (e.g., MAIHDA, probability samples, qualitative tools) are best-suited to small sample sizes (Evans et al., 2018; Harding and Seefeldt, 2013; Tipton et al., 2019)? When is pre-specification or pre-registration necessary? How should HTEs be reported? Several recent articles on performing, reporting, and assessing the credibility of subgroup analyses may help in this effort (Gil-Sierra et al., 2020; Lesko et al., 2018; Schandelmaier et al., 2019, 2020; Sun et al., 2010; Tipton et al., 2019; Varadhan et al., 2013). We found that substantial heterogeneity was more common (43% vs 11%) when HTE evaluations were not specified *a priori*. This finding is consistent with research on the replicability crisis that suggests increases in spurious findings due to multiple hypothesis testing (Austin et al., 2006; Berger et al., 2009; Gelman & Loken, 2014). Research indicates methods for evaluating HTEs vary considerably across disciplines and that these differences may lead to differing conclusions (Breck & Wakar, 2021; Inglis et al., 2018; Loh et al., 2019). Work to develop guidelines for the conduct and reporting of HTEs specific to studies of the health effects of social policies is needed, especially given the unique implications of these studies for public policy and health equity.

Many potentially relevant subgrouping dimensions were not evaluated in the studies reviewed here. For example, health insurance status and disability status were not considered in any of the studies in our sample, yet important heterogeneity along these dimensions may exist. Differential impacts of social policies on racial/ethnic subgroups are of particular interest, but there were insufficient studies including consistent definitions of racial/ethnic categories to examine results for this dimension separately; this is a priority for future work. Furthermore, no subgroups were explicitly defined using intersectionality theory (Crenshaw, 1989); almost all studies treated subgrouping dimensions as independent and mutually exclusive (only 12 of the study-outcome-dimensions considered cross-classification of multiple demographic characteristics, e.g., age by gender). Future HTE evaluations must consider subgroups based on intersectionality theory (e.g., race by gender) to illuminate how a person's multiple identities and social positions might be embedded within systems of inequality. Guidance is needed on how to determine which dimensions are relevant and should be evaluated based on theoretical and/or statistical principles (Boyd et al., 2020).

4.1. Limitations

We excluded 8% of study-outcome-subgroup-specific estimates due to incomplete reporting of HTEs in the original research contributing to our meta-analyses. The small number of studies and subgroups in our analysis also limited our ability to make claims about differences in the frequency or magnitude of heterogeneity in social policy effects by specific population characteristics or social policy domains. Furthermore, the subgroups examined were quite variable across studies and we lacked consistent observation of the same subgrouping dimensions (e.g. race/ethnicity) many times across different studies.

Also, note that the estimated HTEs may not be the true, unbiased HTEs. In this paper, we treat estimates as the investigators' best attempt to estimate the causal effect of the social policy on the population subgroups, but we acknowledge that all estimates likely depart from the true causal effect to some degree. Further assessment of the methodological quality of studies evaluating HTEs is necessary; this work is best-done within subsets of studies that are more homogeneous with respect to policy type, study context, study design, and outcome measures. We

view this paper as a first step towards these goals. Finally, some assumptions and approximations were necessary to convert reported measures of association to a common scale and to include consistent and complete variance estimates for all effect estimates.

4.2. Conclusions

This is the first study to systematically advance our understanding of the frequency, magnitude, and distribution of HTEs in research on the health effects of social policies at scale. We found that social policies can have considerably different health effects on subgroups and that the frequency and magnitude of heterogeneity varied by social policy domain, subgrouping dimensions, and *a priori* HTE specification. While this study does not provide recommendations on specific policy domains or population subgroups for which evaluating HTEs is a priority, it provides a novel methodological framework for quantifying HTEs and lays the groundwork for future investigations. Researchers and policymakers should be aware that social policies may have differential impacts across population subgroups in ways that can either exacerbate or mitigate health disparities. Consistently conducting HTE evaluations across all relevant population subgrouping dimensions is essential for adequate evidence-based policymaking that promotes health equity. This includes assessing HTE in future policy studies as well as revisiting prior social policy studies that did not examine HTEs. Yet in social policy research, HTE evaluations remain rare, methods are not standardized, and there is no available guidance on best practices. Given the small number of studies and diversity of subgrouping dimensions in this review, more research on a larger sample of the literature is needed to definitively characterize the policies and population subgrouping dimensions that are most important to evaluate.

Sources of financial support

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

Data access

Data and computing code may be made available upon reasonable request.

CRediT author statement

Dakota W. Cintron: Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, 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.

Declaration of competing interest

There are no conflicts of interest to report.

Data availability

Data will be made available upon reasonable request.

Acknowledgements:

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ssmph.2023.101352.

