

A practical tutorial on mediation analysis with clustered data in health psychology

Xiaoai Wen

University of Colorado Denver <https://orcid.org/0009-0002-6614-8473>

Chun-Chieh Hu

chun-chieh.hu@ucdenver.edu

University of Colorado Denver <https://orcid.org/0009-0009-7337-885X>

Method Article

Keywords: health psychology, mediation analysis, psychosocial pathways, clustered data, bootstrap, multiple comparisons, reproducibility, suicide-related vulnerability

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11370584/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A practical tutorial on mediation analysis with clustered data in health psychology

Xiaoai Wen^{1,2,3}, Chun-Chieh Hu^{1*}

¹ Department of Economics, University of Colorado Denver, Denver, CO 80204, USA.

² International College Beijing, China Agricultural University, Beijing 100083, China.

³ Kenneth C. Griffin Department of Economics, MACSS – Economics, University of Chicago, Chicago, IL, USA

*Corresponding author: Chun-Chieh Hu, email: Chun-Chieh.Hu@ucdenver.edu

Abstract

Health psychology examines the pathways linking interventions and social conditions to psychological outcomes. This tutorial provides a reproducible workflow for evaluating candidate pathways when observations are clustered. Such settings arise when a psychological mechanism is hypothesized and participants are nested within communities, schools, clinics, or intervention groups. Using five waves (2011–2020) of the China Health and Retirement Longitudinal Study across 94 cities, the Healthy City example links a macro-level environmental perspective with a micro-level focus on social participation. Six candidate indicators are examined, with Green and Social as worked examples. The “could not get going” item is interpreted as an anergia-based, theoretically informed, non-validated proxy for suicide-related vulnerability. The workflow combines three-step path regressions, estimation of $a \times b$, supplementary Sobel testing, 2,000 city-level bootstrap replications with percentile and bias-corrected (BC) intervals, and Benjamini–Hochberg (BH) adjustment across six candidates. Common samples within each candidate keep the path estimates comparable, while city clustering follows the level of policy exposure. Under city resampling, Green’s two unadjusted intervals include zero; Social’s BC interval excludes zero and its percentile interval includes zero. BH adjustment raises Social’s Sobel p value from .04505 to .27029. Bootstrap sign-tail summaries receive a separate descriptive BH adjustment, without a claim of validated false discovery rate control. The accompanying Stata and Python code connects these decisions to their calculations and outputs. Researchers can adapt the workflow to health psychology studies in community, school, or care settings, selecting measures and inference methods that match the proposed psychological process and study design.

Keywords: health psychology; mediation analysis; psychosocial pathways; clustered data; bootstrap; multiple comparisons; reproducibility; suicide-related vulnerability

1 Introduction

Health psychology seeks to explain how interventions and social conditions influence psychological well-being. Mediation analysis helps turn these explanatory questions into hypotheses about the processes linking an exposure to an outcome [1]. With clustered observations, the proposed pathway, its measured indicator, and the unit that supplies independent information must fit the same research question. This tutorial makes those decisions explicit through a reproducible analysis of the Healthy City study.

China's Healthy City Pilot Policy, launched in 2016, seeks to improve health through changes in the natural environment, community life, and health services [2]. This setting connects a policy implemented across cities with residents' everyday environments and social experiences. The health psychology question is whether these changes are associated with psychological vulnerability, and which candidate pathways could explain that association. We examine this question using an anergia-based proxy score and indicators of city conditions and individual participation. The methodological task is to connect the proposed pathways to their measurements, analysis samples, and uncertainty estimates.

Two theoretical perspectives guide the candidate pathways. Stokols's social ecological framework links health to the interaction between people and their physical and social surroundings [3]. It motivates a macro-level question: could changes in the urban environment help explain the policy–outcome association? Green infrastructure is relevant to this question because restorative surroundings may support recovery from stress and mental fatigue [4]. This provides a psychological rationale for examining green coverage alongside an individual symptom-based outcome. In a cluster randomized trial, South et al. found that greening vacant urban land reduced reported feelings of depression and worthlessness [5].

At the micro level, Klonsky and May's Three-Step Theory of suicide emphasizes pain and hopelessness in the development of suicidal ideation, and connectedness as a protection against its escalation [6]. This perspective motivates a candidate social pathway: opportunities for contact may help sustain engagement and a sense of connection in everyday life. Social participation provides an observable behavioral indicator for examining this possibility. Lin et al., using longitudinal Chinese data, linked participation in activities such as meeting friends and playing mahjong, chess, or cards to later depressive-symptom patterns [7]. These environmental and social perspectives motivate candidate pathways that can be examined through observed indicators.

This tutorial uses the China Health and Retirement Longitudinal Study (CHARLS) [8], drawing on five waves from 2011 to 2020 to compare urban adults aged 45 years and older in treated and untreated cities. The outcome is the “could not get going” item from the Center for Epidemiologic Studies Depression Scale (CES-D) [9]. Validation of the ten-item CES-D in older Chinese adults concerns the scale as a whole [10]. We interpret the item as an anergia-related indication of reduced everyday activation and use it as a theoretically informed, non-validated proxy for suicide-related vulnerability. The proposed psychological link is that reduced activation may accompany depressive distress and withdrawal from everyday engagement. The Three-Step Theory links the development of suicidal ideation to pain and hopelessness [6]; longitudinal evidence also links hopelessness to later ideation, while distinguishing ideation from progression

to attempts [11]. These findings provide a theoretical context for the exploratory proxy interpretation. Validation of this single item against suicidal ideation or behavior remains a separate measurement question.

The analysis therefore combines a macro-level view of the conditions surrounding psychological functioning with a micro-level view of residents' social engagement. City green coverage (Green) and individual social participation (Social) are the main examples. Four further candidates are retained: road surface area per person (Road), industrial sulfur-dioxide emissions (SO₂), annual mean fine particulate matter concentration (PM_{2.5}), and hospitals and health centers per population (Medical). These city measures represent infrastructure, pollution, and health-service supply, respectively, and each is interpreted according to its own substantive meaning.

The three-step regression approach provides a starting point for examining these candidate pathways [12,13]. The first regression estimates the association between policy exposure and the proxy score. The second estimates the policy–mediator association, a . The third includes both policy exposure and the mediator to estimate the mediator–outcome association, b , and the remaining policy–outcome association. For Green, these steps connect the overall policy comparison to changes in city green coverage and their association with the proxy score. The product ab is calculated from the two path coefficients estimated on the same observations within each candidate pathway. Each product is assessed directly, irrespective of whether the overall policy association reaches a significance threshold [14]. Because mediators and outcomes are paired within the same survey wave, the product is interpreted here as an indirect association. A causal interpretation additionally requires appropriate temporal ordering and assumptions about confounding [15,16].

The policy and the data also determine how uncertainty should be calculated. Repeated responses from one person may be related, and different people in the same city share policy exposure and local conditions. Every respondent interviewed in a city and year receives the same Green value; interviewing more residents does not create more independent changes in city greening. Individual clustering accounts for repeated observations from each person. City clustering also allows for connections between different residents of the same city [17,18]. This distinction matters even when both analyses produce the same regression coefficients.

After estimating the paths, we assess the uncertainty in their product and consider the fact that six candidates are examined. The Sobel test provides a supplementary test of the product ab [19]. The city bootstrap repeatedly draws cities and refits the regressions, producing percentile and bias-corrected (BC) confidence intervals. Both intervals are reported because they can give different judgments about whether zero is included [20]. The Benjamini–Hochberg (BH) procedure then adjusts sets of results for multiple comparisons [21]. These steps address different parts of the same analysis: path estimation, testing the product, estimating its uncertainty, and considering the full candidate set.

The tutorial connects each analytical decision to executable Stata and Python code and a corresponding result. Holding the observations and regressions fixed shows how individual and city clustering produce the same path estimates but different standard errors. The Green and Social examples then show how two types of interval from 2,000 city-bootstrap replications and BH

adjustment across six candidates answer distinct questions about the evidence. City-level inference follows the level of policy exposure and shared conditions. For health psychology, the contribution is an executable link between a proposed psychosocial explanation and the evidence needed to assess it. Researchers can use the workflow to define samples, calculate indirect associations, and identify which interpretations depend on clustering, interval construction, or the candidate set.

2 Methodological framework

The Healthy City example connects shared living conditions and individual social engagement to an energia-based proxy for suicide-related vulnerability. It asks whether policy-associated changes in these candidate pathways could help explain differences in the proxy score. This turns a health psychology question about context and psychological functioning into a sequence of observable associations. Four methods contribute to the answer. The three regressions estimate the paths. The Sobel test provides a supplementary test of their product. The city bootstrap shows how much the estimated product varies when cities are resampled. BH adjustment considers the evidence when several candidates are tested.

From candidate mechanisms to observable indicators

The macro-level environmental pathway follows the social ecological framework: city policy may change the surroundings in which residents live [3]. The psychological rationale concerns opportunities for restoration and recovery from stress [4]. Green is the percentage of built-up urban area covered by greenery. It describes city-wide coverage; access, use, and experienced restoration are separate characteristics. The micro-level social pathway draws on the importance of connectedness in the Three-Step Theory of suicide and on research linking social activities to psychological outcomes [6,7]. Social counts participation during the preceding month in two activities: interacting with friends, and playing mahjong, chess, or cards or attending a community activity room. The score ranges from 0 to 2 and records participation in these activities. Participation records behavioral engagement, while perceived belonging, relationship quality, and received support are distinct dimensions [22]. Green and Social connect the proposed environmental and social pathways to observed indicators. Direct measures of restoration or connectedness would address the intervening psychological processes more specifically.

