How Do Household Energy Transitions Work?*

Jill Baumgartner (Co-PI)[1] Sam Harper (Co-PI)[1]
Chris Barrington-Leigh[1] Collin Brehmer[2] Ellison M. Carter[2]
Xiaoying Li[2] Brian E. Robinson[1] Guofeng Shen[3]
Talia J. Sternbach[1] Shu Tao[3] Kaibing Xue[4] Wenlu Yuan[1]
Xiang Zhang[1] Yuanxun Zhang[4]

2024-10-15

Table of contents

1 Introduction **6**

2 Background **6**
2.1 Context for the policy 6
2.2 Prior evidence on household energy interventions and air pollution 7
2.3 Prior evidence on clean energy interventions and cardiovascular outcomes 8
2.4 Evaluating the mechanisms through which policies may affect health outcomes. 9

3 Specific Aims and Overarching Approach **9**

4 Study Design and Methods **10**
4.1 Study area 10
4.2 Location and participant recruitment and enrolment 10
4.3 Data Collection Overview 12
4.3.1 Air Pollution 12
4.3.2 Outdoor and indoor (household) air temperature 21
4.3.3 Objective measurement of household stove use using sensors 21
4.3.4 Questionnaires 22

*Affiliations [1] McGill Univeristy; [2] Colorado State University; [3] Peking University; [4] University of the Chinese Academy of Sciences

4.3.5 Blood pressure . . . 23
4.3.6 Self-reported respiratory symptoms and airway inflammation . . . 23
4.3.7 Blood inflammatory and oxidative stress markers . . . 24
4.3.8 Anthropometric measurements . . . 25
4.4 Measuring policy impacts . . . 25
4.5 Measuring pathways and mechanisms . . . 27

5 Data Analysis **28**
5.1 Total Effect . . . 29
5.2 Mediation Analysis . . . 29
5.3 Identification of potential confounders and model covariates . . . 30
5.4 Multiple imputation for covariates and indoor $PM_{2.5}$ in analyses with BP outcomes . . . 32

6 Results **32**
6.1 Description of study sample by treatment . . . 33
6.2 Summary of PM and BC measurements . . . 36
6.3 Policy uptake . . . 36
6.4 Aim 1: Policy impacts and potential mediation . . . 39
6.4.1 Impact of policy on potential mediators of air pollution and indoor temperature . . . 39
6.4.2 Impact of the policy on health outcomes . . . 42
6.4.3 Mediated impact on health outcomes . . . 43
6.5 Aim 2: Source contributions . . . 44
6.5.1 Source analysis using positive matrix factorization . . . 45
6.5.2 Description of $PM_{2.5}$ sources identified . . . 48
6.5.3 Impact of policy on outdoor and personal exposure to the mixed combustion source 51
6.6 Aim 3: Mediation by source contribution . . . 52

7 Discussion and Conclusions **52**
7.1 Adoption of the heat pump technology and adherence to the policy . . . 53
7.2 Impacts of the policy on health . . . 54
7.3 Impacts of the policy on air pollution and its sources . . . 56
7.4 Assumptions, strengths, and limitations . . . 58

8 Implications of Findings **62**

9 Data Availability Statement **63**

10 Acknowledgements **63**

11 References **63**

A Appendices **77**
A.1 Biomarker descriptives . . . 77
A.2 Missing data . . . 78
A.2.1 Missingness across imputed variables . . . 78
A.2.2 Missing data by enrollment cohort and outcome . . . 80
A.2.3 Imputation results . . . 82
A.3 Participant flow diagram . . . 83
A.4 Sample sizes . . . 84
A.5 District-level statistics . . . 85
A.6 Policy uptake . . . 86
A.7 Heterogeneity in treatment effects . . . 87
A.7.1 Personal exposure . . . 87
A.7.2 Indoor $PM_{2.5}$ . . . 88
A.7.3 Indoor temperature . . . 89
A.7.4 Blood pressure outcomes . . . 90
A.7.5 Mediation analyses for blood pressure . . . 92
A.7.6 Self-reported respiratory outcomes . . . 94
A.7.7 FeNO . . . 97
A.7.8 Outdoor and personal mixed combustion . . . 98
A.8 Impact of adjustment for district on air pollution estimates . . . 99
A.9 Impact of sample composition on brachial and central blood pressure results. . . . 100
A.10 Impact of including Season 3 data . . . 101
A.11 Impact of sample composition on FeNO results . . . 102
A.12 Alternative PMF analyses . . . 103
A.12.1 Disaggregated analyses . . . 103
A.12.2 PMF results disaggregated by day . . . 105
A.12.3 PMF results disaggregated by month . . . 106
A.12.4 Personal exposure sample diagnostics . . . 107
A.12.5 Outdoor exposure sample diagnostics . . . 108
A.12.6 $PM_{2.5}$ constituents for outdoor samples . . . 109
A.12.7 $PM_{2.5}$ constituents for personal samples . . . 110
A.13 Pre-trends . . . 111
A.13.1 Blood pressure . . . 111
A.13.2 Personal exposure and black carbon . . . 113
A.13.3 Self-reported respiratory outcomes . . . 115
A.14 Impact of group and time fixed effects . . . 116
A.15 Retrospective design analysis . . . 117
A.15.1 All outcomes . . . 117
A.15.2 Blood pressure . . . 118

Abbreviations and other terms **119**

About the authors **120**

Other publications **121**

Abstract

Introduction

Since 2015, thousands of rural and peri-urban villages across Beijing and northern China have been treated by a household Clean Heating Policy (CHP) that banned household coal burning and subsidized the costs of electric heaters and electricity. Whether this large-scale policy was successful in improving air quality and health remains an important and unresolved question. We estimated the effects of the CHP policy on air quality and cardiopulmonary health in Beijing villages, and quantified how much of the policy's effects on health were mediated by changes in air pollution and indoor temperature.

Methods

In winter 2018-19 we enrolled 1003 participants in 50 Beijing villages that were eligible for, but not currently treated by, the CHP and followed them over four consecutive winter data collection waves. In waves 1, 2 and 4, we administered questionnaires and measured participants' anthropometrics, blood pressure (BP), airway inflammation (FeNO), and 24-h personal exposure to fine particulate matter ($PM_{2.5}$). Fasting whole blood samples were obtained at clinic visits in waves 1 and 2 for analysis of glucose, lipid profile, and markers of inflammation and oxidative stress. Wintertime outdoor $PM_{2.5}$ was measured in all 4 waves, and wintertime indoor temperature and indoor $PM_{2.5}$ were measured in waves 2, 3 and 4. The $PM_{2.5}$ filters were analyzed for their mass, black carbon, and chemical composition, which were used for source apportionment. To estimate the impacts of the policy we used a difference-in-differences design that accommodated the staggered roll-out of the CHP. We used 'extended' two-way fixed effects models and marginal effects to quantify the effect of the policy on air pollution and health outcomes. We further evaluated whether villages treated by the policy in different years respond differently to the policy, and whether the observed health impacts of the policy were mediated through changes in air pollution or home (indoor) temperature.

Results

At baseline (wave 1), mean participant age was 60 y (SD=9.2), 60% were female, and most (63%) worked in agriculture. Geometric mean personal exposures to $PM_{2.5}$ were twice as high as outdoor $PM_{2.5}$ (72 versus 36 $\mu g/m^3$), and the main source contributors were local and transported dust,

regional and domestic coal and biomass burning, and secondary pollutants. By waves 2, 3, and 4 there were a cumulative total of 10, 17, and 20 villages (out of 50 total) exposed to the CHP. Uptake and adherence to the policy was high: among villages treated in wave 2, the proportion of households using heat pumps and coal heaters, respectively, changed from 3% and 97% in wave 1 to 94% and 3% in wave 4, with similar clean energy transitions in villages exposed to the policy in later waves. Marginal effects derived from multivariable extended two-way fixed effects models showed that exposure to the policy increased indoor temperature by 1-2°C and reduced indoor seasonal $PM_{2.5}$ by approximately 20 µg/m^3. Treatment by the policy also reduced contributions to $PM_{2.5}$ from solid fuel sources, including household coal burning, and improved blood pressure (~1.5 mmHg lower systolic and diastolic) and self-reported respiratory symptoms (~8 percentage point reduction in any symptoms). There was notable heterogeneity in effects across treatment cohorts, with larger benefits to indoor $PM_{2.5}$ and health in villages treated in earlier relative to later years. In the mediation analysis, indoor $PM_{2.5}$ and indoor temperature explained most of the total effect of the policy on systolic BP and roughly half of the total effect on diastolic BP, but did not explain improvements in self-reported respiratory symptoms. We did not find evidence of meaningful effects of the policy on outdoor or personal exposure to $PM_{2.5}$, or on biomarkers of inflammation and oxidative stress.

Conclusions

In this comprehensive field-based assessment of a large-scale household energy policy in Beijing, we observed high fidelity and compliance with the CHP. Exposure to the policy reduced blood pressure and self-reported chronic respiratory symptoms, and the effects for blood pressure were mediated by reductions in indoor $PM_{2.5}$ and improvements in home temperature, providing empirical evidence that clean household energy policies can provide population health benefits.

1 Introduction

China is deploying an ambitious clean energy policy to transition up to 70% of households in its northern provinces from residential coal heaters to electric or gas "clean" space heating, including a large-scale roll out across rural and peri-urban Beijing villages, referred to in this document as the Clean Heating Policy (CHP). To meet this target the Beijing municipal government announced a two-pronged program that designates coal-restricted areas and simultaneously offers subsidies to night-time electricity rates and for the purchase and installation of electric-powered heat pumps to replace traditional coal-heating stoves. The policy was piloted in 2015 and, starting in 2016, was rolled out on a village-by-village basis. The variability in when the policy was applied to each village allowed us to treat the roll-out of the program as a quasi-randomized intervention and evaluate its impacts on air quality and health. Household air pollution is a well-established risk factor for adverse health outcomes over the entire lifecourse, yet there is no consensus that clean energy interventions can improve these health outcomes based on evidence from randomized trials (Lai et al. 2024). Households may be differentially affected by the CHP due to factors such as financial constraints and user preferences, and there is uncertainty about whether and how the policy may affect indoor and outdoor air pollution, as well as heating behaviors and health outcomes.

2 Background

2.1 Context for the policy

The CHP builds on China's long history of launching ambitious, large-scale policies and programs to promote clean household energy transition and support rural energy infrastructure development (Zhang and Smith 2007). China was a relatively early initiator of rural electrification projects in the 1950s and achieved complete (100%) electrification of households by 2016 (Yang 2021), which undoubtedly facilitated the current policy option to replace coal stoves with electric-powered heat pumps. Several decades later, China achieved what is likely still the largest improvement in household energy efficiency in history with regards to the population affected by a single program. The National Improved Stove Program (NISP) and its provincial- and county-level counterparts were initiated in the early 1980s and are credited with introducing 180 million improved cooking and heating stoves by the late 1990s. All NISP stoves had chimneys and some had manual or electric blowers to promote more efficient combustion (Zhang and Smith 2007), with the primary goal of increased biomass fuel efficiency to promote rural welfare and reduce pressure on local forests and a secondary goal of improving indoor air quality (Sinton et al. 2004). Because NISP focused mainly on biomass cookstoves, it had limited impacts on the rapid increase in coal heating stove installation during that same period, most of which were implemented without chimneys and with rudimentary designs (Zhang and Smith 2007). Though NISP was a significant achievement in early clean energy transition, especially for biomass cookstoves, the rural energy demands and

air pollution challenges of 21st century China required a renewed effort to promote transition to cleaner rural energy, particularly for rural heating where progress significantly lagged behind energy transition for cooking.

Most of northern China has a temperate continental monsoon climate that is characterized by cold, dry winters and hot, humid summers. Access to central heating is limited to urban areas and thus most peri-urban and rural households have historically heated their homes using coal heaters and biomass *kangs* (a traditional Chinese energy technology that integrates at least four different home functions including cooking, a bed for sleeping, space heating, and home ventilation). Household coal burning was a major contributor to indoor and outdoor air pollution in northern China, especially in winter. Prior to the CHP, over 100 million rural households consumed ~200 million tons of coal to meet more than 80% of northern China's residential space heating demand (Dispersed Coal Management Research Group 2023), which contributed to roughly 30% of wintertime air pollution (GBD MAPS Working Group 2016). In 2013, emissions inventories indicated that coal combustion from industrial, electricity, and residential heating sources was the single largest estimated contributor to population exposures to $PM_{2.5}$ in China and responsible for an estimated 366,000 annual premature deaths (GBD MAPS Working Group 2016).

Banning residential coal burning and providing homes with clean heating alternatives through the CHP was considered a potentially important intervention to improve rural development, reduce local and regional fine particulate matter ($PM_{2.5}$), and mitigate air pollution-related health impacts. A number of clean heating options, including electric heat pumps, gas heaters, and electric resistance heaters with thermal storage, were promoted by the Chinese government (Dispersed Coal Management Research Group 2023). By 2021, over 36 million households in northern China were treated by the CHP and an estimated 21 million additional households are expected to be treated by 2025. Whether this large-scale energy policy yielded air quality and health benefits remains a critical and unresolved question.

2.2 Prior evidence on household energy interventions and air pollution

Household energy interventions, mostly cooking-related, that replace traditional solid fuel stoves with more efficient and less-polluting alternatives have been implemented and studied extensively in countries including China over the past several decades. While the introduction of more efficient household stoves and fuels is expected to reduce indoor air pollution and exposures, evidence of their real-world effectiveness in achieving health-relevant air pollution reductions has been mixed, with some studies actually finding worse air quality in homes that received the intervention (Quansah et al. 2017). Further, most previous studies evaluated smaller-scale interventions implemented by civil society organizations or investigators themselves, and the indoor and local air quality benefits of large-scale household energy policies like the CHP have been rarely empirically investigated, especially at a sub-city spatial resolution or in countries in the Global South. In Ireland, county-level residential coal bans in the 1990s were associated with 40-70% decreases in black smoke

concentrations in ban-affected areas (Dockery et al. 2013). In Australia, a wood-burning stove exchange lowered daily wintertime PM_{10} from 44 to 27 $\mu g/m^3$ (Johnston et al. 2013), and clean energy policies in New Zealand were associated with 11-36% reductions in winter PM_{10} (Scott and Scarrott 2011). The few previous evaluations of the CHP reported small decreases in outdoor $PM_{2.5}$ (-7 to -2.4 $\mu g/m^3$) in municipalities or prefectures treated by the policy compared with untreated neighboring regions (Niu et al. 2024; Song et al. 2023; Tan et al. 2023; Yu et al. 2021), and a recent modeling study estimated 36% lower personal exposure to $PM_{2.5}$ based on household-reported changes in fuel use (Meng et al. 2023). These studies captured wide geographic areas, but none included field-based measurements of air pollution or personal exposures, which can differ considerably from modeled estimates (Thompson et al. 2019), and few accounted for secular changes in air quality over time, limiting conclusions about the causal effect of the policy on air quality.

2.3 Prior evidence on clean energy interventions and cardiovascular outcomes

Most previous health assessments of household energy interventions have focused on cookstoves rather than heating technologies, though in many settings cookstoves are also used for space heating. Randomized trials of less polluting cookstoves generally indicate a cardiovascular benefit. In older Guatemalan women, a chimney stove intervention lowered exposure to air pollution and reduced the occurrence of nonspecific ST-segment depression (McCracken et al. 2011). Randomized trials in Guatemala, Nigeria, and Ghana also showed reductions in blood pressure (systolic range: -3.7 to -1.3 mmHg) in women assigned to gas, ethanol, or improved combustion biomass stoves. In contrast, recent single country (Peru) and large multi-country (Household Air Pollution Intervention Network, HAPIN) randomized trials found no benefit of LPG stoves on gestational blood pressure (Checkley et al. 2021; Ye et al. 2022) despite much large reductions (~66% lower) in exposure to $PM_{2.5}$ and black carbon than what was observed in trials showing a BP benefit of intervention (Johnson et al. 2022).

The few population-based evaluations of large-scale residential energy policies also suggest a cardio-respiratory benefit of clean energy transition. Residential wood-burning bans were associated with reductions in cardiovascular hospitalizations (-7%) in California (Yap and Garcia 2015) and with reduced cardiovascular (-17.9%) and respiratory (-22.8%) mortality in Australia (Johnston et al. 2013), though neither study fully controlled for secular changes in health that were unrelated to the policy. Most relevant to our study are two quasi-experimental assessments of coal replacement policies. In Ireland, reductions in respiratory not but cardiovascular mortality were observed following their coal ban (Dockery et al. 2013). A multi-city study of Chinese adults in cities where the CHP was piloted compared with adults in cities not in the pilot observed small decreases in chronic lung diseases (-3.0 to -1.1%) but no change in physician-diagnosed cardiovascular diseases, potentially due to the short (one-year) post-policy evaluation period or confounding by other unmeasured municipality-wide air quality or health-related policies (Wen et al. 2023).