References

Austin, P. C., Mamdani, M. M., Juurlink, D. N., & Hux, J. E. (2006). Testing multiple statistical hypotheses resulted in spurious associations: A study of astrological signs and health. *Journal of Clinical Epidemiology, 59*(9), 964–969. https://doi.org/10.1016/j.jclinepi.2006.01.012

Berger, M. L., Mamdani, M., Atkins, D., & Johnson, M. L. (2009). Good research practices for comparative effectiveness research: Defining, reporting and interpreting nonrandomized studies of treatment effects using secondary data sources: The ISPOR good research practices for retrospective database analysis task force report—Part I. *Value in Health, 12*(8), 1044–1052.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Boyd, R. W., Lindo, E. G., Weeks, L. D., & McLemore, M. R. (2020). On racism: A new standard for publishing on racial health inequities. *Health Aff Blog, 10*(10), 1, 1377.

Boykin, C. M., Brown, N. D., Carter, J. T., et al. (2020). Anti-racist actions and accountability: Not more empty promises. *Equal Divers Incl Int J, 39*(7), 775–786. https://doi.org/10.1108/EDI-06-2020-0158

Breck, A., & Wakar, B. (2021). Methods, challenges, and best practices for conducting subgroup analysis. *OPRE Rep*, 17.

Cintron, D. W., Adler, N. E., Gottlieb, L. M., et al. (2022). Heterogeneous treatment effects in social policy studies: An assessment of contemporary articles in the health and social sciences. *Annals of Epidemiology, 70*, 79–88. https://doi.org/10.1016/j.annepidem.2022.04.009

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. *University of Chicago Legal Forum*, (1), 139–167. *Chic IL*. Published online 1989.

Evans, C. R., Williams, D. R., Onnela, J. P., & Subramanian, S. V. (2018). A multilevel approach to modeling health inequalities at the intersection of multiple social identities. *Social Science & Medicine, 203*, 64–73. https://doi.org/10.1016/j.socscimed.2017.11.011

Fan, J., Song, F., & Bachmann, M. O. (2019). Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. *Journal of Clinical Epidemiology, 108*, 17–25. https://doi.org/10.1016/j.jclinepi.2018.12.009

Fernandez y Garcia, E., Nguyen, H., Duan, N., Gabler, N. B., & Kravitz, R. L. (2010). Assessing heterogeneity of treatment effects: Are authors misinterpreting their results? *Health Services Research, 45*(1), 283–301. https://doi.org/10.1111/j.1475-6773.2009.01064.x

Gabler, N. B., Duan, N., Liao, D., Elmore, J. G., Ganiats, T. G., & Kravitz, R. L. (2009). Dealing with heterogeneity of treatment effects: Is the literature up to the challenge? *Trials, 10*(1), 43. https://doi.org/10.1186/1745-6215-10-43

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist, 102*(6), 460.

Gil-Sierra, M. D., Fénix-Caballero, S., Abdel kader-Martin, L., et al. (2020). Checklist for clinical applicability of subgroup analysis. *Journal of Clinical Pharmacy and Therapeutics, 45*(3), 530–538. https://doi.org/10.1111/jcpt.13102

Glymour, M. M., Osypuk, T. L., & Rehkopf, D. H. (2013). Invited commentary: Off-roading with social epidemiology—exploration, causation, translation. *American Journal of Epidemiology, 178*(6), 858–863. https://doi.org/10.1093/aje/kwt145

Harding, D. J., & Seefeldt, K. S. (2013). Mixed methods and causal analysis. In *Handbook of causal analysis for social research* (pp. 91–110). Springer.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope perry preschool program. *Quant Econ, 1*(1), 1–46. https://doi.org/10.3982/QE8

Inglis, G., Archibald, D., Doi, L., et al. (2018). Credibility of subgroup analyses by socioeconomic status in public health intervention evaluations: An underappreciated problem? *SSM - Popul Health, 6*, 245–251. https://doi.org/10.1016/j.ssmph.2018.09.010

Kasenda, B., Schandelmaier, S., Sun, X., et al. (2014). Subgroup analyses in randomised controlled trials: Cohort study on trial protocols and journal publications. *BMJ, 349*.

Kendi, I. X. (2019). *How to Be an antiracist*. Random House Publishing Group.

Lesko, C. R., Henderson, N. C., & Varadhan, R. (2018). Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology, 100*, 22–31. https://doi.org/10.1016/j.jclinepi.2018.04.005

Leventhal, T., & Brooks-Gunn, J. (2003). Moving to opportunity: An experimental study of neighborhood effects on mental health. *American Journal of Public Health, 93*(9), 1576–1582. https://doi.org/10.2105/AJPH.93.9.1576

Loh, W. Y., Cao, L., & Zhou, P. (2019). Subgroup identification for precision medicine: A comparative review of 13 methods. *WIREs Data Min Knowl Discov, 9*(5), e1326. https://doi.org/10.1002/widm.1326

Matthay, E. C. (2020). Do social interventions have different health effects for different people? *Evidence for Action Methods Notes*. https://www.evidenceforaction.org/sites/default/files/2021-04/E4A-Methods-Note-HTEp1.pdf.