The other candidates capture distinct aspects of city conditions. PM2.5 measures annual mean fine particulate matter concentration in $\mu\text{g}/\text{m}^3$. Research on particulate air pollution and mental health motivates examining this candidate [23]. SO2 measures the total tonnage of industrial sulfur-dioxide emissions. Both describe city conditions rather than personal exposure. Road measures road surface area in square meters per person; walkability also depends on other features of the built environment [24]. Medical measures hospitals and health centers per 10,000 registered residents, representing health-service supply. Care received and service quality are separate dimensions [25]. Each candidate is fitted in a separate exploratory model, so the six products cannot be added to obtain a total mediated effect.

Estimating the three regression steps

The analysis uses CHARLS [8] observations from 2011, 2013, 2015, 2018, and 2020 for urban adults aged at least 45 years. Each observation represents one person at one survey wave. The outcome, Y , is the anergia-based proxy score defined above, retaining the item's 1–4 past-week frequency coding. Higher values indicate a higher proxy score. Policy exposure, D , equals the treated-city indicator multiplied by an indicator for years from 2017 onward; the observed post-policy waves are therefore 2018 and 2020. The sample contains 94 cities, including 10 treated cities.

For one candidate mediator, M , the three additive linear regressions are [12]:

$$Y_{ijt} = cD_{jt} + X'_{ijt}\beta_1 + \mu_{1j} + \lambda_{1t} + \varepsilon_{1ijt} \quad (1)$$

$$M_{ijt} = aD_{jt} + X'_{ijt}\beta_2 + \mu_{2j} + \lambda_{2t} + \varepsilon_{2ijt} \quad (2)$$

$$Y_{ijt} = c'D_{jt} + bM_{ijt} + X'_{ijt}\beta_3 + \mu_{3j} + \lambda_{3t} + \varepsilon_{3ijt} \quad (3)$$

Here, i identifies a person, j a city, and t a survey wave. The first regression estimates c , the policy–outcome association. The second estimates a , the policy–mediator association. The third estimates b , the mediator–outcome association after accounting for policy exposure, and c' , the remaining policy coefficient. For example, a asks whether the policy is associated with Green, while b asks whether Green is associated with the proxy score when the policy and controls are included. City effects, μ , account for persistent differences between cities. Survey-wave effects, λ , account for changes shared across cities at each wave. The control coefficients and fixed effects are estimated separately in each equation.

X contains sex, marital status, logged real household income per person, work status, agricultural household registration, education, number of children, retirement, activities-of-daily-living difficulties, smoking, drinking, self-rated health, and age. Education and self-rated health enter as numeric scores. These main pathway regressions use city and survey-year fixed effects and no survey or equal-city weights. Thus, Social's b can reflect both differences between people and changes within a person, and cities with more observed person-waves contribute more to the fitted relationships. A separate sensitivity analysis of the policy–outcome association uses individual and survey-year fixed effects.

The product ab combines the two paths. For Green, $a = 0.378438$ indicates a policy-associated difference of about 0.38 percentage points in green coverage. Its $b = -0.007406$ describes the adjusted difference in the proxy score per percentage point of green coverage. Their product is -0.002803 proxy-score points. For Social, $a = 0.062869$ and $b = -0.041043$, giving $ab = -0.002580$. Here, b describes the adjusted proxy-score difference associated with one additional type of social activity. These score differences do not quantify a change in the probability of suicidal ideation or behavior. A statistically significant overall policy association, c , is not required before estimating either product [12,14].

For each candidate, both path regressions use the same observations with complete values for the outcome, mediator, policy variable, and controls. We call this the candidate's common sample. The same sample is used for individual and city clustering, the Sobel test, and the starting estimate for both bootstraps. Green, Road, and PM2.5 each use 16,661 observations; Social uses 16,660; SO2 uses 15,430; and Medical uses 11,141. Medical has no usable 2020 observations. A common sample is therefore defined within each candidate, rather than requiring all six indicators to be present for every observation.

In Python, `common_sample` selects the observations for one candidate. `fit_paths` estimates its three regressions with city and survey-year effects and city-clustered standard errors. The policy coefficient from the mediator model is stored as `a`; the mediator coefficient from the outcome model is stored as `b`. The final line multiplies them:

```
sample = common_sample(base, required + [mediator])
a_fit, b_fit, total_fit, result, _ = fit_paths(sample, mediator)
a = result["a"]
b = result["b"]
indirect = a * b
```

Comparing individual and city clustering

The same person can contribute responses at several survey waves, and their proxy scores may be related over time. Individual clustering allows for this connection when calculating standard errors. An individual bootstrap similarly resamples respondents, keeping each selected person's observations together.

On the common samples, both Green and Social have `a` and `b` *p* values below .05 under individual clustering. They remain below .05 when the six `a`-path tests and six `b`-path tests are adjusted separately using BH. Their individual-bootstrap percentile and BC intervals also lie below zero before adjustment for multiple comparisons. The city analysis evaluates these same pathways while also accounting for shared conditions among residents of a city.

Different people in the same city may also have related outcomes because they share local conditions. For example, two residents interviewed in the same city and year receive the same Green value and the same policy exposure. These residents share one observed city-level change in greening [26]. Individual clustering accounts for dependence within each respondent; city clustering also allows dependence across residents and survey waves within a city [17]. This is why the city analysis provides the main basis for policy inference, including for Social, whose individual activity scores are linked to a shared city policy.

City and year effects remove persistent city differences and changes shared across cities, but people in the same city may still experience common influences [27]. Clustering allows for this connection when calculating standard errors, which summarize the uncertainty in the estimates. With the observations and regression model held fixed, changing the clustering level leaves `a`, `b`, and `ab` unchanged. It changes the standard errors and can therefore change whether a *p* value falls below .05.

Checking the policy association with individual fixed effects

Individual fixed effects address a different question from individual clustering. They control for each person's characteristics that remain constant across the observed waves, so estimation uses changes within the same person [28]. Individual clustering allows that person's regression errors to be related over time when calculating uncertainty. We add an individual-and-year fixed-effects model as a sensitivity analysis of the policy–outcome association, and report both individual- and city-clustered standard errors. Because respondents remain in the same city in these data, individual fixed effects also absorb persistent city differences. The sensitivity analysis estimates the overall policy coefficient, c , with individual and year effects; the estimates of a , b , and ab retain city and year effects.

Using the Sobel test as supplementary evidence

The Sobel test compares the estimated product ab with its approximate standard error [19]. Both coefficients and their city-clustered standard errors are estimated on the same complete observations used for the bootstrap. With s_a and s_b denoting the standard errors of a and b , the statistic is:

$$Z_{\text{Sobel}} = \frac{ab}{\sqrt{b^2 s_a^2 + a^2 s_b^2}} \quad (4)$$

The p value is calculated using the standard normal distribution. All tests in this article consider departures from zero in either direction, positive or negative. In the code, `se_a` and `se_b` are the city-clustered standard errors, and `indirect` is the estimated product:

```
sobel_se = np.sqrt(b**2 * se_a**2 + a**2 * se_b**2)
sobel_z = indirect / sobel_se
sobel_p = 2 * stats.norm.sf(abs(sobel_z))
```

For Green, $ab = -0.002803$, the Sobel standard error is 0.004623, and $p = .54433$. For Social, $ab = -0.002580$, the standard error is 0.001287, and $p = .04505$. Social's 95% Sobel interval is $[-0.005104, -0.000057]$. It is the only candidate with an unadjusted Sobel p value below .05; the other four candidates have p values between .53989 and .93590. The adjustment across all six Sobel tests is reported below.

The Sobel calculation gives a supplementary normal-approximation assessment of ab . It combines the two path standard errors while omitting their cross-equation covariance. Because the product's sampling distribution can be asymmetric [12,29], the bootstrap intervals provide a complementary assessment using resampled path estimates.

Estimating uncertainty by resampling cities

The city bootstrap repeats the analysis with different selections of cities. Each of 2,000 repetitions draws 94 cities with replacement, so a city can be selected more than once or not at all. Every selected city contributes all its included respondents and survey waves. Both path regressions are fitted again, and their coefficients are multiplied. Treated and untreated cities are drawn together, rather than sampled separately.

To form a 95% percentile interval, we sort the 2,000 products and take the values at the 2.5th and 97.5th percentiles. BC uses the same products but adjusts the percentile positions used for the endpoints according to where the original estimate falls in the bootstrap distribution [30]. This is the bias-corrected method, abbreviated BC; it does not include the extra acceleration adjustment used in BCa. In the following code, counts records how often each city is selected, and `weighted_solutions` refits the regressions for those selections:

```
draw_a, _ = weighted_solutions(a_fit, counts)
draw_b, _ = weighted_solutions(b_fit, counts)
products = draw_a[:, 0] * draw_b[:, 0]
```

For Social, the city percentile interval is $[-0.005465, 0.000120]$, which includes zero. Its BC interval is $[-0.005866, -0.000173]$, which lies below zero. Both use the same data and estimated product; the different rules for choosing endpoints cause the different judgments. Green's city percentile interval is $[-0.013908, 0.009700]$ and its BC interval is $[-0.017191, 0.005406]$; both include zero. Thus, Social's evidence depends on the interval method, while neither city interval excludes zero for Green.