Though household air pollution is considered a well-established health risk factor, which energy interventions can most effectively reduce air pollution exposures and improve health and are also scalable and sustainable remain critical and unanswered questions. In a recent Official American Thoracic Society Statement, for example, the committee did not reach a consensus that household energy interventions (including gas, ethanol, solar, and improved biomass cookstoves) improved health outcomes (including respiratory symptoms and blood pressure), with 55% of the committee saying no and 45% saying yes (Lai et al. 2024).

2.4 Evaluating the mechanisms through which policies may affect health outcomes.

With several notable exceptions (Alexander et al. 2018; Gould et al. 2023; McCracken et al. 2007; McCracken et al. 2011), decades of household energy intervention studies have found limited or no health benefit, which demonstrates the complexity of both implementing and evaluating interventions on cooking or space heating that are central to daily life (Ezzati and Baumgartner 2017; Lai et al. 2024). Energy interventions and policies, particularly those implemented at the household- or village-scales, can produce multiple behavioral, environmental, and health-related changes, making it important to investigate the mechanisms through which such policies exert their health impacts or lack of impact (Dominici et al. 2014). The health benefits achievable with transition from traditional coal stoves to a new electric home heating system, for example, may be influenced by factors including outdoor air quality (Lai 2019), the desirability and usage patterns of new and traditional stoves (Ezzati and Baumgartner 2017), average or variability in indoor temperature (Lewington et al. 2012), and behaviors including physical activity or time spent in the home (Lindemann et al. 2017). Only recently were these mediating factors considered in health assessments of household energy interventions, and rarely in a comprehensive or formalized way (Rosenthal et al. 2018). Understanding such mechanisms can provide valuable insights into the success (or failure) of clean energy programs or policies like the CHP in meeting their air quality and health targets, and may answer questions that can inform the design of more effective future energy interventions (Lai et al. 2024). For example, is there successful uptake of the policy? Are there cardiovascular-enhancing effects of improved air quality in homes that are treated by the policy? Does the policy lead to heating behavior changes that result in colder homes and thus offset any cardiovascular-enhancing effects of improved air quality? Answers to these questions are facilitated by the analysis of mediating pathways, a key aim of this study.

3 Specific Aims and Overarching Approach

We used three data collection waves in winter 2018/19, winter 2019/20, and winter 2021/22, as well as a partial wave in winter 2020/21, to advance the following aims:

1. Estimate how much of the CHP's overall effect on health, including respiratory symptoms and cardiovascular outcomes (blood pressure, blood inflammatory and oxidative stress markers), can be attributed to its impact on changes in $PM_{2.5}$;
2. Quantify the impact of the policy on outdoor air quality and personal air pollution exposures, and specifically the source contribution from household coal burning;
3. Quantify the contribution of changes in the chemical composition of $PM_{2.5}$ from different sources to the overall effect on health outcomes.

4 Study Design and Methods

4.1 Study area

Beijing is the capital of China (population of 21.9 million in 2020) and covers a large geographic area (~16,000 km^2) that includes a highly developed and densely-populated urban core that is surrounded by several satellite towns and thousands of peri-urban and rural villages in the periphery. Beijing winters typically begin in early November and tend to be cold, dry, and windy, with the lowest temperatures mostly often occurring in January (-3°C, on average), thus requiring space heating (An et al. 2021). Most urban areas of Beijing are connected to a central heating grid that supplies home heating from central locations, whereas rural and many peri-urban areas have historically relied on individual space heating units that, prior to 2015, were largely fueled by unprocessed coal (Duan et al. 2014).

4.2 Location and participant recruitment and enrolment

Between December 2018 and January 2019 we recruited 50 villages across 4 administrative districts (Fangshan, Huairou, Mentougou, and Miyun) in the Beijing municipality in northern China. The villages predominately used coal for heating at the time of enrollment and were eligible for but not currently participating in the CHP. Roughly half of the villages were expected to enter into the policy during our study (Figure 1). We used local guides in each village to help determine a roster of households that were not vacant during the winter months, from which we randomly selected households to recruit for participation.

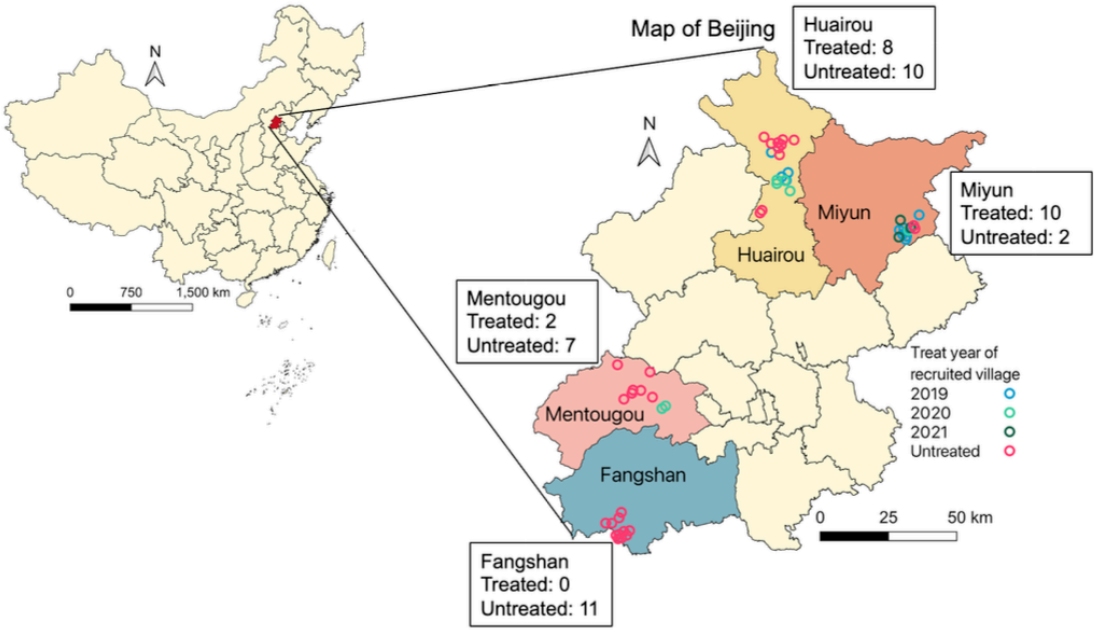


Figure 1: Map of village implementation of CHP. Each circle represents one recruited village. The colors of the circles indicate the year the villages were exposed to the household energy transition policy.

We recruited approximately 20 households in each village and, in each household, obtained a household roster. Our tablet-based survey incorporated a randomization tool than randomly ordered household occupants listed on the roster. We recruited a participant in each household by starting at the top of the randomly ordered list until an eligible participant was identified. Household members were eligible to participate if they were over 40 years old, lived in the study villages, were not planning to move out of the village in the next year, and were not on current immunotherapy or treatment with corticosteroids.

Research staff introduced the study and its measurements to an eligible adult in each household and answered any questions related to the study. In follow-up visits to the study villages, staff first approached households with participants from an earlier wave. Due to study logistics, we were limited to one day for study measurements in each village and wave, such that participants who were outside of the village on the measurement day for work or shopping were not able to participate in that wave. If a previous participant was not at home or refused to participate, staff first tried to randomly recruit the next eligible participant listed on the randomized household roster. If there was not another eligible or willing participant in the same household, we recruited a participant from a new household using the same process for household and participant selection described above. In Wave 2, we recruited 81 new participants from a previously enrolled household

and 189 new households. In Wave 4, we recruited 91 new participants from a previously enrolled household and 68 new households. Our village level study utilizes individual-level data such that each participant is considered independently.

All participants provided written informed consent prior to joining the study. The study protocols were approved by research ethics boards at Peking University (IRB00001052-18090), Peking Union Medical College Hospital (HS-3184) and McGill University (A08-E53-18B).

4.3 Data Collection Overview

We conducted study measurements over four consecutive waves of data collection in winter 2018-19, 2019-20, 2020-21, and 2021-22 (referred to hereafter as Wave 1 [W1], W2, W3 and W4, respectively). Field data collection was conducted by ~20 trained staff members who traveled to participants' homes to conduct tablet-based household and individual questionnaires, measure participant blood pressure, and distribute temperature sensors (for measurement of indoor temperature and stove use) and air pollution monitors in all 50 study villages in W1, W2, and W4. Anthropometrics (height, weight, and waist circumference), measurement of airway inflammation, and whole blood samples were obtained no more than a month later at a village clinic in W1 and W2. In W3 , which was during the height of the COVID-19 pandemic, we limited household measurements to indoor air quality and sensor-based measurement of indoor temperature and stove use in 41 villages, including all 17 treated villages and 24 untreated villages, prior to Beijing-wide COVID-19-related travel restrictions that halted field data collection. In W4, which also occurred during the COVID-19 pandemic, we returned to conducting individual-level assessments. However, unlike in W1 and W2, anthropometric measurements and airway inflammation were assessed in participant homes rather than in clinics to avoid group contact, and blood samples were not collected. Outdoor (community) air pollution was measured in all waves.

4.3.1 Air Pollution

Outdoor air pollution

For outdoor (community) $PM_{2.5}$ monitoring, we deployed between one to three (typically, two) real-time sensors (PMS7003 Plantower, Zefan, Inc.) at different locations in each village. The sensors were assembled with a data logger, electronic screen, and a USB hub into a small metal box that was placed inside an environmental enclosure. One sensor was always placed near the center of the village, and the other one or two sensors were placed no less than 500m away from the centrally-located sensor. Sensors were positioned at least 1.5m above the ground and away from visible point sources of $PM_{2.5}$.

We co-located the real-time sensors with a gravimetric (filter-based) monitor for sensor calibration and analysis of chemical composition for source apportionment. Ultrasonic Personal Aerosol Samplers (UPAS, Access Sensor Technologies, Fort Collins, CO, USA) were used to collect filter-based $PM_{2.5}$ samples with a flow rate of 1.0 L/min (Volckens et al. 2017). The filter-based $PM_{2.5}$ samples were collected and replaced approximately every seven days throughout the winter, and we rotated the UPAS monitors between villages. Each UPAS was placed inside a custom-built environmental enclosure with a tight fit to prevent any resampling of filtered air. Samplers housed 37mm PTFE filters (VWR, 2.0µm pore size) and were equipped with a cyclone inlet with a 2.5µm cut point designed to perform under the sampling flow rate.

In W1, we deployed co-located outdoor PM~2.5 sensors and samplers in 44 of the 50 study villages due to logistical constraints, and obtained sensor data for 40 villages due to instrument failure in 4 villages. Outdoor data for all 50 villages were obtained in waves 2, 3 and 4. In total, we collected 138, 374, 279, and 295 outdoor $PM_{2.5}$ filter samples in W1, W2, W3, and W4, respectively. Field blank PTFE filters were collected at a rate of ~10%, subject to the same field conditions as samples.

To support post-sampling determination of organic carbon (OC) and elemental carbon (EC) fractions of $PM_{2.5}$ mass, quartz filters were co-located with a subset of Teflon filter samples collected outdoors. Quartz filter-based $PM_{2.5}$ samples were collected using UPAS operating with a flow rate of 1.0 L/min. UPAS monitors housed 37 mm quartz filters (VWR, 2.0-µm pore size) and were equipped with a cyclone inlet with a 2.5µm cut point designed to perform under the corresponding sampling flow rate. All quartz fiber filters were baked at 550 °C for a minimum of 8 h to remove organic impurities prior to sample collection. The $PM_{2.5}$ samples collected on quartz filters were analyzed using established thermo-optical methods for quantifying elemental carbon (EC) and organic carbon (OC) to, then, calibrate the colorimetric analysis of EC and OC on Teflon filters (details of this analysis and subsequent calibration are provided in "Optical properties and chemical analysis of $PM_{2.5}$ mass"). We co-located quartz filters with Teflon filter samples for 23 measurements in W2 and 11 measurements in W4, along with 3 quartz field blanks in both seasons.

Figure 2: Calibration of real-time sensors against a reference monitor at University of the Chinese Academy of Sciences.

Indoor $PM_{2.5}$

In study waves 2, 3 and 4, we randomly selected six households from the ~20 recruited in each village for measurement of indoor $PM_{2.5}$. In W3 and W4, we aimed to monitor indoor $PM_{2.5}$ in the same households sampled in W2. If household occupants were not at home or if participants declined indoor $PM_{2.5}$ monitoring, we randomly recruited another household already enrolled in the study. In total, indoor $PM_{2.5}$ was measured in 264 households in W2, 346 households in W3, and 244 households in W4 (Table 1).

Time-resolved indoor $PM_{2.5}$ was measured in all households using the same commercially available sensor (PMS7003 Plantower, Zefan, Inc.) used for outdoor sensor-based $PM_{2.5}$ measurements and recorded $PM_{2.5}$ concentrations every 1 min. The sensor was placed on a table in a room where participants reported spending most of their time, e.g., a living room or bedroom. Indoor $PM_{2.5}$ sensors were deployed between late November and mid January in each wave, with the start time depending on the village visit date. Measurements continued from the time of deployment until sensors were collected from homes in late April.

Table 1: Household recruitment for overall and indoor air quality measurements.

	Overall			Indoor			
Sample	Wave 1	Wave 2	Wave 4	Wave 1	Wave 2	Wave 3	Wave 4
New recruitment	977	196	68	0	300	0	52
Wave 1 households	-	866	780	-	0	0	0
Wave 2 households	-	-	162	-	-	246	248
Total recruitment	977	1062	1010	0	300	246	300

We randomly selected three households from the six with indoor $PM_{2.5}$ measurement to co-locate a filter-based $PM_{2.5}$ sampler. We collected a 24-h $PM_{2.5}$ filter sample during the first 24-h of sensor-based measurement. Filter-based $PM_{2.5}$ samples were collected using UPAS or Personal Exposure Monitors (PEMs, Apex Pro) operating with flow rates of 1.0 and 1.8 L/min, respectively. Both samplers housed 37 mm PTFE filters (VWR, 2.0- m pore size) and were equipped with a cyclone inlet with a 2.5 m cut point designed to perform under the corresponding sampling flow rate. In total, we collected 150 and 151 indoor $PM_{2.5}$ filter samples in W2 and W4, respectively. We did not measure filter-based indoor $PM_{2.5}$ in S3 to avoid contact with household occupants during the COVID-19 pandemic. Field blanks were collected at a rate of approximately 10%.

As with the outdoor air sampling, to support post-sampling determination of organic carbon (OC) and elemental carbon (EC) fractions of $PM_{2.5}$ mass, quartz filters were co-located with a subset of Teflon filter samples collected in homes. Filter-based $PM_{2.5}$ samples were collected using Personal Exposure Monitors (PEMs, Apex Pro) operating with flow rates of 1.8 L/min. PEMs housed 37 mm quartz filters (VWR, 2.0µm pore size) and were equipped with a cyclone inlet with a 2.5µm cut point designed to perform under the corresponding sampling flow rate. All quartz fiber filters were baked at 550°C for a minimum of 8 h to remove organic impurities prior to sample collection. $PM_{2.5}$ samples collected on quartz filters were analyzed using established thermo-optical methods for quantifying EC and OC to, then, calibrate the colorimetric analysis of EC and OC on Teflon filters. In W2, 71 quartz-based indoor $PM_{2.5}$ samples and 14 field blanks were successfully collected. In W4, indoor $PM_{2.5}$ samples for gravimetric analysis had to be collected on two types of PTFE sample media (Zefluor and Teflo filters), due to the discontinuation of Zefluor filters. To ensure that quartz filters were deployed with both types of Teflon-based filter media, 73 quartz-based indoor $PM_{2.5}$ samples were collected concurrently with Zefluor samples, and 47 quartz indoor $PM_{2.5}$ samples were collected alongside Teflo samples. For indoor quartz $PM_{2.5}$ mass sampling in W4, 18 field blanks were collected.

Personal exposure to $PM_{2.5}$ and black carbon

In study waves 1, 2 and 4, we randomly selected approximately ten study participants in each village for 24-h personal exposure measurement using two types of $PM_{2.5}$ samplers: PEMs and UPAS. The PEMs actively sampled air at a flow rate of 1.8 L/min, and UPAS sampled air at 1.0 L/min (Volckens et al. 2017). Both samplers housed 37 mm PTFE filters (VWR, 2.0µm pore size) and were equipped with a cyclone inlet with a 2.5µm cutpoint. Sampler flow rates were calibrated the night before deployment and measured immediately after the sampling period. Only 2% of the post-sampling measurements deviated from the target flow rate by greater than +/-10%. Participants were instructed to wear the sampler in either a small waistpack (for the PEM and sampling pump), an arm band, or a cross-body sling (for the UPAS) for 24-h, which they could remove from their body and place within 2 meters while sleeping, sitting, or bathing. Field blanks for personal air pollution exposure measurements were collected at a rate of ~10% in each village. Across the study waves 1, 2 and 4, study participants contributed 494, 498, and 499 personal $PM_{2.5}$ measurements, respectively.