Matthay, E. C., & Glymour, M. M. (2022). Causal inference challenges and new directions for epidemiologic research on the health effects of social policies. *Curr Epidemiol Rep, 9*(1), 22–37. https://doi.org/10.1007/s40471-022-00288-7

Matthay, E. C., Gottlieb, L. M., Rehkopf, D., Tan, M. L., Vlahov, D., & Glymour, M. M. (2022a). What to do when everything happens at once: Analytic approaches to estimate the health effects of Co-occurring social policies. *Epidemiologic Reviews, 43*(1), 33–47. https://doi.org/10.1093/epirev/mxab005

Matthay, E. C., Hagan, E., Gottlieb, L. M., et al. (2021). Powering population health research: Considerations for plausible and actionable effect sizes. *SSM - Popul Health., 14*, Article 100789. https://doi.org/10.1016/j.ssmph.2021.100789

Matthay, E. C., Hagan, E., Joshi, S., et al. (2022b). The revolution will Be hard to evaluate: How Co-occurring policy changes affect research on the health effects of social policies. *Epidemiologic Reviews, 43*(1), 19–32. https://doi.org/10.1093/epirev/mxab009

Nguyen, Q. C., Rehkopf, D. H., Schmidt, N. M., & Osypuk, T. L. (2016). Heterogeneous effects of housing vouchers on the mental health of US adolescents. *American Journal of Public Health, 106*(4), 755–762. https://doi.org/10.2105/AJPH.2015.303006

Petticrew, M., Tugwell, P., Kristjansson, E., Oliver, S., Ueffing, E., & Welch, V. (2012). Damned if you do, damned if you don't: Subgroup analysis and equity. *Journal of Epidemiology & Community Health, 66*(1), 95–98.

Rojas-Saunero, L. P., Labrecque, J. A., & Swanson, S. A. (2022). Invited commentary: Conducting and emulating trials to study effects of social interventions. *American Journal of Epidemiology, 191*(8), 1453–1456. https://doi.org/10.1093/aje/kwac066

Ross, C. E., & Mirowsky, J. (2006). Sex differences in the effect of education on depression: Resource multiplication or resource substitution? *Social Science & Medicine, 63*(5), 1400–1413. https://doi.org/10.1016/j.socscimed.2006.03.013

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). In *others. Modern Epidemiology* (Vol. 3). Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins.

Schandelmaier, S., Briel, M., Varadhan, R., et al. (2020). Development of the instrument to assess the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses. *Canadian Medical Association Journal, 192*(32), E901–E906.

Schandelmaier, S., Chang, Y., Devasenapathy, N., et al. (2019). A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. *Journal of Clinical Epidemiology, 113*, 159–167. https://doi.org/10.1016/j.jclinepi.2019.05.014

Starks, M. A., Sanders, G. D., Coeytaux, R. R., et al. (2019). Assessing heterogeneity of treatment effect analyses in health-related cluster randomized trials: A systematic review. *PLoS One, 14*(8), Article e0219894. https://doi.org/10.1371/journal.pone.0219894

Sun, X., Briel, M., Busse, J. W., et al. (2012). Credibility of claims of subgroup effects in randomised controlled trials: Systematic review. *BMJ, 344*.

Sun, X., Briel, M., Walter, S. D., & Guyatt, G. H. (2010). Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ*, 340.

Thomson, R. M., Igelström, E., Purba, A. K., et al. (2022). How do income changes impact on mental health and wellbeing for working-age adults? A systematic review and meta-analysis. *The Lancet Public Health, 7*(6), e515–e528. https://doi.org/10.1016/S2468-2667(22)00058-5

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. In *Experimental methods in survey research* (pp. 435–456). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119083771.ch22.

Vable, A. M., Canning, D., Glymour, M. M., Kawachi, I., Jimenez, M. P., & Subramanian, S. V. (2016). Can social policy influence socioeconomic disparities? Korean war GI Bill eligibility and markers of depression. *Annals of Epidemiology, 26*(2), 129–135.e3. https://doi.org/10.1016/j.annepidem.2015.12.003

Varadhan, R., Segal, J. B., Boyd, C. M., Wu, A. W., & Weiss, C. O. (2013). A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology, 66*(8), 818–825. https://doi.org/10.1016/j.jclinepi.2013.02.009

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48.