FDR adjustment across six candidate indicators

Testing several candidates creates more chances of finding a small p value by chance. BH adjusts a group of p values, called a testing family [21]. With valid p values and appropriate relationships between the tests, it controls the false discovery rate: the average proportion of false findings among the results declared significant over repeated studies [31]. If no result is declared significant, that proportion is counted as zero.

Each group contains all six candidates. We adjust the six a-path p values together and the six b-path p values together, separately for individual and city clustering. The six city Sobel p values form another group. The bootstrap summaries are adjusted in two further groups, one for individual resampling and one for city resampling. There are seven groups in total, each adjusted separately. All path regressions use the common sample for the relevant candidate.

The bootstrap summaries need a different interpretation from regression p values. For each candidate, we count how many resampled products are negative and how many are positive. A smaller sign-tail summary means that fewer products fall on the less common side of zero. We apply the BH calculation to the six summaries to describe how considering all candidates changes these numbers:

```
city["sign_tail_BH"] = bh_adjust(city["sign_tail"])
```

For Green, the individual-bootstrap summary changes from .005997 before adjustment to .035982 afterward. For Social, it changes from .041979 to .125937. Both unadjusted values are below .05, but only Green remains below that threshold after accounting for all six candidates. In the city bootstrap, the adjusted values are .958521 for Green and .341829 for Social; none of the six candidates is below .05.

For Social, 1,944 of the 2,000 city-bootstrap products are negative and 56 are positive. This shows that most resampled estimates are negative. It does not directly give the probability of obtaining such evidence if the true indirect association were zero [32]. The sign-tail summary has not been

validated as a hypothesis-test p value, so applying BH to it does not establish FDR control [21]. The percentile and BC intervals remain ordinary 95% intervals for each candidate; BH does not change their endpoints.

The six common-sample Sobel p values are also adjusted together, as a separate supplementary comparison. Social's $p = .04505$ becomes $.27029$ after BH adjustment, and none of the six Sobel tests remains below $.05$. These values are kept separate from the adjusted bootstrap summaries. Social's unadjusted Sobel p value is below $.05$ and its city BC interval excludes zero, while its percentile interval includes zero and its adjusted values exceed $.05$.

Interpreting the proposed pathways

The estimates describe indirect associations with the proxy score. A causal interpretation additionally requires mediator changes to precede outcome changes and relevant sources of confounding to be addressed [15,16,33]. For example, another city initiative might change both greening and residents' well-being. In this case, mediators and outcomes are measured within the same survey wave, and some controls may respond to the policy. The policy comparison also assumes that treated and untreated cities would have followed comparable trends without the policy. City clustering estimates uncertainty while allowing related observations within cities; temporal ordering, confounding, and comparable trends are separate requirements for causal interpretation.

3 A step-by-step tutorial using the Healthy City study

The case asks how a change in shared surroundings might relate to individual psychological vulnerability through environmental conditions or social engagement. This connects the macro-level policy setting with the micro-level health psychology question. Green and Social provide the worked examples, while Road, SO₂, Medical, and PM_{2.5} remain in every six-candidate comparison. The outcome is the anergia-based proxy for suicide-related vulnerability defined above. We first select observations with the necessary measurements, then use those same observations throughout each candidate's analysis.

The following steps estimate the paths, calculate ab and its supplementary Sobel test, construct bootstrap intervals, and apply BH adjustment. Code excerpts show the calculations used for the reported results. The accompanying programs include the full loops and functions needed to run them for all six candidates.

Step 1 Identify the data and select the analysis samples

CHARLS is a longitudinal survey based at Peking University [8]. This study uses its 2011, 2013, 2015, 2018, and 2020 waves. Its publicly released data are available through the CHARLS website (<https://charls.pku.edu.cn/>). Researchers register an account, accept the data-use agreement, and apply for the survey waves they need [34]. In this study, respondent records are matched by city and year to policy status and city indicators. The city data sources include the China Urban-Rural Construction Statistical Yearbook and the China City Statistical Yearbook. One analysis row represents one person at one survey wave, with their outcome, controls, social activity score, and the corresponding city measures.

The outcome, `goingl`, records how often the respondent felt they “could not get going” during the previous week. It is interpreted as an anergia-related indication of reduced activation and used as the theoretically informed, non-validated proxy for suicide-related vulnerability introduced above. Renaming it outcome retains its coding: 1 for less than one day, 2 for 1–2 days, 3 for 3–4 days, and 4 for 5–7 days. Higher values indicate a higher anergia-based proxy score. The linear regression coefficients are expressed in points on this 1–4 scale; they are not changes in the probability of a suicidal event. Table 1 gives the source and analysis names of the candidate indicators.

Table 1 Candidate measures and their variable names

Candidate	Source variable	Analysis variable	Meaning
Green	GreenCoverageRateBD	m_green	Green coverage of built-up urban area, percent
Road	RoadSurAreaPerCap	m_road	Road surface area, m ² per person
SO2	Tons of SO2 emissions	m_so2	Annual industrial sulfur-dioxide emissions, tons
Medical	hosper	m_medical	Hospitals and health centers per 10,000 registered residents
PM2.5	pm	m_pm25	Annual mean PM2.5 concentration, µg/m ³
Social	act_12	m_social	Number of two activity types reported, 0–2

Notes: Green, Road, SO2, Medical, and PM2.5 are city measures. Social is measured for each respondent. Renaming a variable does not change its values. The original variable name in the analysis code is 工业二氧化硫排放量吨. For clarity, it is presented here as “Tons of SO2 emissions”

The indicators distinguish city conditions from residents’ activities. Green records city-wide green coverage, Road records road area per person, and Medical records health-service supply. SO2 describes industrial emissions, whereas PM2.5 describes ambient concentration. Social sums participation in two activity categories during the preceding month: interacting with friends, and playing mahjong, chess, or cards or attending a community club. Its 0–2 score records the breadth of participation in these categories. Access to green space, walkability, personal pollution exposure, care received, activity frequency, and relationship quality require separate measures.

The Stata commands below load the study data, give the variables shorter names, select urban adults aged at least 45, and create the policy variable. `input_path` specifies the location of the analysis data. `rural` identifies urban residence; `rural2` is a separate control for household registration.

Stata

```
use "`input_path'", clear
rename (ID iwy treat rgoingl) (person_id year treated outcome)
rename (GreenCoverageRateBD RoadSurAreaPerCap act_12) ///
      (m_green m_road m_social)

rename 工业二氧化硫排放量吨 m_so2

rename (hosper pm) (m_medical m_pm25)
replace city=strtrim(city)
drop if missing(person_id) | missing(city) | missing(year)
drop if missing(wave)
keep if age>=45 & age<. & rural==0
capture drop did
generate double did=treated*(year>=2017)
egen long city_id=group(city)
egen long respondent_id=group(person_id)
local controls gender marry log_real_inc_per rwork rural2 edu ///
      hchild retire adlab_c smoken drinkl srh age
local labels Green Road SO2 Medical PM25 Social
local mediators m_green m_road m_so2 m_medical m_pm25 m_social
```

In these commands, `did` equals `treated` multiplied by an indicator for $\text{year} \geq 2017$. A treated-city resident therefore has `did = 0` in 2015 and `did = 1` in 2018. An untreated-city resident has `did = 0` at both waves. The `group` commands create numerical identifiers from the city and respondent identifiers. The `controls` macro contains sex, marital status, logged real income per person, work status, household registration, education, number of children, retirement, activities-of-daily-living difficulties, smoking, drinking, self-rated health, and age.

Python uses the same variables and selection rules. `read_input` reads the observations and constructs policy exposure. `common_sample` keeps observations with complete values for the listed variables and applies the same fixed-effects sample rules as Stata. The following lines create base, the sample with complete outcome, policy, and control values:

Python

```
parser = argparse.ArgumentParser()
parser.add_argument("--data", required=True)
parser.add_argument("--out", default="results/python")
args = parser.parse_args()
out = Path(args.out)
out.mkdir(parents=True, exist_ok=True)
d, flow = read_input(args.data)
required = ["outcome", "did", "year", "treated"] + CONTROLS
base = common_sample(d, required)
```

Here, `d` contains eligible observations and `base` contains complete observations for the outcome, policy variable, and controls. There are 34,228 eligible person-wave observations. The baseline complete sample contains 16,661 observations from 7,995 respondents in 94 cities, including 10 treated cities. Logged income is missing for 14,469 eligible observations, or 42.3%. The panel is unbalanced, with respondents contributing at different combinations of waves. The resulting estimates describe residents with complete measurements for the relevant analysis.

The calculations use Stata 18 with `reghdfe` and `ftools`, or Python with NumPy, pandas, and SciPy. The accompanying code includes the estimation functions. The supplied coefficients and bootstrap products also allow the interval and BH calculations to be repeated without downloading respondent records.

Step 2 Estimate the three regression paths with individual clustering

The first model asks whether policy exposure is associated with the proxy score. City fixed effects account for persistent differences between cities; survey-year fixed effects account for changes shared across cities. The option `vce(cluster respondent_id)` allows repeated observations from the same person to be related when calculating standard errors. This unweighted model uses city and survey-year fixed effects. Step 8 adds a separate sensitivity analysis with individual and survey-year fixed effects.