Table 2: Count of total outdoor and personal exposure $PM_{2.5}$ samples (filters) collected over the course of the project and number included for analysis.

	Wave 1		Wave 2		Wave 3		Wave 4	
PM2.5 sample type	Total	Included[a]	Total	Included[a]	Total	Included[a]	Total	Included[a]
Outdoor	138	126	374	363	279	213	295	266
Indoor			150	150			151	138
Personal	494	448	498	429			499	418
Blank	52	52	56	56	27	24	101	95

[a] Number of samples that met inclusion criteria for analysis (see text).

Gravimetric analyses of PTFE filter-based $PM_{2.5}$ samples

All filters were gravimetrically analyzed (weighed) for their mass before and after deployment at a laboratory at Colorado State University. Briefly, the filters were placed in an environmentally-controlled equilibration chamber (21-22°C, 30-34% relative humidity) for at least 24-h before tare and gross weighing. Before each weight was taken, we neutralized static charges by passing the filters over a polonium-210 strip. Filters were weighed on a microbalance (Mettler Toledo Inc., XS3DU, USA) with 1 µg resolution in triplicate or more, until the differences among the last three weights were less than 3 g. The filters were stored in individually labeled cases and sealed in plastic bags to avoid contamination during transportation and storage. After deployment, filter samples and blanks were immediately stored in a -20°C freezer and, at the end of each field campaign, were transported to Colorado State University, where they were stored in a -20°C freezer prior to gravimetric and chemical analysis. The difference in the average filter weights from before versus after deployment was used to determine $PM_{2.5}$ mass, which was then blank-corrected using the

median value of blank filters (3 µg for UPAS-collected filters [53% of filter samples]; 33µg for PEM-collected filters [47% of filter samples]), and $PM_{2.5}$ concentrations were calculated by dividing the mass by the sampled air volume.

We excluded gravimetric (filter) samples meeting any of the following four criteria from the statistical analysis: (1) run time of less than 80% of a 24-h target for personal exposure measurement, as this is a commonly used cutoff for establishing whether a sample is considered representative of a typical day; (2) negative mass or extremely high mass values (e.g., > 2000 µg/m^3) that indicate a potential error in data collection or data entry; (3) missing information on sampling volume of air; (4) filters were damaged including punctures, tears, or holes; or (5) filters were missing during gravimetric analysis and therefore had no mass data. The final counts of gravimetric $PM_{2.5}$ samples and blanks that met our criteria for inclusion in statistical analysis are given in Table 2.

Adjusting sensor-based $PM_{2.5}$ using filter-based gravimetric measurements

Pre- and post- campaign sensor calibration was conducted to assess whether low-cost sensors responded linearly to $PM_{2.5}$ concentrations measured by co-located federal equivalent method (FEM) instruments. Sensors were deployed alongside the FEM instruments for 7-10 days before and after each field campaign. In waves 1, 2 and 3, we co-located the sensors with a rooftop Thermo Electron Synchronized Hybrid Ambient Real-Time Particulate (SHARP) Monitor (model 5030) at Peking University, which is a typical urban site (urban site). In waves 2, 3 and 4, we additionally co-located the sensors with a rooftop Tapered Element Oscillating Microbalance Method (TEOM, Thermo Scientific™ 1405 TEOM™) at the University of the Chinese Academy of Sciences, which is located in a peri-urban area of Beijing (peri-urban site) where we also have study villages. The FEM instrument at Peking University was not functioning after wave 1 data collection, so we instead calibrated the sensors using data from the nearest China National Environmental Monitoring Centre (CNEMC) monitor (publicly available here). The closest distance from the government monitoring stations to Peking University and Chinese Academy of Sciences University campuses are 1.7 and 9.9 km, respectively.

We evaluated the performance of all $PM_{2.5}$ sensors for the 7-10 day deployments described above that occurred before and after each study wave (Figure 2). Sensor-measured $PM_{2.5}$ values were highly correlated with the FEM instruments (Spearman correlation (rho) >0.75 in all pre- and post-calibration campaigns). Daily collections of 24-hour Zeflour (Teflon) and quartz filter samples accompanied the sensors' measurements. For pre- and post-campaign calibration periods, as described above, the filter-based data were supplemental to the FEM data; whereas sensor calibration during field campaigns, as described below, was achieved using calibration with concurrently collected filter-based data (since FEM references were not available in the field). We also monitored the sensor-collected data throughout each study wave to identify sensors in need of repair or replacement (e.g., logging data with the wrong time stamps or only "0" values) and exclude them from deployment. This approach aimed to maintain consistent and accurate measurements from the PM sensors throughout the study.

Sensor calibration during each wave was conducted by deploying filter-based measurements concurrently with sensor-based measurements, to establish the linear regression between the low-cost sensor monitored data and the reference data, and then apply the slope of the linear regression to adjust the low-cost sensor monitored data. To calibrate the outdoor and indoor measurements of the low-cost sensors in the field, we collocated an Ultrasonic Personal Aerosol Samplers (UPAS, Access Sensor Technologies) (Volckens et al. 2017) or Personal Exposure Monitors (PEMs, Apex Pro) with low-cost sensors to collect filter-derived $PM_{2.5}$ samples during each field season (wintertime) (Li et al. 2022). The UPAS and PEM were equipped with a cyclone inlet with a 2.5 m cut point designed to perform under the sampling flow rate of 1 and 1.8 L/min, respectively, and housed a 37 mm PTFE filter (VWR, 2.0- m pore size). The filter samples were transported to Colorado State University, where they were stored in a -20 °C freezer prior to $PM_{2.5}$ mass measurement.

We established linear regression models between the filter-based $PM_{2.5}$ mass (i.e., the 'gold standard' reference) and the sensor-based $PM_{2.5}$ averaged over the same sampling periods. Separate regression models were conducted for indoor and outdoor sensors and for each study wave given the sensitivity of the sensors to relative humidity, temperature, and particle sources, which may differ for indoor versus outdoor conditions and across years. The model slopes were used as the adjustment factors for the sensor-based $PM_{2.5}$ concentrations for that wave. In W3, where only sensor-based measurements were conducted for indoor $PM_{2.5}$, we applied an adjustment factor that was developed from paired indoor filter-sensor data from W2 and W4.

We identified larger than normal biases and root mean square errors (RMSE) between the sensors and FEM instruments during the post-W3 calibration, however further scrutiny of these data indicate that the differences can be attributed to atypically low air pollution and high humidity during co-location rather than a malfunction of the sensors themselves. Sensor calibration results directly informed the data correction processes applied to account for biases. Generally, the higher correlations between the $PM_{2.5}$ sensor response and FEM data allowed for the application of linear regression models to adjust sensor measurements, and filter-based gravimetric samples were used to validate and correct the sensor data.

We used the adjusted sensor-based $PM_{2.5}$ measurements to calculate a wintertime seasonal mean for indoor and outdoor $PM_{2.5}$ for the period of January 15 to March 15 in each wave to facilitate consistent comparisons across villages in each wave and over time. Additionally, we captured a 24-hour indoor $PM_{2.5}$ concentration that was temporarily matched with the timing of personal exposure assessments in the same household to facilitate a comparison between indoor $PM_{2.5}$ and personal exposure results taken during the same period.

Optical properties and chemical analysis of $PM_{2.5}$ mass

We analyzed the optical properties and chemical composition of outdoor and personal exposure gravimetric $PM_{2.5}$ samples to quantify the individual components and species. For each sample,

the components were determined by dividing the quantified component mass by the sampled air volume, after correcting for field blanks collected in the corresponding study wave.

Following gravimetric analysis, all PTFE filters were analyzed non-destructively for black carbon (BC) using an optical transmissometer data acquisition system (SootScan^TM OT21 Optical Transmissometer; Magee Scientific, Berkeley, CA, USA). Light attenuation through each filter was measured before and after deployment in the field campaign. To calculate BC mass, the difference between the pre- and post- light attenuation was converted to a mass surface loading using the classical Magee mass absorption cross-sections of 16.6 m^2/g for the 880 nm channel optical BC (Ahmed et al. 2009). BC concentrations were calculated by multiplying surface loadings by the sampled surface area of the filters (8.6 cm^2 for UPAS-collected filters; 7.1 cm^2 for PEM-collected filters), correcting for the field blank mass using the median value of blanks (0.31 g for UPAS-collected filters; 0.01 g for PEM-collected filters), and finally dividing by the sampled air volume.

Organic (OC) and elemental carbon (EC) on PTFE filters were non-destructively measured using an optical color space sensing system. The CIE-Lab color space optical sensing system measures the optical properties of the $PM_{2.5}$ samples which are used to develop EC and OC predictive models. The CIE-Lab color system is a color-opponent space that includes all of the color models, with dimension L* for lightness and a* and b* for the color-opponent dimensions. More information about the CIE Lab color space system, its formulation, and its specific application to the analysis of OC and EC fractions of fine particulate matter pollution is provided in Khuzestani et al. (Khuzestani et al. 2017). Briefly, all PTFE and quartz filters samples and blanks were analyzed using the i1Pro Colorimeter (X-Rite, INC. Grand Rapids, MI). The colorimeter sensor was placed directly over the filters, and the color components were measured under the D65 instrument internal illumination light source. Each filter sample was analyzed in triplicate, and the average value of each color coordinate was applied as the optical property of the sample (Olson et al. 2016). CIE Standard Illuminant D65 simulates average midday light and is a commonly used standard illuminant, as defined by the International Commission on Illumination (CIE). The CIE-Lab color space response variables were used in separate random forest models for EC and OC.

The reference measurements for the random forest model development were EC and OC measured on quartz filters collected both in study homes and outdoors (as described above). The quartz filters were analyzed for OC and EC with a Sunset Laboratory OC/EC Lab instrument (Sunset Laboratories, Inc., MODEL, USA) using the default Sunset Analyzer protocol. A section of each quartz filter underwent a combined thermal desorption-optical transmittance measurement based on NIOSH methods 5040 to differentiate and quantify the EC and OC components in $PM_{2.5}$ mass. For the thermal desorption component, the filter is oxidized twice using a strict temperature regime. The first oxidation stage thermally removes OC in a mobile phase of pure helium gas that is converted from carbon dioxide (CO_2) to methane (CH_4) gas and measured by a flame ionization detector (FID). The second oxidation stage proceeds in a mixture of helium and oxygen to oxidize EC, which is also quantified by the FID. The FID is internally calibrated with methane, and external quality control checks are made with sucrose standards. To correct for the potential production of

EC during OC pyrolysis in the first oxidation stage, light transmission from a laser through the filter section was monitored throughout analysis. Reduced light transmittance corresponds to EC generated by the laboratory analysis.

Elemental analysis of $PM_{2.5}$ mass was performed using a Thermo Scientific Quant'X Evo energy-dispersive X-ray fluorescence (EDXRF) spectrometer with Wintrace software version 10.3 using standard methods (RTI International 2009). Quantitative mass concentrations of 22 individual elements (Mg, Al, Si, S, K, Ca, Ti, Cr, Mn, Fe, Ni, Cu, Zn, Ga, As, Se, Cd, In, Sn, Sb, Te, I) were determined empirically using linear standard curves. Standard curves were generated from commercial, single and dual element, thin film standards from MicroMatter Technologies Inc. (Montreal, Canada) in addition to blank films. The quality of the analysis method was evaluated by analyzing a National Institute of Standards and Technology (NIST) standard reference material (SRM) 2783 Air particulate on filter media (Gaithersburg, MD, USA). Elements for which at least 80% of $PM_{2.5}$ mass samples yielded quantifiable element mass were included for positive matrix factorization and for source analysis and apportionment. Those elements were: Si, Mg, Fe, S, Ca, Al, K, Pb.

For analysis of water-soluble ions, a portion of each PTFE filter was extracted in 15 mL deionized water (DI Water) in a Nalgene Amber HDPE bottle using sonication without heat for 40 min. The extracts were filtered to ensure that insoluble particles were removed using a 0.2 m PTFE syringe filter. Water-soluble ions were measured using a dual channel Dionex ICS-3000 ion chromatography system. Specifically, a Dionex IonPac CS12A analytical (3 × 150 mm) column with eluent of 20 mM methanesulfonic acid at a flow rate of 0.5 mL/min was used to measure cations (Ca^{2+}, Mg^{2+}, Na^{+}, NH^{4+}, K^{+}), while a Dionex IonPac AS14A analytical (4 × 250 mm) column with an eluent of 1 mM sodium bicarbonate/8 mM sodium carbonate at a flow rate of 1 mL/min was used to measure anions (SO_4^{2-}, NO^{3-}, Cl^{-}) (Sullivan et al. 2008).

- *wi (Water Insoluble Species)*: 'wi' refers to the fraction of particulate matter (PM) that does not dissolve in water. These species typically include elements such as potassium (K), calcium (Ca), and magnesium (Mg), among others, that remain as particulate matter after water extraction. In this study, due to budget constraints, we did not analyze water-soluble organic carbon. Instead, we determined the water insoluble fraction by subtracting the amount of the elemental species measured using ion chromatography (IC) from the total elemental amount determined by X-ray fluorescence (XRF). For instance, the total amount of potassium (K) was measured using XRF, and the water-soluble potassium fraction was quantified using IC. The difference between these two values was taken as the water insoluble fraction of potassium. This approach is consistent with practices in air quality research that use these analyses.
- *ws (Water Soluble Species)*: 'ws' refers to the fraction of particulate matter that dissolves in water, typically including major ions such as sulfate, nitrate, ammonium, and certain soluble forms of metals. The water soluble fraction was extracted from particulate samples using deionized water, and the extract was analyzed using ion chromatography (IC) to determine the concentrations of individual water soluble ions, such as sulfate (SO_4^{2-}), nitrate (NO^{3-}),

and soluble metal ions. This approach is well-documented in the scientific literature and follows established protocols for the determination of anions in PM.

- *ns-S (Non-Sulfate Sulfur)*: 'ns-S' refers to the sulfur present in particulate matter that is not in the form of sulfate – i.e., non-sulfate (ns) sulfur (S). This includes species such as elemental sulfur, organosulfur compounds, and other non-sulfate sulfur-containing compounds. Total sulfur content in particulate matter was determined using X-ray fluorescence (XRF), consistent with the method we employed for this project (RTI International 2009). The sulfate (SO_4^{2-}) content was quantified using ion chromatography. The non-sulfate sulfur (ns-S) was then calculated by subtracting the sulfate sulfur from the total sulfur determined using XRF analysis. This approach is supported by studies such as those by Shakya and Peltier (2015) and Secrest et al. (2016), which have utilized similar methodologies to distinguish sulfur sources in PM studies.

4.3.2 Outdoor and indoor (household) air temperature

Hourly outdoor temperature and relative humidity data were obtained from the extensive network of meteorological stations in Beijing. We used digital thermometers (Tianjianhuayi Inc., Beijing, China) to measure indoor 'point' temperature in the five minutes prior to BP measurement. Staff measured temperature in a centrally located room, away from heating sources and direct sunlight, by placing the probe in mid-air at a height that approximated the participant's shoulder height. In a random 75% subsample of households in each wave, we also conducted long-term measurements of indoor temperature by placing a real-time temperature sensor (iButton DS1921G-F5; Thermochron, Maxim Inc., USA) in the room where participants reported spending most of their daytime hours when indoors. Sensors were wall-mounted at a standardized height (~1.5 to 2 meters), away from major heating sources, windows, and doors, and were programmed to log a temperature reading every 125 minutes for up to 4 months to capture the full winter period and early spring weeks when heating may still intermittently occur. Prior to the start of each wave, we co-located all of the sensors and measured temperature over two days and compared the readings. Sensors recording values >1°C from the group median value were not deployed for data collection.

4.3.3 Objective measurement of household stove use using sensors

Following methods used in a previous intervention evaluation study in rural China (Clark et al. 2017), we objectively measured household heating stove use in a random sample of households selected, also at random, for either short- or long-term measurement. We measured short-term (24-h) stove use for all household heating stoves in 315 and 227 households in W2 and W3, respectively. Long-term stove use was assessed in 324, 273, and 585 homes in W2, W3, and W4, respectively, for a period of ~6 months. We measured stove use using the same real-time temperature data loggers used to measure seasonal indoor temperature (iButton DS1921G-F5; Thermochron, Maxim Inc.,

USA). Field staff placed the sensors on stoves and programmed them to record surface temperature every 125 minutes, a timing decision based on pilot assessments showing that shorter time intervals did not affect the number of heating events detected or heating time recorded. Sensors were placed on the surfaces of biomass and coal-fueled stoves and radiators. For heat pumps, sensors were placed on the heat exchanger coil on air-to-air units and on the radiator of air-to-water units.

The number and duration of stove combustion events were identified from the temperature data using criteria defined based on the observed changes in the peak shape of the time series temperature curves (i.e., changes in the slope or in absolute temperature compared with the indoor ambient temperature). This approach was specific to heating stoves but developed based on stove use identification for cookstoves in previous studies by us and others (Clark et al. 2017; Ruiz-Mercado et al. 2013; Snider et al. 2018). We developed separate criteria for each stove type given the observed stove-specific differences in heating patterns. These criteria were coded into stove-specific algorithms to systematically identify the number and duration of heating events across households. A stratified random sample of stove use temperature files (15% for each stove type and measurement duration - short-term/24-h or long-term/~6 month - combination) were manually coded to develop the test criteria. The number and duration of heating events were identified by the algorithms in the remaining 85% of files. We compared heating periods identified manually with those identified by the algorithm to check for systematic differences and possible overfitting.

4.3.4 Questionnaires

Field staff administered household and individual-level questionnaires to assess household demographic information and educational attainment, household assets, house structure, stove and fuel use patterns (including a complete roster of heating methods and their contributions in each room), and individual health behaviors including exercise frequency, smoking, alcohol consumption, medication use, and clinician-diagnosed health conditions. We used Surveybe computer-assisted personal interview (CAPI) software to collect survey data via handheld electronic tablets. Questions were read to participants in Mandarin-Chinese, and their responses were recorded into tablets.

Prior to the start of data collection, all questions were translated from English into Chinese and then back-translated to English for quality assurance. Many questions were adapted from previous field studies of household energy and blood pressure conducted in rural Beijing or other rural sites in China (Baumgartner et al. 2018; Yan et al. 2020), and all questions were iteratively tested with study staff and adapted before implementation. Prior to each wave in this study, the questionnaire and all other study measurements were tested in 12 households located in a Beijing village that was eligible for our study but was instead selected for testing. We used the test village to train study staff, assess whether the questions were understandable and being interpreted as intended, and to identify any problems with the study measurements or their implementation. Study protocols were subsequently adapted prior to the start of data collection.

In addition to household and individual participant questionnaires, we also conducted village surveys with one representative from each village committee to understand how the policy was implemented in that village and to inquire about any other rural development or health programs being implemented in the village. Committee members answered questions about committee and villager interest in the policy and, for treated villages, assignment versus application to the policy, any home or village renovations required by the upper-level government prior to heat pump installation, decision-making for the type and brand of heating technology, level of subsidies provided for heaters and electricity, and technical and logistical guidance to villagers.

4.3.5 Blood pressure

Following 5 min of quiet rest, at least three brachial and central systolic (bSBP/cSBP) and diastolic (bDBP/cDBP) blood pressures (BPs) were taken by trained staff at 1 min apart on the participant's supported right arm. We used an automated oscillometric device (BP+; Uscom Ltd, New Zealand) that estimates central pressures from the brachial cuff pressure fluctuations. Central pressures were previously validated against invasive cBP measurements in earlier studies (Costello et al. 2015; Lowe et al. 2009). The BP devices were factory calibrated by the manufacturer prior to the start of the first and fourth waves. Up to five measurements were taken if the difference between the last two was >5 mmHg or staff were unable to obtain a reading. The BP measurements were conducted in the participant's home and staff were trained to follow strict quality control procedures, including use of an appropriately sized cuff, correct positioning of the arm, both feet on the ground, and ensuring 5 min of quiet rest before measurement. Details are described in the standard operating procedures (SOP). The average of the final two measurements was used for statistical analysis unless only one BP measurement was obtained (n = 13 observations), in which case a single measurement was used. The time of day, day of the week, and indoor temperature prior to BP measurement were also recorded.

4.3.6 Self-reported respiratory symptoms and airway inflammation

During questionnaire assessment, participants were asked about chronic airway symptoms including cough, phlegm, wheeze, and tightness in the chest using questions validated for use in Mandarin-Chinese and developed from the standard St. George's Respiratory Questionnaire (Xu et al. 2009). The Mandarin-Chinese questions were extensively piloted with rural and peri-urban Beijing residents to ensure that the health terminology and symptom time patterns were adequate and understandable to the local population.

In a ~25% random subsample of participants, we also measured the fractional concentration of exhaled nitric oxide (FeNO), a non-invasive and established marker of airway inflammation, using a portable handheld device (Aerocrine, Solna, Sweden) fit with a NIOX VERO® sensor, following ATS recommendations and guidelines (ATS/ERS 2005). Briefly, FeNO measurement was performed

with participants in a standing position. They inhaled NO-free air through a mouthpiece with an NO-scrubber attached, followed by controlled expiration for 10 s through the mouthpiece at 50±5 mL/s. A nose clip was used to avoid nasal inhalation, and accurate flow rate was achieved using visual and auditory cues generated by the device. Detailed methods are provided in our previous study of air pollution and FeNO in Beijing adults (Shang et al. 2020). At least two measurements were obtained for each participant.

4.3.7 Blood inflammatory and oxidative stress markers

Trained nurses collected 20 ml of whole blood in a labeled vacutainer via venipuncture using standard techniques (Tuck et al. 2009). Details are described in our published SOP. Briefly, fasting blood samples were collected by experienced phlebotomists (nurses) in the morning and stored at 4-10°C prior to centrifugation. Two serum aliquots from each participant were then placed in a -30°C freezer for temporary storage. Collection-to-storage time was <4 hrs for all samples in both waves where blood samples were collected. Within 3-5 days of collection, the samples were transported in styrofoam containers with dry ice to a -80°C freezer with a backup generator and alarm system at Peking University.

The first aliquot was analyzed for glucose and a complete lipid profile within two months of collection, and results were communicated to participants. The second aliquot was stored in the -80°C freezer for analysis of biomarkers of systemic inflammation (C-reactive protein [CRP], interleukin-6 [IL-6], tumour necrosis factor alpha [TNF-α] and malondialdehyde [MDA]) at the University of the Chinese Academy of Sciences between July and September of 2023. These biomarkers were selected because they are associated with the development of cardiovascular disease and events (e.g., Danesh et al. 2008; Emerging Risk Factors Collaboration 2012; Pearson et al. 2003; Ridker 2001; Ridker et al. 2000), and both acute and longer-term exposures to air pollution have been associated with changes in inflammatory and oxidative stress markers (e.g., Huang et al. 2012; Kipen et al. 2010; Pope III et al. 2004; Rich et al. 2012; Rückerl et al. 2007).

We followed standard methods for analysis (Food and Drug Administration 2018). For inflammatory markers (IL-6, TNF-α, CRP), the optic densities (OD) of all samples were measured using an automated ELISA reader. Every plate had 8 standard samples used to generate a standard curve that related OD and standard inflammatory marker concentration. A standard curve for each microplate was generated by a computer software program based on a 4-parameter method. Each plate included at least 3 control samples to ensure the stability of standard curves. All samples , standards, and controls were measured in duplicate, and the average was used for statistical analysis. For oxidative stress biomarkers (MDA), the chromatographic peak areas of all samples were measured using HPLC with UV detector and HPLC-MS/MS. Every plate had 7 standard samples used to generate a standard curve that related peak area and concentration of the oxidative stress marker. A standard curve for each plate was generated using a computer software program based on a linear method. Each plate included at least 3 control samples to ensure the stability

of standard curves. Standards and controls for MDA were measured in duplicate and samples were measured once due to high precision in our pre-analysis pilot study with duplicate testing and evidence from many previous studies showing high stability in measurement (Food and Drug Administration 2018).

4.3.8 Anthropometric measurements.

Body weight, height, and waist circumference were measured at the clinic visit in W1 and W2 and in participant homes in W4 to avoid unnecessary contact during the COVID-19 pandemic. Weight was measured in light indoor clothing without shoes in kilograms to one decimal place, using standing scales supported on a steady surface. The scales were calibrated prior to the start of each wave, and the same staff member weighed themself on the scale each morning to ensure that it was functioning properly. Height was measured without shoes in centimeters to one decimal place with a stadiometer. Waist circumference was measured without clothing obstruction at one centimeter above the participant's navel at minimal respiration in centimeters to one decimal place. The measuring tapes were replaced at the start of each wave to avoid stretching.

4.4 Measuring policy impacts

To understand how Beijing's policy works we used a difference-in-differences (DiD) design (Callaway 2020), leveraging the staggered roll-out of the policy across multiple villages to estimate its impact on health outcomes and understand the mechanisms through which it works. Simple comparisons of treated and untreated (i.e., control) villages after the CHP has been implemented are likely to be biased by unmeasured village-level characteristics (e.g., migration, average winter temperature, wealth) that are associated with health outcomes. Similarly, comparisons of only treated villages before and after exposure to the program are susceptible to bias by other factors associated with changes in outcomes over time (i.e., secular trends, potential health impacts of the COVID-19 pandemic). By comparing *changes* in outcomes among treated villages to *changes* in outcomes among untreated villages, the DiD approach controls for any unmeasured time-invariant characteristics of villages as well as for any general secular trends affecting outcomes in all villages that are unrelated to the policy.

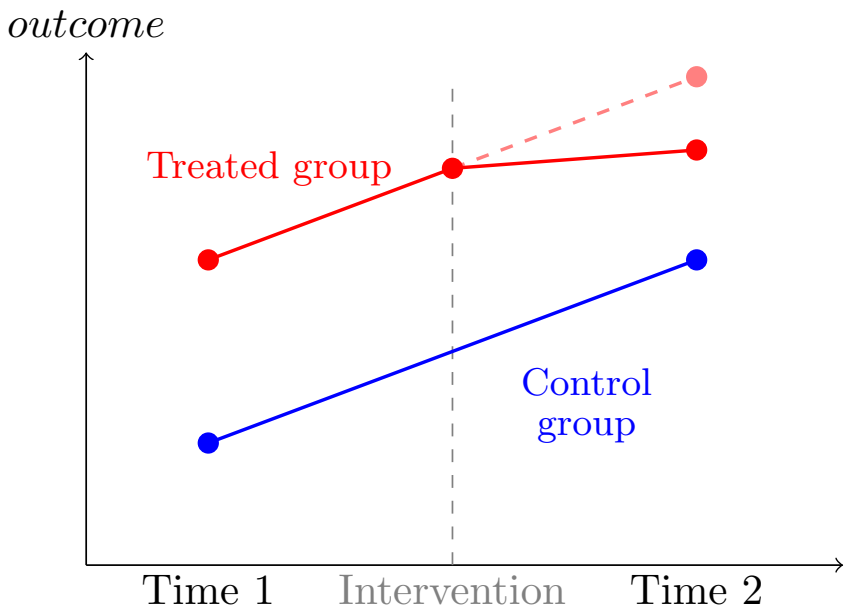


Figure 3: Stylized example of difference-in-differences

The DiD design compares outcomes before and after an intervention in a treated group relative to the same outcomes measured in a control group. The control group trend provides the crucial "counterfactual" estimate of what *would have happened* in the treated group had it not been treated. By comparing each group to itself, this approach helps to control for both measured and unmeasured fixed differences between the treated and control groups. By measuring changes over time in outcomes in the control group unaffected by the treatment, this approach also controls for any unmeasured factors affecting outcome trends in both treated and control groups. This is important since there are often many potential factors affecting outcome trends that cannot be disentangled from the policy if one only studies the treated group (as in a traditional pre-post design).

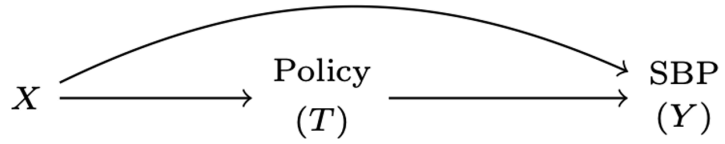
The canonical DiD design (Card and Krueger 1994) compares two groups (treated and control) at two different time periods (pre- and post-intervention, Figure 3). In the first time period both groups are untreated, and in the second time period one group is exposed to the intervention. If we assume that the differences between the groups would have remained constant in the absence of the intervention (the parallel trends assumption), then an unbiased estimate of the impact of the intervention in the post-treatment period can be calculated by subtracting the pre-post difference in the untreated group from the pre-post difference in the treated group. The estimand of interest in a typical DiD analysis is the average treatment effect on the treated (i.e, the ATT), which is a contrast of the post-intervention outcomes in the treated group with the counterfactual estimate of outcomes in the same population in the absence of treatment.

When multiple groups are treated at different time periods, as with the CHP, the most common approach has been to use a two-way fixed effects model to estimate the impact of the intervention which controls for secular trends and differences between villages. However, recent evidence suggests that traditional two-way fixed effects estimation of the treatment effect may be biased in the context of heterogeneous treatment effects, i.e., where the effects of treatment vary for different groups treated at different time periods (Callaway and Sant'Anna 2021; Goodman-Bacon 2021). The bias is due to the fact that when there are multiple groups treated at different times the two-way fixed effects estimate is a weighted average of several '2 x 2' DiD estimates, some of which involve using already treated units as controls for later treated units, which can lead to bias (Baker et al. 2022). We take advantage of new developments in the econometrics literature (Callaway and Sant'Anna 2021; Sun and Abraham 2021; Wooldridge 2021) that relax the assumption of homogeneity in the context of staggered policy roll-outs but also allow straightforward interpretation of ATTs for assessing policy impacts. This decision was motivated by the many behavioral, social, or economic factors that might affect both new heat pump use and coal stove suspension (e.g., energy prices and availability, wintertime temperature, COVID-19 pandemic, user preferences) over time in our study, and thus the possibility that the effect of the policy on air pollution and health may be dynamic over time and/or heterogeneous across treatment cohorts.

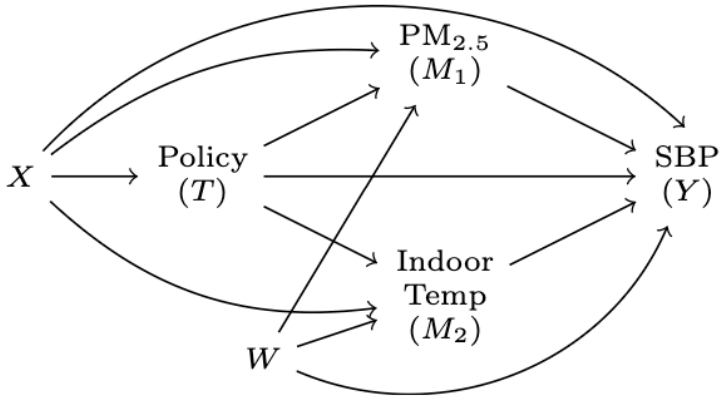
4.5 Measuring pathways and mechanisms

To estimate how much of the CHP may work through different mechanisms, we used causal mediation analysis. Causal approaches to mediation attempt to discern between, and clarify the necessary assumptions for identifying, different kinds of mediated effects. Figure 4 shows directed acyclic graphs (DAGs) to illustrate the total effect (a) and the potential direct and indirect effects of the CHP (b), with T as the policy, X as a set of pre-treatment covariates, and Y as systolic blood pressure as an example outcome. The 'total effect' of the policy is an estimate of how much the overall outcomes (Y) would change for a change in exposure (versus no exposure) to the CHP (T). Part (b) of Figure 4 adds M_1 as $PM_{2.5}$ and M_2 as indoor temperature as potential mediators that are affected by the policy, and we can define the controlled direct effect (CDE) as the effect of the CHP on systolic blood pressure if we fix the values of $PM_{2.5}$ and indoor temperature to a fixed reference level for the entire population. For example, we can estimate the impact of the policy on health outcomes while holding $PM_{2.5}$ and indoor temperature at uniform levels of average background exposure, or some other hypothetical level (VanderWeele 2015).

Figure 4: Hypothetical Directed Acyclic Graphs (DAGs) showing (a) total effect and (b) direct and indirect effects with outcome (Y), pre-treatment covariates (X), policy (T), multiple mediators (M_1, M_2), as well as covariates for the mediators (W).



(a) DAG for Total effect



(b) DAG for direct and indirect effects

Although other mediated effects such as "natural" direct and indirect effects are theoretically estimable (VanderWeele 2015), they involve challenging "cross-world" assumptions that are difficult to anchor in policy (Naimi et al. 2014). Other approaches to mechanisms have focused on principal stratification (e.g., Zigler et al. 2016), although conceptual difficulties with identifying the (unverifiable) principal strata make it challenging for questions of mediation. Because controlled direct effects are considered more directly policy relevant for public health, we focused on estimating these mediated quantities.

5 Data Analysis

To understand how the policy's impact on health may be mediated by different potential mediators, we need to first estimate the total effect of the policy on the outcomes (shown in Figure 4 part (a), then estimate the CDEs after adjustment for potential mediators and any residual mediator-outcome confounding. As discussed above, in order for the mediators to 'explain' the total effects

of the policy on health, the policy should affect the mediators, and the mediators should also affect the outcomes.