Stata

```
reghdfe outcome did `controls', ///
    absorb(city_id year) vce(cluster respondent_id)
generate byte baseline=e(sample)
```

The policy coefficient is -0.086679 proxy-score points, with an individual-clustered standard error of 0.027856. The command `e(sample)` identifies the observations used in this model, and `baseline` stores that selection. For Green, we retain baseline observations with a Green value and use this same selection for both path regressions:

Stata

```
generate byte green_sample=baseline & !missing(m_green)
reghdfe m_green did `controls' if green_sample, ///
    absorb(city_id year) vce(cluster respondent_id)
scalar green_a=_b[did]
scalar green_se_individual=_se[did]
reghdfe outcome m_green did `controls' if green_sample, ///
    absorb(city_id year) vce(cluster respondent_id)
scalar green_b=_b[m_green]
```

The first regression estimates $a = 0.378438$: policy exposure is associated with about 0.38 percentage points more green coverage. The second estimates $b = -0.007406$: one percentage point more coverage is associated with a 0.007406-point lower proxy score after accounting for policy exposure and controls. Both regressions use 16,661 observations. Stata retrieves a coefficient with `_b[variable]` and its standard error with `_se[variable]`. The same procedure is repeated for each candidate in Table 2.

Table 2 Path estimates on the common samples with individual clustering

Candidate	a	a p value	b	b p value
Green	0.378438	3.41e-06	-0.007406	0.00505
Road	-2.305227	4.85e-77	-0.000214	0.92473
SO2	-8292.462473	1.43e-06	-1.79e-07	0.23399
Medical	0.145363	1.29e-80	-0.017084	0.61022
PM2.5	-4.850085	9.14e-55	-0.000901	0.42773
Social	0.062869	0.04724	-0.041043	1.45e-07

Notes: Both paths for each candidate use the same observations. The same samples are used for city clustering, Sobel tests, and bootstrap estimates. These p values have not yet been adjusted for multiple comparisons.

For Social, $a = 0.062869$ indicates a higher activity score associated with policy exposure. Its $b = -0.041043$ indicates a lower proxy score associated with one additional activity type, after adjustment for policy and the controls. Both Green and Social have a and b p values below .05. The other four candidates have a-path p values below .05, but their b-path p values exceed .05. These tests describe the two component associations; the product ab and its uncertainty connect them to the evidence for the candidate pathway [12].

Step 3 Calculate the indirect association on the same observations

The product ab combines the two coefficients from Step 2. Because both coefficients use the same observations, they describe the same candidate pathway in the same sample. The calculation for Green is:

Stata

```
scalar green_ab=scalar(green_a)*scalar(green_b)
display scalar(green_a) " " scalar(green_b) " " scalar(green_ab)
```

For Green, $0.378438 \times (-0.007406)$ gives $ab = -0.002803$. For Social, $0.062869 \times (-0.041043)$ gives $ab = -0.002580$. The calculations use the full coefficient precision, rather than the rounded values shown here. Each product is measured in points on the 1–4 proxy scale. These are the same coefficients and products used in the path tables, Sobel tests, and starting estimates for both bootstraps.

Table 3 Common samples used throughout each candidate analysis

Candidate	Observations	Respondents	Survey period
Green	16,661	7,995	2011–2020
Road	16,661	7,995	2011–2020
SO2	15,430	7,673	2011–2020
Medical	11,141	6,353	2011–2018
PM2.5	16,661	7,995	2011–2020
Social	16,660	7,995	2011–2020

Notes: Each candidate retains 94 cities, including 10 treated cities. Medical covers 2011–2018 because usable 2020 values are unavailable. Other candidates cover all five waves.

Defining a common sample within each candidate retains the observations available for that analysis. A respondent can therefore contribute to Green even when Medical is missing. Requiring all six indicators simultaneously would remove such observations and exclude the 2020 wave. The candidate-specific samples preserve this information while keeping the two paths comparable within each analysis. Because samples and indicator scales differ, the products do not provide a direct ranking of pathway importance.

Step 4 Choose the clustering level from the policy design

Individual clustering accounts for repeated answers from one person. It does not account for different residents sharing unexplained changes within a city. For example, two residents interviewed in the same city and year receive the same Green value and policy exposure. Their responses add information about individual outcomes, but do not represent two independent changes in city policy or greening. City clustering allows dependence across both residents and waves within a city [17].

Stata

```

reghdfe m_green did `controls' if green_sample, ///
    absorb(city_id year) vce(cluster city_id)
display "a = " _b[did]
display "Individual SE = " scalar(green_se_individual)
display "City SE = " _se[did]

```

Green's α coefficient remains 0.378438, while its standard error changes from 0.081426 with individual clustering to 0.598211 with city clustering. The observations and regression model are identical. This comparison makes the effect of the uncertainty calculation visible: allowing shared city influences gives less precise evidence about this coefficient. City clustering matches the level of the policy and shared local conditions. The same reasoning applies to Social, whose individual activity score is linked to a city-level policy exposure.

For Social, the following commands fit the outcome model, store β and its city-clustered standard error, retain that model's observations, and fit α on exactly those observations. `_se[variable]` retrieves the standard error of the named coefficient.

Stata

```
reghdfe outcome m_social did `controls' if baseline, ///
    absorb(city_id year) vce(cluster city_id)
generate byte social_sample=e(sample)
scalar path_b=_b[m_social]
scalar path_sb=_se[m_social]
reghdfe m_social did `controls' if social_sample, ///
    absorb(city_id year) vce(cluster city_id)
scalar path_a=_b[did]
scalar path_sa=_se[did]
```

In Python, mediator takes each of the six variable names in turn. `fit_paths` estimates the three regressions on that candidate's common sample. The result stores the coefficients and city-clustered standard errors. The two fitted path models are also retained for bootstrap resampling:

Python

```
sample = common_sample(base, required + [mediator])
a_fit, b_fit, total_fit, result, _ = fit_paths(sample, mediator)
a = result["a"]
b = result["b"]
se_a = result["SE_a"]
se_b = result["SE_b"]
indirect = a * b
```

For Social, $a = 0.062869$, $b = -0.041043$, and $\text{indirect} = -0.002580$. The city-clustered standard errors are 0.028196 and 0.008974. Python removes city means and includes survey-year indicators to fit the same models as Stata's `absorb(city_id year)`. No respondent changes city in these samples, so drawing whole cities also keeps each person's repeated observations together.

Step 5 Calculate the supplementary Sobel test

The Sobel test assesses the size of ab relative to an approximate standard error calculated from the uncertainty in a and b [19]. The Stata code uses the coefficients and city-clustered standard errors stored in Step 4. `scalar(...)` retrieves a stored number.

Stata

```

scalar path_ab=scalar(path_a)*scalar(path_b)
scalar path_se=sqrt(scalar(path_b)^2*scalar(path_sa)^2 ///
+ scalar(path_a)^2*scalar(path_sb)^2)
scalar path_z=scalar(path_ab)/scalar(path_se)
scalar path_p=2*normal(-abs(scalar(path_z)))
scalar path_lower=scalar(path_ab)-invnormal(.975)*scalar(path_se)
scalar path_upper=scalar(path_ab)+invnormal(.975)*scalar(path_se)
display scalar(path_ab) " " scalar(path_se) " " scalar(path_p)
display scalar(path_lower) " " scalar(path_upper)

```

The first two lines calculate the product and its Sobel standard error. Dividing the product by that standard error gives the test statistic. The p value allows for an association in either direction; the factor of 2 accounts for both directions. The lower and upper limits are the product minus or plus 1.96 standard errors. This conventional formula uses the separate uncertainties in a and b, without estimating how their errors move together. Python performs the same calculation:

Python

```

sobel_se = np.sqrt(b**2 * se_a**2 + a**2 * se_b**2)
sobel_z = indirect / sobel_se
sobel_p = 2 * stats.norm.sf(abs(sobel_z))
sobel_ci = indirect + np.array([-1, 1]) * 1.959963984540054 * sobel_se

```

Table 4 Unadjusted Sobel results on the common samples

Candidate	ab	Sobel SE	Sobel p value
Green	-0.002803	0.004623	0.54433
Road	0.000494	0.006138	0.93590
SO2	0.001488	0.003124	0.63377
Medical	-0.002483	0.006800	0.71498
PM2.5	0.004369	0.007127	0.53989
Social	-0.002580	0.001287	0.04505

Notes: Both paths use the common sample for that candidate and city-clustered standard errors. No multiple-testing adjustment is included in this table.

Social has a Sobel standard error of 0.001287, $p = .04505$, and a 95% interval of $[-0.005104, -0.000057]$. Green has a standard error of 0.004623 and $p = .54433$. These results illustrate how the estimated product and its standard error jointly determine the test result. The Sobel calculation is supplementary because the sampling distribution of a product can be asymmetric [12]. Step 6 examines uncertainty through resampling; Step 7 assesses all six candidates together.

Step 6 Resample clusters and construct percentile and BC intervals

Generate one product in every city resample

Each of the 2,000 city-bootstrap replications draws 94 cities with replacement. If a city is selected twice, all of its included residents and waves contribute twice. If it is not selected, none contributes. The a and b models are refitted using the same city selections, and their coefficients are multiplied. Treated and untreated cities are not sampled separately. Changing the resampling unit changes how uncertainty is represented; the original-sample product remains the same.

A resample can be represented by counts. Each row of counts is one repetition, and each column is one of the 94 cities. A count of 2 means the city is selected twice; the counts in each row sum to 94. The Stata and Python analyses use the same city selections, generated with NumPy's PCG64 generator and seed 2025. Using the same selections makes the results comparable across the two programs.