5.1 Total Effect

To estimate the total effect of the policy we used a DiD analysis that accommodates staggered treatment roll-out. To allow for heterogeneity in the context of staggered roll-out we used 'extended' two-way fixed effects (ETWFE) models (Wooldridge 2021) to estimate the total effect of the CHP. The mean outcome (replaced by a suitable link function $g(\cdot)$ for binary or count outcomes) was defined using a set of linear predictors:

$$Y_{ijt} = g(\mu_{ijt}) = \alpha + \sum_{r=q}^{T} \beta_r d_r + \sum_{s=r}^{T} \gamma_s fs_t + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}(d_r \times fs_t) + \varepsilon_{ijt} \quad (1)$$

where Y_{ijt} is the outcome for individual i in village j at time t, d_r represent treatment cohort dummies, i.e., fixed effects for each cohort of villages r that were first exposed to the policy at the same time q (e.g., in 2019, 2020, or 2021), fs_t are fixed effects for each time period s corresponding to different winter data collection waves (2018-19, 2019-20, or 2021-22), and τ_{rs} are the cohort-time ATTs in the context of a linear model. For non-linear outcomes the cohort-time ATTs are derived by estimating marginal effects from non-linear models (Arel-Bundock 2024). For binary and count outcomes we used logit and Poisson models, respectively, and for skewed outcomes (e.g., black carbon, $PM_{2.5}$, inflammatory markers) we used generalized linear models with a Gamma distribution and a log link, based on the specification tests recommended by Manning and Mullahy (2001). For all models we clustered standard errors at the village level, consistent with the unit of treatment assignment (Cameron and Miller 2015). The ETWFE and other approaches that allow for several (potentially heterogeneous) treatment effects may also be averaged to provide a weighted summary ATT. Several potential possibilities are feasible, including weighting by treatment cohorts or time since policy adoption (Goin and Riddell 2023). We generally focus on two types of ATTs for this report: simple averages across all treatment cohorts and the full set of cohort-time ATTs to evaluate heterogeneous treatment effects. Although we primarily focus on reporting the simple average ATT for most outcomes, we also used omnibus joint F-tests to assess whether there was sufficient evidence to reject the assumption of homogeneity across the ATTs.

5.2 Mediation Analysis

As noted above, with respect to the mediation analysis we are chiefly interested in the CDE, which can be derived by adding relevant mediators M to Equation 1. If we also allow for exposure-mediator interaction and potentially allow for adjustment for confounders W of the mediator-outcome effect, we can extend equation Equation 1 as follows:

$$
\begin{aligned}
Y_{ijt} = g(\mu_{ijt}) = \alpha + \sum_{r=q}^{T} \beta_r d_r + \sum_{s=r}^{T} \gamma_s fs_t + \sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}(d_r \times fs_t) \\
+\delta M_{it} + \sum_{r=q}^{T} \sum_{s=r}^{T} \eta_{rs}(d_r \times fs_t \times M_{it}) + \zeta \mathbf{W} + \varepsilon_{ijt}
\end{aligned} \tag{2}
$$

where now δ is the conditional effect of the mediator M at the reference level of the treatment (again, represented via the series of group-time interaction terms), and the collection of η_{rs} terms are coefficients for the product terms allowing for mediator-treatment interaction. Finally, ζ is a vector of coefficients for the set of confounders contained within $\mathbf{W}$. As noted above, in the staggered DiD framework that allows for heterogeneity we do not have a single treatment effect but a collection of group-time treatment effects that may be averaged in different ways. This extends to the estimation of the CDE, in which case we will also have several CDEs that can be averaged to make inferences about the extent to which the policy's impact is mediated by $PM_{2.5}$ or temperature. Based on the setup in Equation 2 the CDE is estimated as: $\delta + \eta_{rt} MT$. In the absence of interaction between the exposure and the mediator (i.e., $\eta_{rs} = 0$) the CDE will simply be the estimated treatment effects $\sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rs}$, i.e., the effect of the policy holding M constant. For a valid estimate of the CDE we must account for confounding of the mediator-outcome effect, represented by W in the equation above. The inclusion of baseline measures of both the outcome and the proposed mediators inherent in our DiD strategy help to reduce the potential for unmeasured confounding of the mediator-outcome effect (Keele et al. 2015). Given the large number of outcomes of interest in this study, as well as the potential for heterogeneous treatment effects, we limited the mediation analysis to health outcomes for which we observed some evidence of a total effect of the CHP.

5.3 Identification of potential confounders and model covariates

In contrast to typical analytic approaches such as regression adjustment or propensity scores that solely focus on measured covariates, our DiD approach helps to minimize the risk of some sources of *unmeasured* confounding. Treatment cohort fixed effects control for measured and unmeasured time-constant factors that may differ between treatment cohorts (e.g., genetics, altitude), and time fixed effects control for secular trends, capturing any unmeasured factors that affect outcomes in all treatment cohorts (including the untreated) similarly over the study period (e.g., background improvements in ambient air quality or household transition to more efficient heating unrelated to the CHP). The latter are particularly helpful in the context of the documented declines in ambient $PM_{2.5}$ in the Beijing region attributable to air quality improvement programs and policies other than the CHP policy (Van Donkelaar et al. 2021; Zhang et al. 2019).

For models estimating the effect of the policy on indoor temperature and health outcomes, we used DAGs (Pearl 2000) to identify potential time-varying causes of both treatment by the policy

and our study outcome(s) that could differ between treatment groups, and adjusted for those potential confounders in the regression models. For the mediation analysis, we identified potential mediator-outcome confounders using the same approach. These variables were identified from the relevant peer-reviewed literature and our team's substantive knowledge about the CHP. For models estimating the effect of the policy on air pollution outcomes, the main predictors of personal exposures and indoor air quality in rural China are inconsistent across studies (e.g., Lee et al. (2021); Ni et al. (2016)). Thus, we considered the following covariates as potential determinants of air pollution in our study setting: village population and total number of households in the village; temperature, relative humidity, dew point, wind direction, wind speed, boundary layer height; home area and home area heated; home insulation; smoking status of the participant; whether or not the household reported residential wood (i.e., biomass) burning, and if so, self-reported quantity used.

Exposure to tobacco smoke is important for both air pollution and health outcomes, and we used the participant responses to survey questions related to their current smoking status (i.e., is the participant a current smoker, former smoker, or never smoker) and, for never smokers, history of passive smoking (i.e., has the participant ever lived with a smoker in the same house for at least 6 months, with possible responses of never, yes but not currently, and yes at present). The survey responses were used to create the following four distinct tobacco smoking categories: current smoker defined as currently smoking at the time of survey; former smoker defined as previously smoking but no longer smoking at the time of the survey (not accounting for duration of cessation); never smoker with a history of living with a smoker defined as a never smoker who is currently living with a smoker or has previously lived with a smoker for at least 6 months (i.e., 'passive' smoking exposure); and never smoker defined as no history of smoking and no history of living with a smoker for more than 6 months.

Ultimately, we included the following measured time-varying covariates in the final DiD models for each outcome-specific model. Models for air pollution outcomes were adjusted for household size, tobacco smoking, outdoor temperature, and outdoor humidity. As a sensitivity analysis, we additionally adjusted for district of residence given the baseline district-level differences in energy use, socioeconomic status and altitude, especially for villages in Fangshan compared with the other three districts. Temperature models were adjusted for the number of rooms, wintertime occupants in the household, age of the primary respondent, and wealth index. Models for blood pressure were adjusted for age, sex, waist circumference, tobacco smoking, alcohol consumption, and use of blood pressure medication. For self-reported respiratory outcomes we adjusted for age, gender, tobacco smoking, occupation, frequency of drinking, frequency of farming. Measured respiratory outcome (FeNO) models included adjustment for age, gender, body mass index, frequency of drinking, tobacco smoking, and frequency of exercise, occupation, time of measurement. Inflammatory marker outcome models were adjusted for age, waist circumference, occupation, wealth index quantile, frequency of drinking, tobacco smoking, and frequency of farming. For the final covariate-adjusted DiD model for personal exposure 'mixed combustion' source contributions, we adjusted for: temperature (represented by a spline with 2 degrees of freedom), tobacco smoking, and whether or not

the household reported using biomass fuel. For the final covariate-adjusted DiD model for outdoor (community) 'mixed combustion' source contributions, the following covariates were included: total number of households in the village, village population, and ambient relative humidity (represented by a spline with 2 degrees of freedom).

5.4 Multiple imputation for covariates and indoor $PM_{2.5}$ in analyses with BP outcomes

Blood pressure was measured at household visits but several key covariates like waist circumference, height, and weight were measured at the clinic visits in W1 and W2. Thus, we were missing covariate information for individuals who were unable to attend the clinic visits (~15-20% of participants in each wave). Additionally, since we only measured indoor $PM_{2.5}$ in a subsample of 300 homes in W2 and W4, we were missing indoor $PM_{2.5}$ for all participants in w1 with BP measures, as well as for a sub-sample of participants in W2 and W4. To prepare data for the BP outcomes analysis we used multiple imputation with chained equations (MICE) to impute missing indoor $PM_{2.5}$ and missing covariate data values for individuals who participated in the household visit but not the clinic visit. This allowed us to retain observations with BP measurements that would have otherwise been dropped in adjusted and mediation models using complete-case analysis. Imputation was performed with the *MICE* package (van Buuren and Groothuis-Oudshoorn 2011) in *R* (m = 30 imputation datasets, with 30 iterations each), and the difference-in-differences and mediation analyses were conducted for each of the 30 datasets. We then used Rubin's Rules to combine point estimates and standard errors while accounting for both within- and between-dataset variances (Rubin 1987).

Appendix Table A1 shows that most measures had no or <1% missing data, with the exception of measured waist circumference, height and weight (all around 15% missing). Appendix Table A2 also shows the number and percent of missing observations by treatment enrollment cohort and outcome, and we found little evidence that the percent of missing observations differed substantially between treatment cohorts. In Appendix Figure A2 we show kernel density plots for the distribution of imputed values for BMI, waist circumference, and indoor $PM_{2.5}$, all of which closely approximated the observed values.

6 Results

We retained all 50 study villages during this four-year longitudinal assessment of village treatment by the CHP, though we were only able to visit 41 villages in winter 2020-21 (W3) during which we were limited to village and household-level measurements of air quality, indoor temperature, and stove use due to travel restrictions during the COVID-19 pandemic.

By W2, W3, and W4 there were a cumulative total of 10, 17, and 20 (out of 50 total) study villages treated by the CHP policy, respectively. All of the treated villages in our study selected to install electric-powered air-source heat pumps with 200 RMB per meter square (up to 24,000 RMB) in subsidies and were also provided with 80% night-time electricity subsidies up to 10,000 kWh per heating season. To limit coal use, villages enrolled in the policy were no longer allowed to place orders for subsidized coal with the district-level governments that manage the procurement and distribution of coal for residential heating in Beijing. In addition, village committee leaders in treated villages reported feeling accountable to the Environmental Protection Department for limited coal-related air pollution, and were motivated to encourage residents to not burn coal. Some villages were equipped with government air pollution monitors and the Environmental Protection Department conducted village inspections and issued warnings about coal burning. Households burning coal in treated villages were at risk of losing their electricity subsidy.

Appendix Figure A3 and Table A3 show the participation of villages, households, and participants across the four waves of data collection and the number of sampled participants, households, and villages by study wave. Appendix Table A4 also shows selected demographic characteristics by district. We conducted measurements in over 1000 participants in each of the three measurement waves that included individual-level measurements. In total, we enrolled 1438 participants into the study, of which 630 (43%) individuals contributed 1890 observations across all the three waves where health measurements were conducted, 443 (31%) individuals contributed 886 observations across two waves and 365 (25%) individuals participated in a single wave. Table 3 shows selected demographic characteristics and health behaviors between participants who contributed to each study wave. Table 3 shows selected demographic characteristics and health behaviors between participants who contributed to each study wave. We found no differences in gender or smoking across waves, but overall BMI and waist circumference increased over time. Table 4 shows similar measures according to whether or between participants in each of the three waves with individual measurements, and we found some evidence that individuals contributing more than 1 wave of data had slightly higher BMI and lower waist circumference.

6.1 Description of study sample by treatment

Table 5 shows the distribution of selected demographic, health, and environmental characteristics from the baseline survey, prior to any villages being enrolled in the CHP. We provide means and standard deviations separately for villages that eventually enter into the policy with those that never do so. As noted above, although our DiD identification strategy allows for fixed differences between treated and untreated villages, overall the differences at baseline are generally small and the groups seem well balanced on most measures, with the exception of personal exposure to $PM_{2.5}$, which was lower in villages that were eventually treated.

Table 3: Selected demographic and health characteristics of participants in each study wave.

	Estimates			Test for Equality	
Characteristic	Wave 1 (2018-19) N=1003	Wave 2 (2019-20) N=1110	Wave 4 (2021-22) N=1028	Statistic[a]	p-value
Female, n (%)	597 (59.5)	654 (58.9)	617 (60.0)	0.270	0.874
Current smoker, n (%)	257 (25.6)	295 (26.6)	265 (25.8)	0.292	0.864
Passive smoke exposure, n (%)	486 (48.4)	538 (48.5)	486 (47.3)	0.234	0.890
Any smoke exposure, n (%)	795 (79.3)	898 (80.9)	857 (83.4)	5.616	0.060
Age in years, Mean (SD)	60.1 (9.3)	61.1 (9.1)	63.3 (9.0)	31.980	0.000
BMI (kg/m2), Mean (SD)	26.1 (3.7)	25.7 (3.5)	26.2 (3.8)	4.209	0.030
Waist circumference (cm), Mean (SD)	86.8 (10.2)	87.4 (9.4)	91.3 (10.4)	54.171	0.000

[a] Chi-square test for categorical and F-test for continuous characteristics.

Table 4: Selected demographic and health characteristics of participants who contributed to different numbers of study waves.

	Estimates			Test for Equality	
Characteristic	1 Wave N=365	2 Waves N=886	3 Waves N=1890	Statistic[a]	p-value
Female, n (%)	211 (57.8)	532 (60.0)	1125 (59.5)	0.542	0.763
Current smoker, n (%)	110 (30.1)	230 (26.0)	477 (25.2)	3.817	0.148
Passive smoke exposure, n (%)	172 (47.1)	425 (48.0)	913 (48.3)	0.099	0.952
Any smoke exposure, n (%)	293 (80.2)	732 (82.6)	1526 (80.7)	1.607	0.448
Age in years, Mean (SD)	26.3 (3.6)	25.8 (3.6)	26.0 (3.7)	1.532	0.433
BMI (kg/m2), Mean (SD)	59.8 (9.3)	61.0 (8.9)	62.1 (9.3)	11.718	0.000
Waist circumference (cm), Mean (SD)	90.3 (9.8)	88.1 (10.2)	88.7 (10.2)	4.438	0.024

[a] Chi-square test for categorical and F-test for continuous characteristics.

Table 5: Descriptive statistics for selected demographic, health, and environmental measures at baseline, by treatment status.

	Never enrolled (N=603)		Ever enrolled (N=400)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	Std. Error
Demographics:						
Age (years)	59.9	9.4	60.4	9.2	0.5	0.6
Female (%)	59.5	49.1	60.0	49.1	0.5	3.2
No education (%)	11.5	31.9	12.3	32.9	0.9	2.1
Primary education (%)	75.5	43.0	77.6	41.7	2.1	2.8
Secondary+ education (%)	12.6	33.2	9.8	29.7	−2.9	2.0
Wealth index (bottom 25%)	26.9	44.4	22.3	41.7	−4.6	2.8
Wealth index (25-50%)	23.6	42.5	27.0	44.5	3.4	2.9
Wealth index (50-75%)	24.7	43.1	25.5	43.6	0.8	2.9
Wealth index (top 25%)	24.8	43.2	25.2	43.5	0.4	2.9
Health measures:						
Never smoker (%)	21.8	41.3	19.1	39.4	−2.7	2.6
Former smoker (%)	11.9	32.4	15.1	35.8	3.2	2.2
Passive smoker (%)	39.6	49.0	40.2	49.1	0.6	3.2
Current smoker (%)	26.2	44.0	25.4	43.6	−0.8	2.8
Never drinker (%)	55.9	49.7	52.5	50.0	−3.4	3.2
Occasional drinker (%)	26.0	43.9	25.5	43.6	−0.5	2.8
Daily drinker (%)	17.8	38.3	21.9	41.4	4.1	2.6
Systolic (mmHg)	131.4	16.8	128.7	14.3	−2.7	1.0
Diastolic (mmHg)	82.7	11.6	82.1	11.3	−0.6	0.8
Waist circumference (cm)	87.7	10.5	85.4	9.5	−2.3	0.8
Body mass index (kg/m2)	26.3	3.7	25.8	3.6	−0.5	0.3
Frequency of coughing (%)	18.7	39.0	19.7	39.8	1.0	2.6
Frequency of phlegm (%)	27.6	44.7	23.7	42.6	−3.8	2.8
Frequency of wheezing (%)	6.2	24.2	6.6	24.8	0.3	1.6
Shortness of breath (%)	29.2	45.5	34.3	47.5	5.1	3.0
Chest trouble (%)	11.6	32.0	14.1	34.9	2.5	2.2
Any respiratory problem (%)	50.6	50.0	54.3	49.9	3.7	3.2
Environmental measures:						
Temperature (°C)	13.8	3.6	13.5	3.3	−0.3	0.2
Personal PM2.5 (ug/m3)	127.1	145.3	102.3	105.5	−24.7	11.9
Black carbon (ug/m3)	4.4	5.3	3.3	3.4	−1.1	0.4

6.2 Summary of PM and BC measurements

At baseline before the policy was rolled-out in any study villages, $PM_{2.5}$ and BC concentrations were higher, on average, for personal exposures compared with outdoor concentrations. From W2 onward, with the inclusion of indoor air pollution measurements, personal exposure air pollution concentrations were still higher than indoor or outdoor concentrations, with indoor levels being higher than outdoors (Table 6). This trend (personal > indoor > outdoor) was observed among households in treated and untreated villages. Personal, indoor, and outdoor geometric mean (95% confidence interval) concentrations of $PM_{2.5}$ were 72 (65, 80), 45 (39, 53), and 33 (29, 36) $\mu g/m^3$, respectively, and elevated relative to health-based guidelines. The current World Health Organization (WHO) guidelines state that annual average exposures to $PM_{2.5}$ should not exceed 5 $\mu g/m^3$, while 24-hour average exposures should not exceed 15 $\mu g/m^3$ for more than 3 to 4 days per year (World Health Organization 2021). Interim targets have been set to support the planning of incremental milestones toward cleaner air, particularly for cities, regions, and countries with higher air pollution levels. For $PM_{2.5}$, the four interim (IT) targets for annual and 24-h means are: IT-1: 35 and 75 $\mu g/m^3$; IT-2: 25 and 50 $\mu g/m^3$; IT-3: 15 and 37.5 $\mu g/m^3$; and IT-4: 10 and 25 $\mu g/m^3$ (World Health Organization 2021). The baseline personal exposures to $PM_{2.5}$ in our study aligned with IT-1, indicating considerable opportunity for air quality exposure reduction with intervention.

We also present the geometric and arithmetic means (and 95% confidence intervals) for $PM_{2.5}$ and BC in each measurement wave (Table 6). Wave 3 (2020/2021) was a partial wave that took place over a time period impacted by the COVID-19 pandemic and did not involve filter-based air pollution sample collection.

6.3 Policy uptake

Each year of the study, participants reported the types of fuels and stoves and the amount of fuel used for space heating in winter. Based on these data, heating energy types were classified into four categories: exclusive use of a heat pump ("Heat pump exclusively"), use of a heat pump and a biomass-fueled kang ("Heatpump with biomass kang"), use of solid fuel heater with an electric heating devices other than heat pumps ("Coal stove and/or biomass kang with electric heater'), and exclusive use of solid fuel ('Coal stove and/or biomass kang'). In villages treated by the policy, Figure 5 shows meaningful transitions from solid fuel to electric-powered heat pumps for all treatment cohorts. For example, the proportion of households in the group treated in 2019 (W2) using heat pumps increased from 3% in W1 to 93% in W2 and 96% in W4. Conversely, use of coal stoves decreased from 97% in W1 to 8% in W2 and 3% in W4. We observed similar stove use transitions for households in villages treated in 2020 (W3). In the three villages treated in 2021, we observed overall less exclusive use of the heat pump and a slightly larger proportion of households continuing to use coal.

Table 6: Arithmetic and geometric means for air pollutant concentrations (micrograms per cubic meter) by wave.

			Wave 1		Wave 2		Wave 3		Wave 4	
			Est.	CI	Est.	CI	Est.	CI	Est.	CI
Personal measurements										
Filter-derived	24h PM2.5	Mean	117	[105, 129]	97	[87, 107]			84	[72, 96]
		GM	72	[65, 80]	60	[54, 66]			47	[42, 52]
	24h BC	Mean	3.9	[3.5, 4.4]	3.6	[2.9, 4.2]			3.7	[2.9, 4.5]
		GM	2.6	[2.4, 2.8]	1.9	[1.7, 2.1]			1.7	[1.5, 1.9]
Indoor measurements										
Sensor-derived	Seasonal PM2.5	Mean			94	[84, 103]	84	[75, 94]	67	[59, 75]
		GM			71	[65, 77]	63	[57, 70]	47	[42, 52]
Filter-derived	24h PM2.5	Mean			69	[59, 78]			58	[48, 68]
		GM			45	[39, 53]			33	[27, 40]
	24h BC	Mean			2.7	[2.1, 3.2]			2.9	[2.2, 3.5]
		GM			1.6	[1.3, 2.0]			1.6	[1.3, 1.9]
Outdoor measurements										
Sensor-derived	Seasonal PM2.5	Mean	47	[45, 48]	55	[54, 56]	33	[32, 34]	33	[32, 34]
		GM	36	[35, 37]	40	[39, 41]	23	[22, 23]	22	[22, 23]
Filter-derived	Seasonal PM2.5	Mean	38	[34, 42]	38	[34, 41]	25	[23, 28]	26	[24, 28]
		GM	33	[29, 36]	30	[28, 32]	21	[19, 23]	22	[21, 24]
	Seasonal BC	Mean	1.5	[1.3, 1.6]	1.4	[1.3, 1.5]			1.2	[1.1, 1.2]
		GM	1.3	[1.1, 1.4]	1.1	[1.0, 1.2]			1.0	[0.9, 1.1]

Note: Est. = Estimate, CI = 95 percent confidence interval, GM = Geometric Mean

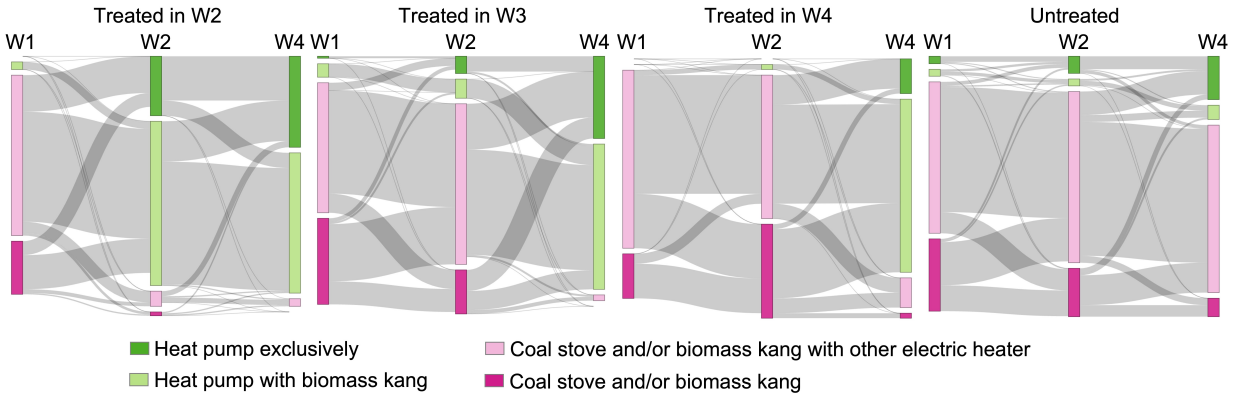


Figure 5: Transitions to different energy sources across study waves

We also observed a substantial decline in the amount of self-reported coal used in villages treated by the CHP (Figure 6), though the reduction in coal use was smaller with each subsequent treatment cohort (Appendix Table A5). Biomass (i.e., wood logs/twigs or charcoal), usually burned in kangs for both cooking and space heating, was not expressly targeted by the CHP. We observed declines in self-reported biomass use in villages treated in 2019 and 2020, but there was a small increase in biomass consumption in the cohort treated last (2021).

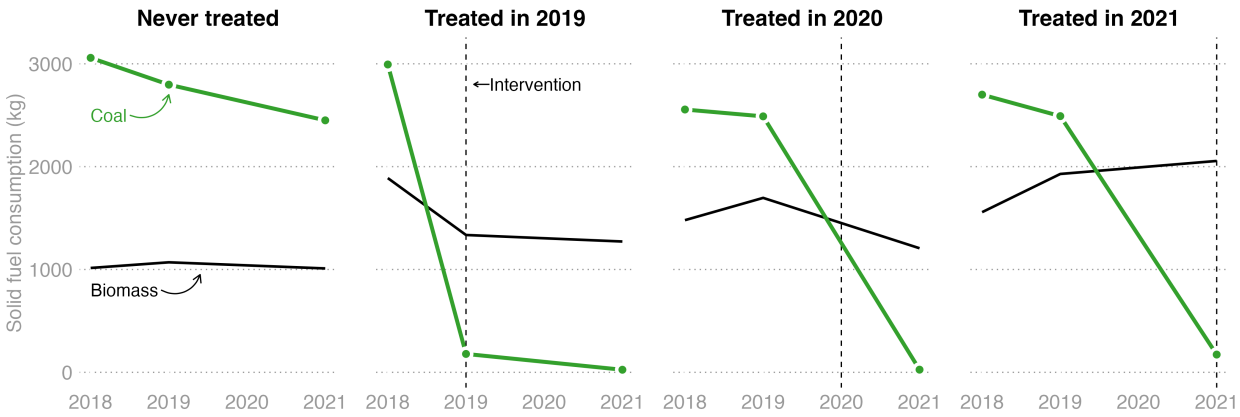


Figure 6: Trends in self-reported coal and biomass, by treatment season.

In never treated villages, we also observed a transition from solid fuel to clean energy over the four year study but it was much slower than in villages exposed to the CHP. The proportion of households that reported using electric heat pumps increased from 5% in W1 to 10% in W2 and 25% in W4, and those who adopted heat pumps tended to use them exclusively. Commensurately,

the reported expenditures on electricity increased gradually over time in the untreated villages. The percentage of untreated households using solid fuel with other types of electric devices remained relatively stable, ranging from 64% to 70% across waves. Self-reported use of biomass also remained stable, at approximately one ton of fuel each winter, whereas exclusive use of solid fuel decreased from 30% in W1 to 7% in W4.

6.4 Aim 1: Policy impacts and potential mediation

6.4.1 Impact of policy on potential mediators of air pollution and indoor temperature

The average marginal effect (ATT) from the basic ETWFE model (Table 7) shows that exposure to the CHP reduced 24-h indoor $PM_{2.5}$ by -18.9 µg/m^3 (95%CI: -56.1, 18.4). After adjusting for outdoor temperature, dew point, household smoking status, and the number of residents in each household, the ATT was -20.0 µg/m^3 (95%CI: -45.6, 5.5). The basic DiD impact was stronger on seasonal indoor $PM_{2.5}$, with an average ATT of -30.9 µg/m^3 (95%CI: -53.2, -8.7). After adjustment the average ATT on seasonal indoor $PM_{2.5}$ was -20.3 µg/m^3 (95%CI: -37.5, -3.0). This finding likely reflects the direct benefit of the policy in replacing coal stoves and air quality improvement. We found little evidence of heterogeneity in ATTs across cohort and time (all p-values > 0.4 for tests of heterogeneity, see Appendix Table A7). Overall we found little evidence of an impact of the CHP on different measures of outdoor (local, community-level) $PM_{2.5}$ or personal exposures to $PM_{2.5}$ and BC, with adjusted ATTs of -1.7 and 0.4 for 24-hr and seasonal outdoor $PM_{2.5}$, respectively. Adjusted estimates for personal $PM_{2.5}$ and personal BC were 0.5 and -0.4, respectively, and generally all of the estimates for outdoor and personal exposure impacts were imprecise (Table 7), and we provide the full set of cohort-time ATTs for personal exposure in Appendix Table A6. Appendix Table A19 and Table A20 show the impact of including Wave 3 data on the estimates of the impact of the policy on indoor (seasonal) and outdoor (24-hr and seasonal) $PM_{2.5}$, respectively. Generally this improved precision but did not affect the magnitude of our estimates. Further adjustment for district of residence in the covariate-adjusted models had little impact on our results (Appendix Table A17).

With respect to the other potential mediator of temperature, Table 7 shows that exposure to the CHP increased mean household point temperature by 1.9°C (95%CI: 0.9, 2.9), with similar impacts on mean seasonal temperatures during the heating season. The CHP had considerably stronger impacts on average seasonal minimum temperatures, which increased by 3.8°C (95%CI: 2.3, 5.4). Additionally adjusting models for the number of rooms and wintertime occupants in the household, age of the primary respondent, and wealth index had little impact on ATTs.

Table 7: Treatment effect on outdoor and indoor $PM_{2.5}$, personal exposure to $PM_{2.5}$ and black carbon, and measures of indoor temperature. Outdoor and indoor $PM_{2.5}$ were derived from sensor measurements after being adjusted based on co-located gravimetric $PM_{2.5}$ measurements. 24h indicates the mean $PM_{2.5}$ concentrations during the 24 hours when personal exposure samples were collected in each village. 'Seasonal' indicates the seasonal mean $PM_{2.5}$ concentrations in each village, from Jan. 15th to Mar. 15th.

			DiD		Adjusted DiD	
		Obs	ATT	(95% CI)	ATT[a]	(95% CI)
Air pollution						
Personal	PM2.5	1270	-3.0	(-26.1, 20.1)	0.2	(-19.6, 19.9)
	Black carbon	1161	-0.6	(-1.7, 0.6)	-0.4	(-1.5, 0.6)
Indoor	24-hr PM2.5[b]	399	-18.9	(-56.1, 18.4)	-20.0	(-45.6, 5.5)
	Seasonal PM2.5	366	-30.9	(-53.2, -8.7)	-20.3	(-37.5, -3.0)
Outdoor	24-hr PM2.5	11174	-0.5	(-5.5, 4.4)	-2.1	(-10.0, 5.8)
	Seasonal PM2.5	139	1.7	(-3.4, 6.7)	0.5	(-4.8, 5.9)
Indoor temperature						
Point	Mean	2999	1.9	(0.9, 2.9)	1.9	(0.9, 2.9)
Seasonal	Mean (all)	1350	0.7	(-0.1, 1.4)	0.7	(-0.1, 1.4)
	Mean (daytime)	1346	0.8	(0.0, 1.5)	0.8	(0.0, 1.5)
	Mean (heating season)	1350	1.8	(0.9, 2.7)	1.8	(0.9, 2.7)
	Mean (daytime heating season)	1346	2.0	(1.0, 2.9)	1.9	(1.0, 2.9)
	Min. (all)	1350	4.2	(2.3, 6.0)	4.2	(2.4, 6.1)
	Min. (heating season)	1350	4.2	(2.3, 6.0)	4.2	(2.4, 6.0)

Note: ATT = Average Treatment Effect on the Treated, DiD = Difference-in-Differences, ETWFE = Extended Two-Way Fixed Effects.

[a] ETWFE models for air pollution outcomes were adjusted for household size, smoking, outdoor temperature, and outdoor dewpoint. Temperature models adjusted for the number of rooms and wintertime occupants in the household, age of the primary respondent, and wealth index

[b] The indoor 24-hr PM2.5 concentration was determined over the time period concurrent with when the personal PM2.5 concentration was determined.

Table 8: Overall impacts of the CHP on blood pressure, respiratory outcomes, inflammatory markers and MDA.

			DiD		Adjusted DiD	
		Obs	ATT	(95% CI)	ATT[a]	(95% CI)
Blood pressure						
Systolic BP (mmHg)	Brachial	3082	-0.8	(-2.6, 1.0)	-1.4	(-3.3, 0.5)
	Central	3081	-1.0	(-2.8, 0.7)	-1.6	(-3.4, 0.3)
Diastolic BP (mmHg)	Brachial	3082	-1.3	(-2.6, 0.0)	-1.6	(-3.0, -0.2)
	Central	3081	-1.4	(-2.7, -0.0)	-1.7	(-3.0, -0.3)
Pulse Pressure	Brachial	3082	0.5	(-0.7, 1.7)	0.2	(-1.0, 1.4)
	Central	3081	0.3	(-0.8, 1.5)	0.1	(-1.0, 1.2)
BP Amplification x100	Pulse pressure	3081	0.1	(-1.1, 1.4)	-0.0	(-1.2, 1.2)
	Systolic BP	3081	0.2	(-0.2, 0.5)	0.1	(-0.2, 0.4)
Respiratory outcomes						
Self-reported (pp)	Any symptom	3076	-7.7	(-12.8, -2.5)	-7.5	(-12.7, -2.3)
	Coughing	3076	-2.6	(-7.2, 2.0)	-2.7	(-7.1, 1.7)
	Phlegm	3076	-1.3	(-5.5, 2.9)	-1.6	(-5.6, 2.4)
	Wheezing attacks	3076	0.7	(-2.3, 3.8)	1.0	(-1.9, 3.9)
	Trouble breathing	3076	-4.4	(-9.9, 1.0)	-3.4	(-9.2, 2.4)
	Chest trouble	3076	-4.2	(-8.8, 0.5)	-3.4	(-8.1, 1.3)
Measured	FeNO (ppb)	793	0.9	(-1.6, 3.3)	0.3	(-2.2, 2.8)
Inflammatory markers						
Measured	IL6 (pg/mL)	1603	0.9	(-0.2, 1.9)	0.8	(-0.3, 2.0)
	TNF-alpha (pg/mL)	1603	1.0	(0.1, 1.8)	0.8	(-0.1, 1.7)
	CRP (mg/L)	1603	0.1	(-0.4, 0.6)	0.1	(-0.5, 0.6)
	MDA (μM)	1603	0.3	(-0.1, 0.8)	0.2	(-0.2, 0.6)

Note: ATT = Average Treatment Effect on the Treated, DiD = Difference-in-Differences, ETWFE = Extended Two-Way Fixed Effects, Obs = observations, pp = percentage points, ppb = parts per billion.