Python

```
B = 2000
counts = weights(a_fit["g"], B, 2025,
                 "code/shared_draw_counts_94.csv")
draw_a, _ = weighted_solutions(a_fit, counts)
draw_b, _ = weighted_solutions(b_fit, counts)
products = draw_a[:, 0] * draw_b[:, 0]
result.update(bootstrap_summary(products, indirect))
```

`weighted_solutions` fits the regressions for every row of counts. The first column of `draw_a` contains the policy coefficients from the mediator regressions; the first column of `draw_b` contains the mediator coefficients from the outcome regressions. Multiplying them gives 2,000 products. A city selected twice receives twice the weight. All observations in that city receive the same weight, so its within-city means stay unchanged.

The corresponding Stata program loads the same count matrix and calls the supplied Mata least-squares functions. Mata is Stata's matrix programming language; no Python process is needed for the Stata city analysis. After fitting the original models, the central resampling operations are:

Stata Mata

```
a_draws=hc_solve(a_fit,counts)
b_draws=hc_solve(b_fit,counts)
products=a_draws[,1]:*b_draws[,1]
hc_csv(filename,"draw,a,b,ab",
        ((1::rows(counts)),a_draws[,1],b_draws[,1],products))
```

In this code, the first columns of `a_draws` and `b_draws` are multiplied row by row. Each saved row contains the repetition number, a, b, and ab. All 2,000 repetitions produced an estimate for every candidate.

Calculate the two intervals from the 2000 products

The percentile method sorts the 2,000 products and uses the 2.5th and 97.5th percentiles as the interval endpoints. BC uses the same sorted products [30]. It first calculates the proportion that is at or below the original estimate, then uses this position to adjust which percentiles define the endpoints. This changes the interval, not the original estimate of *ab*.

Python

```
def bootstrap_summary(products, observed):
    products = np.asarray(products, dtype=float)
    B = len(products)
    pct = np.quantile(products, [0.025, 0.975], method="linear")
    proportion = np.mean(products <= observed)
    z0 = stats.norm.ppf(np.clip(proportion, 0.0001, 0.9999))
    bc_prob = stats.norm.cdf(
        2 * z0 + stats.norm.ppf([0.025, 0.975])
    )
    bc = np.quantile(products, bc_prob, method="linear")
    return {"B": B, "Pct_L": pct[0], "Pct_U": pct[1],
            "BC_L": bc[0], "BC_U": bc[1], "z0": z0,
            "BC_prob_L": bc_prob[0], "BC_prob_U": bc_prob[1]}
```

products contains the 2,000 resampled estimates, while *observed* is *ab* from the original sample. The variable *proportion* locates that original estimate within the bootstrap results. *norm.ppf* and *norm.cdf* convert between probabilities and positions on the standard normal distribution to calculate the BC percentile positions. *np.quantile* then retrieves the interval endpoints, interpolating between adjacent products when needed. The small bounds in *np.clip* avoid undefined calculations at probabilities of exactly 0 or 1; they do not affect the Social example. The function calculates BC, not BCa.

For Social, $z_0 = -0.085329$. The BC method uses approximately the 1.6560th and 96.3217th percentiles, instead of the 2.5th and 97.5th percentiles. This gives $[-0.005866, -0.000173]$ for BC, compared with $[-0.005465, 0.000120]$ for the percentile method. The estimated product is unchanged. The endpoints differ because the two methods select different positions in the same set of bootstrap results.

The Stata calculation makes the same steps explicit. Here, *ab* is the column of bootstrap products, and *scalar(indirect)* is the original-sample product retained before those products are loaded:

Stata

```
quietly count
scalar B=r(N)
quietly count if ab<=scalar(indirect)
scalar fraction=r(N)/B
scalar z0=invnrmal(max(.0001,min(.9999,fraction)))
scalar prob_lower=normal(2*z0+invnrmal(.025))
scalar prob_upper=normal(2*z0+invnrmal(.975))
mata: v=st_data(., "ab")
mata: st_numscalar("pct_lower", hc_q(v, .025))
mata: st_numscalar("pct_upper", hc_q(v, .975))
mata: st_numscalar("bc_lower", hc_q(v, st_numscalar("prob_lower")))
mata: st_numscalar("bc_upper", hc_q(v, st_numscalar("prob_upper")))
```

`hc_q` is the supplied linear-interpolation quantile function. The following is its complete definition. It sorts the products, locates the requested probability between two adjacent positions, and interpolates between them, using the same convention as the Python code.

Stata Mata

```
real scalar hc_q(real colvector v, real scalar p)
{
    real scalar h, l, u
    v=sort(v, 1); h=1+(rows(v)-1)*p; l=floor(h); u=ceil(h)
    return(v[l]+(h-l)*(v[u]-v[l]))
}
```

Social's BC interval lies below zero, while its percentile interval includes zero. Reporting both shows how the endpoint rule affects the judgment near zero. BC adjusts the percentile positions used to construct one candidate's interval. The BH calculation in Step 7 addresses a separate question by adjusting the six candidate results together. Both intervals reported here remain unadjusted for multiple comparisons.

Compare city and individual resampling

The individual bootstrap instead draws respondents with replacement and retains every selected person's observed waves. The Python loop below uses a multinomial draw to record how often each respondent appears. The helper `_two_way_fe_cross` recalculates the city means under those respondent frequencies, because the mix of respondents within a city can change. Reusing the original city means in this procedure would fit a different model.

Python

```
sufficient = individual._prepare_bootstrap_stats(raw, source)
G = sufficient["n_ids"]
observed = individual._mediation_from_cross(
    individual._two_way_fe_cross(sufficient, np.ones(G))
)
rng = np.random.default_rng(2025)
draws = np.empty((2000, 5))
for r in range(2000):
    frequency = rng.multinomial(G, np.full(G, 1/G)).astype(float)
    cross = individual._two_way_fe_cross(sufficient, frequency)
    draws[r] = individual._mediation_from_cross(cross)
products = draws[:, 4]
```

sufficient contains the respondent-level totals needed to refit the models. The fitting function returns a, b, the total policy coefficient, the remaining policy coefficient, and ab, in that order. The last column of draws therefore contains the products. The individual bootstrap is implemented in Python; the city bootstrap is implemented in both Stata and Python. The two resampling levels use different draw sequences.

Table 5 Unadjusted bootstrap intervals for all six candidates

Candidate and unit	ab	95% percentile interval	95% BC interval
Green Individual	-0.002803	[-0.005250, -0.000766]	[-0.005675, -0.001017]
Green City	-0.002803	[-0.013908, 0.009700]	[-0.017191, 0.005406]
Road Individual	0.000494	[-0.009730, 0.009970]	[-0.009760, 0.009966]
Road City	0.000494	[-0.015232, 0.015421]	[-0.013924, 0.017610]
SO2 Individual	0.001488	[-0.000994, 0.004045]	[-0.000604, 0.004356]
SO2 City	0.001488	[-0.011559, 0.008885]	[-0.003163, 0.015469]
Medical Individual	-0.002483	[-0.011875, 0.007049]	[-0.011450, 0.007619]
Medical City	-0.002483	[-0.015828, 0.020088]	[-0.017408, 0.017853]
PM2.5 Individual	0.004369	[-0.006596, 0.014982]	[-0.005902, 0.015620]
PM2.5 City	0.004369	[-0.008562, 0.026777]	[-0.006065, 0.031607]
Social Individual	-0.002580	[-0.005524, -0.000135]	[-0.005720, -0.000258]
Social City	-0.002580	[-0.005465, 0.000120]	[-0.005866, -0.000173]

Notes: Each row uses 2,000 replications. Within each candidate, both resampling levels use the same observations, variables, controls, and fixed effects. None of these intervals has been adjusted for multiple comparisons.

The comparison holds the original-sample products, observations, and regression specifications fixed while changing the resampling unit. Both individual-bootstrap intervals exclude zero for Green and Social. With city resampling, both Green intervals include zero; Social's BC interval excludes zero and its percentile interval includes zero. The intervals therefore show how the resampling unit, resulting draws, and endpoint rule shape uncertainty around the same estimates. City results provide the primary basis for policy inference because exposure and shared local influences operate at that level.

Step 7 Apply Benjamini–Hochberg FDR adjustment across six candidates

Define what is being adjusted together

When six candidates are examined, there are more opportunities to obtain a small p value by chance. BH adjusts the six results together [21,31]. Its target is the false discovery rate (FDR): over repeated studies, the average share of false findings among the results declared significant. If

none is declared significant, that share is counted as zero. FDR is not the probability that one particular result is false. The usual BH interpretation requires valid p values and either independent tests or appropriate forms of positive dependence.

A testing family is simply the group of results adjusted together. For example, the six city Sobel tests form one family. The six a-path tests with individual clustering form another. Every family includes Green, Road, SO2, Medical, PM2.5, and Social. Table 6 shows the seven groups used here.

Table 6 The families retained for BH calculations

Inputs	Clustering or resampling	Number of families	Members per family
a-path p values	Individual and city, separately	2	6
b-path p values	Individual and city, separately	2	6
Sobel p values	City	1	6
Bootstrap sign-tail summaries	Individual and city, separately	2	6

Notes: Seven groups are adjusted separately, with six candidates in each. All path estimates use the candidate's common sample. The bootstrap inputs are sign-tail summaries, whose interpretation differs from that of hypothesis-test p values.

BH adjustment applies to the specified testing family. In this analysis, the six Sobel tests form one family, and the percentile and BC intervals remain unadjusted. For a new study, define families from the research questions and claims evaluated, retain the full candidate set in each specified family, and report any changes to that definition.