[a] ETWFE models for blood pressure models adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication. Self-reported respiratory outcomes adjusted for age, gender, smoking, occupation, frequency of drinking, frequency of farming. Measured respiratory outcome (FeNO) adjusted age, gender, body mass index, frequency of drinking, smoking, and frequency of exercise, occupation, time of measurement. Inflammatory marker and MDA outcome models adjusted for age, waist circumference, occupation, wealth index quantile, frequency of drinking, tobacco smoking, and frequency of farming.

6.4.2 Impact of the policy on health outcomes

Table 8 shows the impacts of the policy on blood pressure in basic ETWFE models and models further adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication. Overall exposure to the CHP demonstrated reductions in blood pressure of approximately 1.5 mmHg for both systolic and diastolic BP, but we found little evidence of a meaningful impact on pulse pressure or BP amplification. The effects of the policy on brachial and central blood pressures were similar, and were consistent when restricted to only those participants enrolled in W1 (i.e., excluding participants recruited in later waves, see Appendix Table A18). However, the average effects in Table 8 conceal a fair amount of heterogeneity in treatment effects for blood pressure across treatment cohorts and time. Appendix Table A9 shows that treatment impacts were considerably stronger for the earlier compared to the later-treated cohorts. For example, the CHP reduced central DBP in the year of treatment by -2.7 mmHg (95%CI: -4.7, -0.7) for the villages first treated in wave 2, but increased central DBP by 1.1 mmHg (95%CI: -0.1, 2.2) for the villages treated in wave 4 (p-value for heterogeneity <0.0001).

Table 8 shows the impacts on self-reported chronic respiratory symptoms categorized as any symptoms and separately for each individual symptom type. Based on the covariate-adjusted ETWFE models, exposure to the CHP reduced self-report of any poor respiratory symptoms by 7.5 percentage points (95%CI: -12.7, -2.3). This overall effect was mostly due to reductions of roughly 3 percentage points in reports of coughing, having chest trouble, or difficulty breathing, respectively, on several or most days of the week. We found limited evidence that the CHP reduced self-reported symptoms of phlegm (-1.6, 95%CI: -5.6, 2.4) or wheezing (1.0, 95%CI: -1.9, 3.9). Appendix tables A11, A12, A13, A14, A15, A16 show little evidence of any systematic heterogeneity in the cohort-time treatment effects across the outcomes of any symptom, coughing, chest trouble, or trouble breathing, but the overall small ATTs for phlegm and wheezing may be due to heterogeneity as the CHP reduced symptoms in earlier treated cohorts but increased symptoms in later cohorts.

Table 8 also shows the impacts of the CHP on FeNO, which was conducted in a sub-sample of 511 participants, including 274 participants with one measurement, 142 with two measurements, 95 participants with 3 measurements. We did not find evidence that the policy affected changes in FeNO in the covariate-adjusted ETWFE model (0.3 ppb, 95%CI: -2.2, 2.8). There was some evidence of heterogeneity in the FeNO effects of the policy by treatment cohort Appendix A4, though the confidence intervals for each of the cohort-specific effects were wide and overlapping. Our results did not change with sensitivity analyses that limited the analysis to participants with at least two repeated measurements and to those who participated in all three waves (Appendix Table A21)

We also found limited evidence of an impact of the CHP on markers of inflammation or oxidative stress. The basic DiD analyses showed an increase of 1.0 (95% CI: 0.1, 1.8) in TNF-α but this estimate was reduced to 0.8 (-0.1, 1.7) after adjustment for waist circumference, occupation, wealth index quantile, frequency of drinking, tobacco smoking, and frequency of farming. The adjusted

estimates for the effect of the policy on IL-6, CRP, and MDA were also generally small and measured with limited precision.

6.4.3 Mediated impact on health outcomes

As noted above, we aimed to assess whether any health impacts of the CHP may work specifically through pathways involving changes in $PM_{2.5}$ and indoor temperature. Below we show results from several mediation models. We focus the mediation analysis on the BP and respiratory outcomes for which we observed stronger evidence of total effects of the policy. We evaluated potential mediation for each mediator (indoor temperature and exposure to indoor $PM_{2.5}$) separately and in a single model accounting for multiple mediators, and we set the values of both mediators to the mean value for untreated participants at baseline (W1).

In Table 9 we show estimates of the CDEs for different sets of potential mediators. The first column of the first panel shows the covariate-adjusted total ATT of the CHP on brachial SBP (i.e., a 1.4 mmHg decrease as seen in Table 8 above). The second panel shows the CDE, i.e., the effect of exposure (vs. no exposure) to the CHP on brachial SBP in a counterfactual population where we intervene to fix the value of indoor $PM_{2.5}$ to the average value for untreated participants at baseline. The CDE is -0.8 mmHg (95% CI -2.9, 1.3), demonstrating that, under the assumptions outlined in Section 5.2, roughly 40% of the total effect of the CHP is mediated by the impact of the policy on indoor $PM_{2.5}$. The third panel shows the estimated CDE when holding constant the value of indoor temp (without simultaneous adjustment for $PM_{2.5}$) to that of the untreated participants. This CDE is even smaller (-0.3 mmHg, 95%CI -2.2, 1.6) suggesting a somewhat stronger role for indoor temperature in mediating the total effect of the CHP on brachial SBP. Finally, the last panel shows a similar CDE of 0.3 mmHg (95% CI: -1.9, 2.5) when setting both indoor $PM_{2.5}$ and indoor temperature to their respective pre-treatment means. Holding the values of $PM_{2.5}$ and indoor temperature at pre-intervention values effectively eliminates these pathways by which the CHP can effect BP, so the small value of the CDE adjusting for both mediators suggests that the CHP effect on BP would be effectively null were it not for the impact on the mediators. Overall the results in Table 9 indicate that conditioning on indoor $PM_{2.5}$ and indoor temperature largely explain the entire total effect of the CHP on blood pressure for systolic BP, as the CDE conditional on both mediators was reduced to 0.03 for brachial SBP. The CDEs for brachial and central DBP were roughly half the value of the total effect. Appendix Table A10 shows heterogeneous treatment effects for the mediation models for SBP and DBP and are generally consistent with the patterns of mediation for the overall CDEs.

Table 10 shows estimates from similar analyses for the CDE of the policy on respiratory outcomes. For respiratory outcomes we focus on mediation by personal exposure to $PM_{2.5}$ and point temperature and therefore these estimates are derived for the subset of individuals with measures of personal exposure. Thus the total adjusted ATTs in Table 10 are not directly comparable with those in Table 8. We estimate the CDEs holding the values of both mediators to the average levels

Table 9: Controlled direct effects for the CHP on blood pressure.

	Adjusted Total Effect		CDE Mediated By:					
			Indoor PM		Indoor Temp		PM + Temp	
	ATT[a]	(95%CI)	ATT[b]	(95%CI)	ATT[b]	(95%CI)	ATT[b]	(95%CI)
Brachial SBP	-1.4	(-3.3, 0.5)	-0.8	(-2.9, 1.3)	-0.3	(-2.2, 1.6)	0.3	(-1.9, 2.5)
Central SBP	-1.4	(-3.3, 0.4)	-0.8	(-2.9, 1.3)	-0.4	(-2.2, 1.3)	0.2	(-1.9, 2.4)
Brachial DBP	-1.6	(-2.9, -0.3)	-1.1	(-2.7, 0.5)	-1.1	(-2.3, 0.1)	-0.6	(-2.1, 0.9)
Central DBP	-1.6	(-2.9, -0.3)	-1.1	(-2.7, 0.6)	-1.2	(-2.4, -0.0)	-0.7	(-2.2, 0.9)

Note: Results combined across 30 multiply-imputed datasets (average of 3082 observations per dataset). ATT = Average Treatment Effect on the Treated, CDE = Controlled Direct Effect, CI = Confidence Interval, DBP = Diastolic blood pressure, PM = Particulate matter, SBP = Systolic blood pressure.

[a] Adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication.

[b] Mediators were set to the mean value for untreated participants at baseline.

for never treated households at baseline. Overall we find no evidence that any of the total effects we observed for self-reported respiratory outcomes in Table 8 were mediated by personal exposure to $PM_{2.5}$ or indoor temperature. Generally the CDEs for all of the outcomes are statistically indistinguishable from the total effects estimated without controlling for mediators.

6.5 Aim 2: Source contributions

Source analysis for this study was conducted using data from all eligible outdoor and personal exposure $PM_{2.5}$ samples. Eligible samples were those for which both $PM_{2.5}$ mass and chemical components were quantified. Individual chemical species concentrations (means and 95% confidence intervals) for outdoor and personal samples by study wave are provided in Appendix tbl-species-outdoor and tbl-species-personal, respectively. We evaluated factors contributing to community-outdoor and personal exposure $PM_{2.5}$ using the U.S. EPA's source apportionment model PMF (positive matrix factorization) 5.0, which has been widely used for air pollution analyses in China (Gao et al. 2018; Liu et al. 2017; Tao et al. 2017). As an optimum PMF result depends on the appropriate number of input factors, sensitivity analysis using a range of factors (e.g., 3 to 7, based on a combination of the measured chemical species, field observations, and sources previously identified in our study region) were conducted to examine the impact of a different number of factors on the model results. Detailed information on the procedures of PMF analysis can be found elsewhere (Wang et al. 2016; Zíková et al. 2016). Briefly, the scree plot from our principal

Table 10: Controlled direct effects of the CHP on self-reported respiratory outcomes

	Adjusted Total Effect		CDE Mediated By:					
			Personal PM		Indoor Temp		PM + Temp	
	ATT[a]	(95%CI)	ATT[b]	(95%CI)	ATT[b]	(95%CI)	ATT[b]	(95%CI)
Any symptom	-13.0	(-20.5, -5.5)	-12.7	(-20.7, -4.6)	-15.1	(-23.3, -6.9)	-15.0	(-23.7, -6.3)
Coughing	-10.5	(-19.2, -1.8)	-11.9	(-20.7, -3.1)	-14.5	(-23.1, -5.8)	-14.7	(-22.9, -6.6)
Phlegm	-9.5	(-16.6, -2.4)	-8.8	(-16.2, -1.4)	-14.7	(-24.3, -5.1)	-13.3	(-21.3, -5.4)
Wheezing attacks	-4.2	(-11.1, 2.6)	-2.8	(-8.7, 3.1)	-10.2	(-22.0, 1.6)	-3.8	(-10.0, 2.4)
Trouble breathing	-9.5	(-19.3, 0.3)	-9.6	(-19.6, 0.4)	-13.8	(-26.3, -1.3)	-13.1	(-25.2, -1.0)
Chest trouble	-5.0	(-12.2, 2.1)	-4.9	(-11.2, 1.4)	-4.6	(-11.8, 2.5)	-3.0	(-8.9, 2.9)

Note: Estimated for the subset of individuals with measured personal exposure (n=1270). ATT = Average Treatment Effect on the Treated, CDE = Controlled Direct Effect, CI = Confidence Interval.

[a] Adjusted for age, gender, smoking, occupation, freqency of drinking, frequency of farming.

[b] Mediators were set to the mean value for untreated participants at baseline.

component analysis indicated that solutions of between 3 and 5 factors (+/- 1) would be most appropriate, further supporting our evaluation of 3 to 6 factor solutions from PMF. As there was no indication that even moving from 5 to 6 factors would improve our solution, we did not further investigate 7 factors (Figure 7).

6.5.1 Source analysis using positive matrix factorization

The chemical analysis data used in the PMF model were dispersion normalized prior to their inclusion. PMF uses the covariance of compositional variables to separate sources of PM. However, atmospheric dilution also induces covariance. Dilution can be quantified in terms of a ventilation coefficient (VC) and used to normalize the input chemical concentrations and uncertainties in the original data matrix on a sample by sample basis. The dispersion normalized concentrations and uncertainties are used as the inputs to PMF analysis. Dispersion normalization, as conducted in this study, is a relatively new application of this conceptual framework (Dai et al. 2020), developed to adjust for wind speed (dispersion in the x-y plane) and boundary layer height (dispersion in the z-axis). This process involves first calculating the sample specific ventilation coefficient by multiplying the average wind speed by the average boundary layer height over the sampling duration. The average ventilation coefficient is also calculated for the village by averaging all the ventilation coefficients. The dispersion normalized concentration for any species in any sample is equal to the species concentration in that sample multiplied by the ventilation coefficient for that sample and

divided by the average ventilation coefficient for that village. Dividing by the average ventilation coefficient for that village helps curtail any extreme concentrations driven by an outlier in the sample ventilation coefficient.

The meteorological data included hourly boundary layer height, 2-m temperature, 2-m dew point temperature, and 2-m horizontal wind speed components (u, v), which were obtained from the European Center for Midrange Weather Forecasting ERA5 reanalysis dataset (0.25 x 0.25 resolution). Values of these meteorological variables were determined at the village-level by identifying the four surrounding grid points with values available from the ERA5 reanalysis, and then applying inverse distance weighted interpolation from those four grid points to the village. Percent relative humidity was calculated from the 2-m dew point temperature using the "weathermetrics" package (version 1.2.2) in R (Anderson et al. 2016). Total hourly wind speed and wind direction were calculated from the horizontal wind speed components.

The model diagnostics for the 3- to 6-factor PMF solutions are shown in Table 11. Model fit was assessed using a ratio of our model fit (Q) divided by the expected fit (Qexp). As the change in Q/Qexp decreased with more factors, the model may be fitting additional sources that do not improve the overall fit. The largest change in Q/Qexp was from 3 to 4 sources (6.24 to 5.37) while the changes moving from 4 to 5 factors and 5 to 6 factors were similar, which suggests that 4 factors are sufficient and parsimonious to explain the variation in our data. We assessed the random error in our model by randomly sampling blocks of data, fitting new models with the blocks, and comparing how the source profiles compared with the original model (bootstrap mapping). The 3- and 4-factor solutions had high bootstrap mapping (all factors identified in >96.5% of bootstrap runs). The additional sources identified in the 5-factor (lead) and six-factor (chloride) solutions had low bootstrap mapping (> 72%), indicating that these solutions are less consistent than the 3- and 4-factor solutions. The possibility that multiple solutions could result in the same Q value was assessed using displacement. The displacement approach takes the original factor profiles and modifies (+/-) the values for each species to maintain a small change in Q, reruns the solution with the new species values, and then compares the profiles of the new model to the original. Any swaps indicate that small changes to the species values could result in factor profiles that are different from the original solution, suggesting that the original solution is unstable. None of the factors in any of the solutions discussed were swapped during displacement, which indicates that all of the potential solutions are stable. Based on the Q/Qexp, bootstrap mapping, and interpretability of the factors, the 4-factor solution was selected as most appropriate for the data.

The source profiles for the four-factor solution are presented in Figure 7. We sought to develop a defensible source analysis solution, which we found to have 4 factors and then identify and name those sources, jointly informed by our field observations, knowledge of local sources, and relevant previous studies. The pooled PMF analysis, which combined outdoor and personal exposure samples, led to a more robust factor solution due to the increased number of samples. This pooled analysis determined that the optimum solution was a 4-factor model, which identified the major sources as dust, transported dust, secondary sulfur, and a combustion mixture of coal and biomass

Table 11: PMF error estimation diagnostics.