Follow the Sobel example through every BH calculation

BH begins by sorting the six inputs from smallest to largest. For the value in position i , multiply it by $6/i$. Next, work backwards from the largest rank: replace each value by the smaller of its own scaled value and the adjusted value in the next row. This backward step prevents a smaller input from receiving a larger adjusted value than a larger input. Cap adjusted values at 1 and return each value to its candidate. The following table shows all steps for the six Sobel tests.

Table 7 Worked BH calculation for the six city Sobel tests

Candidate	Rank i	Sobel p	6p divided by i	BH value
Social	1	0.045049	0.270291	0.270291
PM2.5	2	0.539895	1.619684	0.857975
Green	3	0.544331	1.088662	0.857975
SO2	4	0.633767	0.950650	0.857975
Medical	5	0.714979	0.857975	0.857975
Road	6	0.935901	0.935901	0.935901

Notes: Inputs are shown in ascending order. The 6p/i column is intermediate and may exceed 1. Final BH values include the backward minimum and the cap at 1. Calculations use full precision, not the displayed rounded values.

Social has the smallest Sobel p value, 0.0450485. Its first scaled value is $6 \times 0.0450485 / 1 = 0.270291$. Every later candidate's final adjusted value is larger, so the backward step leaves Social at 0.270291. None of the six adjusted Sobel p values is below .05. Green illustrates why multiplication alone is insufficient: its intermediate value is 1.088662, but the backward minimum gives 0.857975. BH is therefore not simply multiplying every p value by six.

Implement BH in Python and retain the indicator labels

The complete Python function used in the analysis is below. order records the sorting positions; scaled contains the intermediate 6p/i values; the reverse cumulative minimum performs the backward step. adjusted[order] restores the input order so that every adjusted value remains attached to the correct candidate.

Python

```
def bh_adjust(values):
    values = np.asarray(values, dtype=float)
    m = len(values)
    order = np.argsort(values, kind="stable")
    ranked = values[order]
    scaled = ranked * m / np.arange(1, m + 1)
    adjusted_sorted = np.minimum.accumulate(scaled[::-1])[::-1]
    adjusted_sorted = np.minimum(adjusted_sorted, 1.0)
    adjusted = np.empty(m)
    adjusted[order] = adjusted_sorted
    return adjusted
```

The following code applies the function to the six rows of each result table. Sobel_BH_6 adjusts the six Sobel p values. sign_tail_BH applies the same calculation to the six bootstrap sign-tail summaries, whose interpretation is discussed below. City and individual results are handled separately.

Python

```
city = pd.read_csv(root / "city_results.csv")
individual = pd.read_csv(root / "individual_results.csv")
city["Sobel_BH_6"] = bh_adjust(city["Sobel_p"])
city["sign_tail_BH"] = bh_adjust(city["sign_tail"])
individual["sign_tail_BH"] = bh_adjust(individual["sign_tail"])
print(city[["Mediator", "Sobel_p", "Sobel_BH_6",
            "sign_tail", "sign_tail_BH"]].to_string(index=False))
```

SciPy provides the same BH calculation through `scipy.stats.false_discovery_control` with `method="bh"`. The explicit function above shows how sorting, scaling, and the backward minimum produce each adjusted value.

Implement the same calculation in Stata

This Stata program implements the same algorithm without an additional multiple-testing package. `original_order` preserves the candidate order, `rank` is the sorted position, and the loop works from the next-to-last row back to the first. The program is applied to one complete six-member family at a time.

Stata

```
capture program drop bh_adjust
program define bh_adjust
    syntax varname(numeric), Generate(name)
    tempvar original_order rank scaled
    generate long `original_order'=_n
    sort `varlist' `original_order'
    generate long `rank'=_n
    generate double `scaled'=`varlist'*_N/`rank'
    generate double `generate'=min(1,`scaled')
    local last=_N-1
    forvalues j=`last'(-1)1 {
        quietly replace `generate'= ///
            min(`generate',`generate'[_n+1]) in `j'
    }
    sort `original_order'
end
```

After loading the city summary table, the next commands apply the program to the two distinct product-related inputs. `assert _N==6` checks that the full candidate set is present. Neither call modifies the percentile or BC endpoints.

Stata

```
import delimited using results/stata/city_results.csv, ///
    clear case(preserve)
assert _N==6
bh_adjust Sobel_p, generate(Sobel_BH_6)
bh_adjust sign_tail, generate(sign_tail_BH)
list Mediator Sobel_p Sobel_BH_6 sign_tail sign_tail_BH, noobs
export delimited using results/stata/city_adjusted.csv, replace
```

The same Stata function can be applied to each six-candidate group. It returns the adjusted values alongside the candidate labels. These BH calculations use summary results and do not require respondent records.

Distinguish bootstrap sign-tail summaries from hypothesis-test p values

The bootstrap summary counts how many products fall below and above zero. The calculation used in this study doubles the smaller count after adding one, divides by $B + 1$, and caps the result at 1. The full calculation is:

Python

```
def sign_tail_summary(products):
    products = np.asarray(products, dtype=float)
    negative = int(np.sum(products < 0))
    positive = int(np.sum(products > 0))
    zero = int(np.sum(products == 0))
    B = len(products)
    value = min(1.0, 2 * (min(negative, positive) + 1) / (B + 1))
    return {"Negative": negative, "Positive": positive,
            "Zero": zero, "sign_tail": value}
```

For Social, the city bootstrap gives 1,944 negative products and 56 positive products, with no exact zeros. Its sign-tail summary is $2 \times (56 + 1) / (2,000 + 1) = 0.056972$. BH adjustment across the six city summaries gives 0.341829. These numbers describe how often the resampled products fall on either side of zero. The resampling did not impose a true indirect association of zero, so this count is not automatically a valid p value for that hypothesis. We therefore call it a sign-tail summary. Applying the BH formula does not turn it into a validated test or establish FDR control.

Table 8 City bootstrap sign-tail summaries before and after BH

Candidate	Sign-tail summary	BH-adjusted summary
Green	0.639680	0.958521
Road	0.958521	0.958521
SO2	0.938531	0.958521
Medical	0.756622	0.958521
PM2.5	0.578711	0.958521
Social	0.056972	0.341829

Notes: All six candidates are retained. These are adjusted descriptive summaries; applying BH does not establish valid FDR control for these inputs.

Table 9 Individual bootstrap sign-tail summaries before and after BH

Candidate	Sign-tail summary	BH-adjusted summary
Green	0.005997	0.035982
Road	0.914543	0.914543
SO2	0.235882	0.471764
Medical	0.610695	0.732834
PM2.5	0.418791	0.628186
Social	0.041979	0.125937

Notes: The individual and city families are adjusted separately. All displayed intervals in Step 6 remain unadjusted.

Green's individual summary changes from 0.005997 to 0.035982, and Social's changes from 0.041979 to 0.125937. Both are below .05 before adjustment, but only Green remains below it afterward. None of the adjusted city summaries is below .05.

Apply the adjustment to the component paths as well

The regression p values are adjusted in four groups: individual a, individual b, city a, and city b. All use the common samples. The code selects those results, groups them by clustering level and path, and applies BH within each six-candidate group:

Python

```
paths = pd.read_csv(root / "component_paths.csv")
reported = paths.loc[
    paths["Sample"].eq("Common")
].copy()
family_keys = ["Level", "Sample", "Path"]
reported["BH"] = reported.groupby(family_keys) ["p"].transform(bh_adjust)
print(reported[ [*family_keys, "Mediator", "p",
"BH"] ].to_string(index=False))
```

Table 10 Individual component paths before and after BH

Candidate	a p value	a BH	b p value	b BH
Green	3.41e-06	4.09e-06	0.00505	0.01515
Road	4.85e-77	1.45e-76	0.92473	0.92473
SO2	1.43e-06	2.14e-06	0.23399	0.46798
Medical	1.29e-80	7.71e-80	0.61022	0.73227
PM2.5	9.14e-55	1.83e-54	0.42773	0.64160
Social	0.04724	0.04724	1.45e-07	8.68e-07

Notes: These are the common-sample paths reported in Table 2. The six a-path p values and six b-path p values are adjusted separately.

Green and Social retain p values below .05 for both paths under individual clustering. For Social's a path, $p = .04724$ and the BH-adjusted value is also .04724. This concerns the separate paths, not the uncertainty in their product. Under city clustering, Green's adjusted values are .586805 for a and .109166 for b; Social's are .063911 and .000089. Only Social's city b path is below .05 after adjustment. This b-path result describes the mediator–outcome association; the a path and the product-level analysis address the other parts of the candidate pathway.

BC and BH act on different results. BC changes the endpoints of a bootstrap interval for one candidate. BH adjusts six p values or six sign-tail summaries together. For Social, the city BC interval excludes zero, while its BH-adjusted sign-tail summary is .341829 and its adjusted Sobel p value is .27029. Together, these results distinguish an unadjusted interval for one candidate from adjustments across six candidates. The displayed intervals remain unadjusted.

Step 8 Check the policy association with individual fixed effects

Would the policy association remain negative after accounting for characteristics that are stable within each person? We examine this by replacing city fixed effects with individual fixed effects in the policy–outcome regression. The outcome, policy variable, eligibility criteria, control-variable list, and year effects are retained. The accompanying Stata routine repeats the data preparation in Step 1 and creates baseline from the policy model in Step 2. Both clustering choices apply to the same individual-fixed-effects model:

Stata

```

reghdfe outcome did `controls' if baseline, ///
    absorb(respondent_id year) vce(cluster respondent_id)
generate byte respondent_fe_sample=e(sample)
estimates store policy_fe_individual
display "p value = " 2*ttail(e(df_r),abs(_b[did]/_se[did]))

reghdfe outcome did `controls' if baseline, ///
    absorb(respondent_id year) vce(cluster city_id)
assert e(sample)==respondent_fe_sample
estimates store policy_fe_city
display "p value = " 2*ttail(e(df_r),abs(_b[did]/_se[did]))

```

The command `absorb(respondent_id year)` controls for persistent differences between respondents and changes shared in each survey year. Gender and age are absorbed by these fixed effects in this dataset and therefore have no separately estimated coefficients. Stata also excludes 3,082 observations from respondents who have only one complete observation; those respondents provide no within-person change. The model retains 13,579 observations from 4,913 respondents in 94 cities. The variable `respondent_fe_sample` identifies these observations, and the second regression uses exactly the same sample.

Table 11 Policy association under two fixed-effect specifications

Fixed effects	Clustering	N	Policy coefficient	SE	p value	95% CI
City and year	Individual	16,661	−0.086679	0.027856	0.001867	[−0.141285, −0.032074]
City and year	City	16,661	−0.086679	0.033045	0.010184	[−0.152301, −0.021058]
Individual and year	Individual	13,579	−0.099845	0.037393	0.007607	[−0.173153, −0.026538]
Individual and year	City	13,579	−0.099845	0.046192	0.033222	[−0.191573, −0.008117]

Notes: N denotes person-wave observations. The city-and-year model includes 7,995 respondents; the individual-and-year model includes 4,913. Every specification includes 94 cities. Each interval and p value is unadjusted. The two clustering choices within each fixed-effect specification use identical observations and regressors. These policy-coefficient sensitivity results are separate from the six-candidate BH families in Step 7.

With individual and year effects, the policy coefficient is −0.099845 under both clustering choices. Its standard error is 0.037393 with individual clustering and 0.046192 with city clustering; the corresponding p values are .007607 and .033222. Both confidence intervals lie below zero. City clustering remains the principal inference because residents share city-level policy exposure and local influences, even after individual fixed effects have been included.

The negative overall policy association is also present in the within-person analysis. Its comparison with the main estimate of −0.086679 involves both a different fixed-effect

specification and a smaller sample, so this is a joint comparison of sample and specification. This sensitivity analysis evaluates c after accounting for stable respondent characteristics. The estimates of a , b , and ab , the Sobel tests, bootstrap intervals, and BH calculations remain those of the city-and-year fixed-effects models. Time-varying confounding remains a separate consideration.

Step 9 Return to the Healthy City research question

The case demonstrates how the workflow separates component-path evidence, uncertainty in the product, and multiple comparisons. Green and Social have individual-clustered a - and b -path p values below .05, and both individual-bootstrap intervals exclude zero before adjustment. City resampling gives two intervals containing zero for Green; Social's BC interval lies below zero, whereas its percentile interval includes zero. Social's Sobel p value changes from .04505 to .27029 after BH adjustment, and none of the six adjusted city sign-tail summaries is below .05. These results identify which judgments depend on the interval rule and on considering all six candidates. The individual-fixed-effects sensitivity analysis separately shows a negative overall policy association under both clustering choices. Candidate-pathway estimates remain indirect associations from the city-and-year fixed-effects models.

The design defines the interpretation of these associations. Annual city conditions, past-month activities, and the past-week proxy response are paired within survey waves, leaving their temporal ordering unresolved [15,16,35]. Some controls may respond to policy exposure, and complete observations may differ from those excluded. Missing activity responses may also have different meanings across waves. The comparison includes 10 treated cities. These features concern identification, measurement, and sample coverage; city clustering addresses dependence among observations when estimating uncertainty.

The workflow links each result to a health psychology question: whether policy exposure is associated with a change in a proposed pathway, whether that indicator is associated with the anergia-based proxy, and how strongly the data support their combined association. The product, its intervals, and the six-candidate comparison make these judgments explicit. For the environmental and social explanations, the next substantive questions concern experienced restoration and connectedness, which require direct measures of those processes. An interval containing zero can remain compatible with associations in either direction [36]. Statistical evidence about an association and identification of a causal mechanism therefore require separate interpretation.

4 Discussion

This tutorial connects a health psychology question about psychosocial pathways to an analysis that researchers can implement and adapt. The social ecological framework motivates attention to shared conditions at the macro level, while the role of connectedness in suicide theory motivates social engagement at the micro level [3,6]. Green and Social translate specific features of these perspectives into observed indicators. The practical contribution is to connect this theoretical reasoning to common samples, path estimates, and uncertainty assessments. The accompanying code makes the regressions, supplementary Sobel test, bootstrap intervals, and BH calculations part of the same candidate-specific analysis. This helps researchers determine which part of a

proposed pathway has statistical support and what additional measurement would make its psychological interpretation more specific.

The three regressions answer different questions. The a path asks whether the policy is associated with the candidate mediator. The b path asks whether the mediator is associated with the outcome after accounting for policy exposure. Testing the two paths separately provides evidence about each component association; the Sobel and bootstrap calculations assess the uncertainty in their product ab [37]. Using the same observations for both paths makes the product interpretable and allows the Sobel and bootstrap results to address the same estimate. This common sample is defined within each candidate. Medical, for example, has no usable 2020 observations, so its results cover a different period from those for Green.

The clustering comparison shows why the policy design should guide the calculation of uncertainty. Repeated observations from one person are related, but people in the same city may also share changes in their surroundings. Green's coefficient is unchanged when the same model and sample are analyzed with different clustering levels, yet its standard error changes substantially. Drawing whole cities also keeps their residents and survey waves together. City clustering is therefore the main choice because it matches the policy level and the relationships between observations [17]. The comparison makes the consequences of that design choice visible without changing the estimated paths.

The individual-fixed-effects analysis illustrates why the regression specification and clustering level should be considered separately. Individual effects account for stable differences between people; clustering addresses dependence when calculating standard errors. Both can be used in the same model. In this case, the policy coefficient remains negative with individual and year effects under either clustering choice. The estimate describes respondents with repeated complete observations, and comparison with the main model involves changes in both sample and specification. This check evaluates the overall policy association; the candidate pathways are assessed separately with city and year effects.

Reporting both bootstrap intervals makes the evidence easier to assess. Before multiple-comparison adjustment, the individual-bootstrap percentile and BC intervals lie below zero for Green and Social. With city resampling, both Green intervals include zero. Social's percentile interval is $[-0.005465, 0.000120]$, whereas its BC interval is $[-0.005866, -0.000173]$. The point estimate is the same, but the interval methods give different judgments near zero. Reporting both intervals makes this dependence visible and ties the interpretation to the stated interval method.

Multiple comparisons add a further question: how does the evidence change when all six candidates are considered? For the city Sobel tests, Social's p value rises from .04505 to .27029 after BH adjustment. Among the individual-bootstrap sign-tail summaries, Green and Social are below .05 before adjustment, but only Green remains below it afterward. No adjusted city-bootstrap summary is below .05. These bootstrap summaries have not been validated as hypothesis-test p values, so their adjustment does not establish FDR control. BH requires suitable p values and relationships between tests [21], and it does not adjust the displayed bootstrap intervals.

The workflow offers an analytical starting point for health psychology studies in which participants share an intervention or setting. For example, a school-based stress-management study could examine whether a program is associated with coping confidence and whether that confidence is associated with subsequent psychological distress. A community participation program could examine social engagement or perceived belonging. Hospital or therapy-group studies could assess coping or perceived support as candidate pathways to psychological adjustment. These are possible applications of the sequence of questions and calculations. In each, mediation analysis can inform the assessment of proposed intervention components when the design supports the relevant interpretation [1].

Adaptation starts with the psychological construct, the timing of its measurement, and the outcome the intervention is intended to change. Researchers should select measures of those constructs, define common analysis samples, and specify which candidate pathways will be considered together. The model form, controls, missing-data handling, fixed effects, clustering unit, and testing families should follow the new design, with these choices and their assumptions reported explicitly [38]. Schools, hospitals, or therapy groups may be relevant clustering units when participants share intervention exposure or local influences; assignment and dependence determine the choice [17,18]. The present case uses linear models for an ordered proxy score and city and year effects. Other outcomes or intervention designs require corresponding model choices. The reusable contribution is the sequence of decisions and calculations, with substantive settings chosen for each application.

The accompanying code supports use of the workflow in two ways. Researchers with the required respondent and city measures can estimate the models and repeat the cluster resampling. They can also use the supplied coefficients and bootstrap products to reproduce the interval and BH calculations without respondent records. For a new application, the variable definitions, time periods, model, and cluster selections need to match that study. The functions can be adapted to implement those choices. The Stata examples also fit the individual-fixed-effects policy model with each clustering choice and show how its sample differs from that of the main model.

The analyses estimate indirect associations within the observed survey waves. A causal interpretation additionally requires temporal ordering and control of relevant confounding [15,35]. Past-month activities, the past-week proxy response, and annual city conditions are paired within waves, so their order is unresolved. The main pathway models account for city and year effects; stable respondent differences and changing local influences can still confound the pathways. The individual-fixed-effects sensitivity analysis addresses stable respondent differences in the policy–outcome regression, while time-varying confounding and the separate mediator paths require their own assessment. Controls such as income and health may respond to the policy, affecting their role in adjustment [16,39].

Complete-case selection, particularly through missing income, determines which observations contribute to the estimates. The unweighted estimates are not necessarily population-representative, and the potential for bias in regression estimates depends on the missingness process [40]. The outcome remains a theoretically informed, non-validated proxy for suicide-related vulnerability. Fatigue, physical ill health, and other depressive symptoms can also contribute to the response. Interpreting a lower proxy score as a reduction in suicidal ideation,

behavior, or clinical risk requires validation against direct suicidal-outcome measures. Future applications can pair symptom measures with direct assessments of connectedness, perceived stress, and the psychological outcome of interest, choosing the combination that matches the proposed pathway. The six separate candidate models estimate distinct associations rather than additive components of a total mediated effect.

The case contains 94 cities, including 10 treated cities. These policy units determine the information available for city-level inference. The 2,000 bootstrap replications approximate uncertainty by repeatedly drawing from those units; they do not add independent policy comparisons. Unequal cluster sizes and connections between cities can affect the uncertainty estimates [17,41]. Applications should account for relevant shared influences, including regional shocks and spillovers, when choosing the clustering unit. The workflow separates reproducible computation, the conditions for appropriate inference, and the evidence needed for causal interpretation, making each decision explicit.

5 Conclusion

This tutorial provides a practical way to assess candidate pathways between an intervention and a psychological outcome. The Healthy City example carries path estimates through supplementary Sobel testing, cluster bootstrap intervals, and BH adjustment, with Stata and Python code for each calculation. Researchers can use the code to examine a proposed psychological process by defining suitable measures and a common sample for its path regressions, choosing clustering that reflects the study design, and interpreting the indirect association alongside its uncertainty and the full candidate set.

References

- [1] MacKinnon DP, Luecken LJ. How and for whom? Mediation and moderation in health psychology. *Health Psychol.* 2008;27(2 Suppl):S99–S100.
[https://doi.org/10.1037/0278-6133.27.2\(Suppl.\).S99](https://doi.org/10.1037/0278-6133.27.2(Suppl.).S99)
- [2] Office of the National Patriotic Health Campaign Committee. Notice on launching the Healthy City pilot programme. Published November 7, 2016. [In Chinese].
<https://www.nhc.gov.cn/jkj/c100063/201611/b5c94663f2b348ad85f36fab2b8b39f6.shtml>
- [3] Stokols D. Translating social ecological theory into guidelines for community health promotion. *Am J Health Promot.* 1996;10(4):282–298.
<https://doi.org/10.4278/0890-1171-10.4.282>
- [4] Markevych I, Schoierer J, Hartig T, et al. Exploring pathways linking greenspace to health: theoretical and methodological guidance. *Environ Res.* 2017;158:301–317.
<https://doi.org/10.1016/j.envres.2017.06.028>
- [5] South EC, Hohl BC, Kondo MC, MacDonald JM, Branas CC. Effect of greening vacant land on mental health of community-dwelling adults: a cluster randomized trial. *JAMA Netw Open.* 2018;1(3):e180298. <https://doi.org/10.1001/jamanetworkopen.2018.0298>
- [6] Klonsky ED, May AM. The Three-Step Theory (3ST): a new theory of suicide rooted in the “ideation-to-action” framework. *Int J Cogn Ther.* 2015;8(2):114–129.
<https://doi.org/10.1521/ijct.2015.8.2.114>
- [7] Lin X, Liu S, Hu Z, Xu H. Social activities and long-term depressive-symptoms trajectories among middle-aged and older adults in China: a population-based cohort study. *Front Psychiatry.* 2023;14:1131084. <https://doi.org/10.3389/fpsy.2023.1131084>
- [8] Zhao Y, Hu Y, Smith JP, Strauss J, Yang G. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *Int J Epidemiol.* 2014;43(1):61–68.
<https://doi.org/10.1093/ije/dys203>
- [9] Radloff LS. The CES-D Scale: a self-report depression scale for research in the general population. *Appl Psychol Meas.* 1977;1(3):385–401.
<https://doi.org/10.1177/014662167700100306>
- [10] Boey KW. Cross-validation of a short form of the CES-D in Chinese elderly. *Int J Geriatr Psychiatry.* 1999;14(8):608–617.
[https://doi.org/10.1002/\(sici\)1099-1166\(199908\)14:8<608::aid-gps991>3.0.co;2-z](https://doi.org/10.1002/(sici)1099-1166(199908)14:8<608::aid-gps991>3.0.co;2-z)
- [11] Qiu T, Klonsky ED, Klein DN. Hopelessness predicts suicide ideation but not attempts: a 10-year longitudinal study. *Suicide Life Threat Behav.* 2017;47(6):718–722.
<https://doi.org/10.1111/sltb.12328>
- [12] MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. *Annu Rev Psychol.* 2007;58:593–614. <https://doi.org/10.1146/annurev.psych.58.110405.085542>
- [13] Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol.* 1986;51(6):1173–1182. <https://doi.org/10.1037/0022-3514.51.6.1173>
- [14] Zhao X, Lynch JG Jr, Chen Q. Reconsidering Baron and Kenny: myths and truths about mediation analysis. *J Consum Res.* 2010;37(2):197–206. <https://doi.org/10.1086/651257>
- [15] Cole DA, Maxwell SE. Testing mediational models with longitudinal data: questions and tips in the use of structural equation modeling. *J Abnorm Psychol.* 2003;112(4):558–577.
<https://doi.org/10.1037/0021-843X.112.4.558>

- [16] Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010;15(4):309–334. <https://doi.org/10.1037/a0020761>
- [17] Cameron AC, Miller DL. A practitioner's guide to cluster-robust inference. *J Hum Resour*. 2015;50(2):317–372. <https://doi.org/10.3368/jhr.50.2.317>
- [18] Abadie A, Athey S, Imbens GW, Wooldridge JM. When should you adjust standard errors for clustering? *Q J Econ*. 2023;138(1):1–35. <https://doi.org/10.1093/qje/qjac038>
- [19] Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociol Methodol*. 1982;13:290–312. <https://doi.org/10.2307/270723>
- [20] Fritz MS, Taylor AB, MacKinnon DP. Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behav Res*. 2012;47(1):61–87. <https://doi.org/10.1080/00273171.2012.640596>
- [21] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol*. 1995;57(1):289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [22] Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. *Soc Sci Med*. 2000;51(6):843–857. [https://doi.org/10.1016/s0277-9536\(00\)00065-4](https://doi.org/10.1016/s0277-9536(00)00065-4)
- [23] Braithwaite I, Zhang S, Kirkbride JB, Osborn DPJ, Hayes JF. Air pollution (particulate matter) exposure and associations with depression, anxiety, bipolar, psychosis and suicide risk: a systematic review and meta-analysis. *Environ Health Perspect*. 2019;127(12):126002. <https://doi.org/10.1289/ehp4595>
- [24] Ewing R, Cervero R. Travel and the built environment: a meta-analysis. *J Am Plann Assoc*. 2010;76(3):265–294. <https://doi.org/10.1080/01944361003766766>
- [25] Andersen RM. Revisiting the behavioral model and access to medical care: does it matter? *J Health Soc Behav*. 1995;36(1):1–10. <https://doi.org/10.2307/2137284>
- [26] Moulton BR. An illustration of a pitfall in estimating the effects of aggregate variables on micro units. *Rev Econ Stat*. 1990;72(2):334–338. <https://doi.org/10.2307/2109724>
- [27] Bertrand M, Duflo E, Mullainathan S. How much should we trust differences-in-differences estimates? *Q J Econ*. 2004;119(1):249–275. <https://doi.org/10.1162/003355304772839588>
- [28] Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press; 2010.
- [29] MacKinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: distribution of the product and resampling methods. *Multivariate Behav Res*. 2004;39(1):99–128. https://doi.org/10.1207/s15327906mbr3901_4
- [30] Efron B. Nonparametric standard errors and confidence intervals. *Can J Stat*. 1981;9(2):139–158. <https://doi.org/10.2307/3314608>
- [31] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*. 2001;29(4):1165–1188. <https://doi.org/10.1214/aos/1013699998>
- [32] Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. *Biometrics*. 1991;47(2):757–762. <https://doi.org/10.2307/2532163>
- [33] VanderWeele TJ. *Mediation analysis: a practitioner's guide*. *Annu Rev Public Health*. 2016;37:17–32. <https://doi.org/10.1146/annurev-publhealth-032315-021402>
- [34] China Health and Retirement Longitudinal Study. Frequently asked questions. Peking University. [In Chinese]. Accessed October 9, 2026. <https://charls.pku.edu.cn/Faq/cjw.htm>

- [35] Maxwell SE, Cole DA. Bias in cross-sectional analyses of longitudinal mediation. *Psychol Methods*. 2007;12(1):23–44. <https://doi.org/10.1037/1082-989x.12.1.23>
- [36] Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–350. <https://doi.org/10.1007/s10654-016-0149-3>
- [37] MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. *Psychol Methods*. 2002;7(1):83–104. <https://doi.org/10.1037/1082-989x.7.1.83>
- [38] Lee H, Cashin AG, Lamb SE, et al. A guideline for reporting mediation analyses of randomized trials and observational studies: the AGReMA statement. *JAMA*. 2021;326(11):1045–1056. <https://doi.org/10.1001/jama.2021.14075>
- [39] Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20(4):488–495. <https://doi.org/10.1097/ede.0b013e3181a819a1>
- [40] Hughes RA, Heron J, Sterne JAC, Tilling K. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol*. 2019;48(4):1294–1304. <https://doi.org/10.1093/ije/dyz032>
- [41] MacKinnon JG, Nielsen MØ, Webb MD. Cluster-robust inference: a guide to empirical practice. *J Econom*. 2023;232(2):272–299. <https://doi.org/10.1016/j.jeconom.2022.04.001>