	Potential Factor Solution			
Diagnostic	3	4	5	6
Qexp	27936	26052	24168	22284
Qtrue	187681	147796	123236	100316
Qrobust	174407	139910	117082	95932.5
Qr/Qexp	6.24	5.37	4.84	4.3
Q/Qexp > 6	wi-Ca, ns-S, ws-Na, ws-Ca, Al, Cl, Pb	ns-S, Na, Al, Cl, Pb, Nitrate	Nitrate, ws-Na, Al, Chloride	Nitrate, ws-Na, Al
DISP % dQ	$<$0.1%	$<$0.1%	$<$0.1%	$<$0.1%
DISP swaps	0	0	0	0
BS_mapping	Dust- 98.5%	Transported dust- 95%, Dust- 96.5%, Sulfur secondary- 97.5%, Mixed combustion- 96.5%	Transported dust- 86%, Mixed combustion- 87%, Dust- 86%, Lead- 55%	Transported dust- 84%, Mixed combustion- 87.5%, Dust- 81.5%, Lead- 72% Chloride- 61.5% Sulfur secondary- 98.5%

burning, and possibly some tobacco smoking. The pooled approach was found to be stable, with high bootstrap mapping and no factor swaps during the displacement tests, indicating the reliability of the factor solutions.

As sensitivity analyses , we additionally conducted a disaggregated analysis (Appendix Figure A6) where personal and outdoor samples were analyzed separately. We observed some differences in the number of factors and source attributions, however, the core findings remained largely consistent with the pooled results. For the personal exposure samples, the 3-factor solution identified dust, transported dust, and mixed combustion (a combination of biomass burning and secondary PM), while the 5-factor solution further split the mixed combustion factor into distinct coal combustion and sulfur secondary factors. Similarly, for outdoor samples, the 3-factor solution identified mixed combustion, secondary PM, and dust, while the 5-factor solution introduced transported dust and a refined characterization of secondary species contributions. The primary sources of pollution—dust, secondary sulfur, and mixed combustion—were consistent across pooled and disaggregated analyses. Thus, while the disaggregated analysis does provide additional granularity, it does not fundamentally change the identification of the key pollution sources that we observed in the pooled analysis.

We also evaluated PMF results disaggregated by day and by month (Appendix Figure A7 and Figure A8), where the results are further color-coded by district. Due to yearly field campaign schedules, the timing of sampling in villages and districts was correlated. Therefore, this approach to source analysis does not yield results that allow us to disentangle changes in sources over time, within season, since they also potentially embed changes in sources across villages and districts in this study.

Thus, we concluded that the pooled analysis was the most parsimonious and interpretable approach for explaining the major sources of $PM_{2.5}$ in our study. The pooled results are both representative and stable, making them the most appropriate for addressing the study's primary research questions.

6.5.2 Description of $PM_{2.5}$ sources identified

The first source was identified as dust, characterized by high percentages of crustal elements like wi-Ca, Si, and wi-Mg. The second source contained non-sulfate sulfur and secondary inorganic ions (ammonium, nitrate, and sulfate). Non-sulfate sulfur is a tracer for primary coal combustion, while secondary inorganic ions indicate a secondary source. Given the industrial coal burning in our study area, the secondary source likely combines primary and secondary emissions from coal and other sulfurous fuel combustion. Additionally, the higher outdoor concentrations of the secondary source compared with personal exposures support its identification as 'sulfur secondary' due to its sunlight-driven secondary formation. The factor named "sulfur secondary" was intended to reflect the contribution of sulfur, as measured by the XRF analysis, that was not associated with sulfate.

Had this contribution been coupled with other species associated with direct air pollutant emissions from coal, and had those species not clustered with any other factors, we may have named this source a "household coal combustion" source. However, some species that are typical of direct air emissions from coal combustion are also clustered with species that are typical of direct air emissions from biomass burning, leading to naming the factor as "mixed combustion". It is not surprising that residential coal and biomass emissions were difficult to fully separate, being unable to analyze samples for organic tracers.

The third source had high percentages of ws-Ca nd Al, which in our study region, has been found to be indicative of transported dust from dust storms that can occur in the spring. While our samples were collected during winter months only, it is possible that transported dust from previous years still remained. The fourth source was characterized by high percentages of tracers for both coal (OC, wi-K, chloride, Pb) and biomass combustion (EC, ws-K). Coal and biomass combustion are anticipated sources of $PM_{2.5}$ pollution in our study setting, particularly from domestic cooking and heating activities, so this source is likely a mixture of PM emitted from these two household combustion sources. We extend the source profiles across the different treatment cohorts in Figure 8.

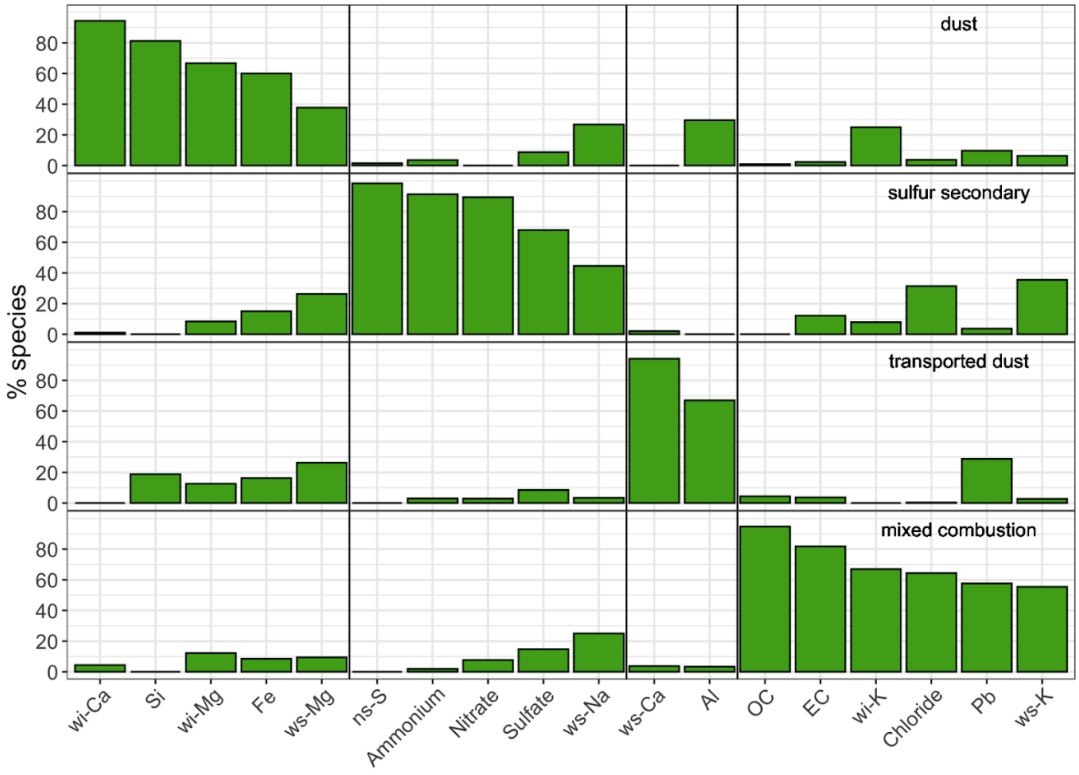


Figure 7: Source profiles for the 4-factor PMF solution to the sum of elements, ions, elemental carbon, and organic carbon for outdoor and personal $PM_{2.5}$ exposure measurements. The lines separate the major contributing species to each source.

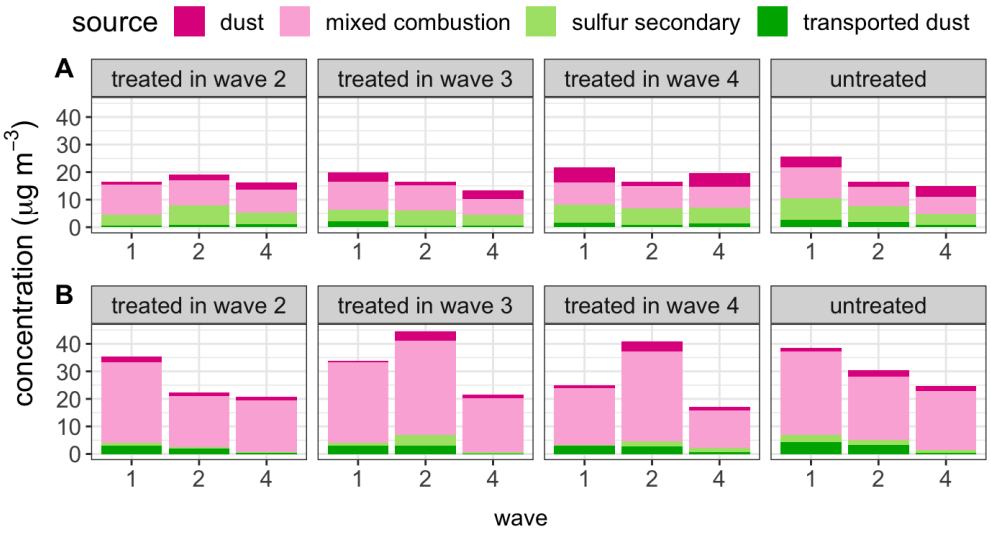


Figure 8: Arithmetic mean dispersion normalized source contributions found from the 4-factor PMF solution for **A** outdoor and **B** personal $PM_{2.5}$ exposure samples by year the group received treatment.

6.5.3 Impact of policy on outdoor and personal exposure to the mixed combustion source

Overall, Table 12 shows that the average treatment effect of the CHP on outdoor (community) levels and personal exposure levels of the mixed combustion source was statistically indistinguishable from the null. Treatment was associated with lower, but statistically imprecise, personal exposures to the mixed combustion source. As with personal exposure to BC, an indicator of combustion pollution, this finding is consistent with the expectation that the policy contributed to reduced solid fuel emissions, as the 'mixed combustion' source most likely reflects solid fuel combustion in our study setting. The results were consistent across the basic and covariate-adjusted models.

When the average treatment effects of the CHP policy on community outdoor levels and personal exposure levels of the mixed combustion source were allowed to vary by treatment year and time, the treatment effect for households most recently treated (i.e., treated in the final wave, W4) was associated with lower personal exposures to the mixed combustion source (Appendix Figure A5). In each wave, treatment by the CHP was associated with a reduction in the source contribution to personal $PM_{2.5}$ mass from the mixed combustion source; however, for villages treated in W2 and W3, the effect was statistically imprecise. Treatment was not associated with a reduction or an increase in the source contribution to community outdoor $PM_{2.5}$ mass from the mixed combustion source. Personal exposure measures of this specific air pollution source were found to be more indicative of treatment effect than community outdoor measures of the same source. This finding aligns with the expectation that the mixed pollution source, identified as a mixture of coal and

Table 12: Average treatment effect (μg/m^3) for outdoor and personal exposure to the mixed combustion source.

		DiD		Adjusted DiD	
	Obs	ATT	(95% CI)[a]	ATT	(95% CI)
Outdoor	717	1.07	(-4.90, 7.04)	1.53	(-4.19, 7.26)
Personal exposure	1158	-5.60	(-13.70, 2.54)	-5.39	(-13.1, 2.35)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval, DiD = Difference-in-Differences.

[a] Personal exposure model adjusted for temperature (represented by a spline with 2 degrees of freedom), participant smoking status, and household reported using biomass fuel. Outdoor model adjusted for the total number of households in the village, total village population, and ambient relative humidity (represented by a spline with 2 degrees of freedom).

biomass combustion, is characteristic of household use of solid fuels. These fuels, including coal and biomass, produce emissions that are likely to be closer to the people using them rather than near the centrally located community outdoor air samplers.

6.6 Aim 3: Mediation by source contribution

Table 13 shows results from the mediation analysis by personal exposure to the mixed combustion source (coal and biomass), estimated for the subset of participants with personal exposure measurements. The CDE in this model estimates the impact of exposure to the CHP on central and systolic blood pressure while holding constant values of mixed combustion source at the mean baseline values for untreated population. The marginal policy effects (ATTs) from the adjusted ETWFE models for this subset of participants were largely similar to those from the full sample for central SBP (around a 1.6 mmHg decrease), but slightly smaller for central DBP (-1.6 mmHg in the full sample vs. -0.9 mmHg in the subset with personal exposure measurements) and were estimated with greater imprecision. We found little evidence that these treatment effects were meaningfully mediated by exposure to the mixed combustion source, as the controlled direct effects were generally of similar magnitude as the adjusted total effects.

7 Discussion and Conclusions

Air pollution emitted from residential space heating with coal has historically been a major contributor to cardio-respiratory disease burden in northern China (Archer-Nicholls et al. 2016; Yun et

Table 13: Average treatment effects and controlled direct effect (mm/Hg) of the CHP on central systolic and diastolic blood pressure with mixed combustion source as the potential mediator.

		DiD		Adjusted DiD		Adjusted CDE	
	Obs	ATT	(95% CI)[a]	ATT	(95% CI)[b]	ATT	(95% CI)
Central SBP	942	-1.6	(-5.3, 2.2)	-1.5	(-5.0, 2.0)	-1.5	(-5.0, 2.1)
Central DBP	942	-0.9	(-3.3, 1.5)	-1.1	(-3.2, 0.9)	-1.3	(-3.6, 0.9)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval, DiD = Difference-in-Differences, CDE = Controlled Direct Effect, DBP = Diastolic Blood Pressure, SBP = Systolic Blood Pressure.

[a] Adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication.

[b] Further adjusted for mediation by mixed combustion source (coal and biomass)

al. 2020). Since the introduction of its 13th 5-Year-Plan (2016-2020), China has successfully implemented numerous large-scale measures to improve air quality including programs that incentivize rural household transition from solid fuels to clean energy sources (Young et al. 2015). The CHP is among the largest and most ambitious household energy policies implemented anywhere in the world in recent decades, and its staggered roll-out provided a unique opportunity to prospectively evaluate this real-world experiment and its effects on air quality and health.

7.1 Adoption of the heat pump technology and adherence to the policy

The CHP was successful in driving a rapid household heating energy transition from coal stoves to electric heat pumps in the treated study villages, with little difference in coal stove suspension or heat pump adoption for those treated before versus during the COVID-19 pandemic. There was high uptake and consistent use of the new heat pump technology, as well as large reductions in coal use in treated villages starting in the first year post-treatment and continuing into the third year of treatment for the villages first treated in 2019. We enrolled rural and peri-urban villages across a wide geographic area and socioeconomic spectrum in Beijing and observed near universal adoption of the heat pump technologies and suspension of coal stove use across the different treatment groups and waves. This contrasts with many previous household energy intervention studies, including several randomized trials, where low fidelity and compliance with the intervention stoves were considered major limitations to achieving their intended air quality or health benefits (Ezzati and Baumgartner 2017; Lai et al. 2024; Rosenthal et al. 2018).

A number of factors contributed to the successful uptake of the new technology and adherence to the policy. The initial uptake of the heat pump technology was influenced by broad support and perceived benefits of village and household participation in the policy. At baseline assessment, 49

of 50 village committee interviewees indicated a desire to participate in the policy by the committee members and their constituents, for reasons including the ease of use of the heat pump, the convenience of no longer having to add coal throughout the day and especially the night, the desire for a cleaner local environment, and a perceived lower risk of carbon monoxide poisoning without coal stoves (data not reported). While the availability and cost of clean fuels are well-established barriers to their adoption and sustained use over time (Rehfuess et al. 2014), in our study, both the upfront costs of the heat pump technology and a portion of electricity use were subsidized by the government, which limited the financial burden of clean energy transition for households. Further, after policy implementation, treated villages no longer had access to government-subsidized coal, and household coal burning was further discouraged with possible punitive measures (e.g., potential loss of electricity subsidies).

7.2 Impacts of the policy on health

One of the key findings from our comprehensive evaluation of the CHP was that exposure to the policy reduced systolic and diastolic blood pressure by ~1.5 mmHg, and that most of the observed BP effects were mediated by improvements in the indoor environment, specifically reductions in indoor $PM_{2.5}$ and increases in indoor temperature. The total effects of the policy are consistent with a small number of randomized trials of gas cookstoves or more efficient biomass cookstoves showing reductions in blood pressure of similar magnitudes (Kumar et al. 2021). In contrast, recent randomized trials of liquefied petroleum gas (LPG) stoves in multiple countries observed no effect or a small (~0.6 mmHg) increase in blood pressure (Checkley et al. 2021; Ye et al. 2022) despite large decreases in personal exposures to $PM_{2.5}$ and black carbon. The inconsistency between our results and the LPG stove trial may stem from large differences in age (mean ages of 25y and 48y in the trials versus 61y in our sample) and that gas stoves can still emit health-damaging air pollutants including benzene and volatile organic compounds (Kashtan et al. 2023), especially in contrast with the zero-emission electric-powered heat pumps introduced to our study villages. Our findings of temperature- and air quality-mediated impacts of the policy on BP are also supported by observational studies showing that increased exposure to household air pollution (Baumgartner et al. 2018, 2011; Dong et al. 2013; Kanagasabai et al. 2022) and to colder indoor temperatures (Lv et al. 2022; Sternbach et al. 2022) are associated with higher blood pressure in rural and peri-urban areas of China, with exposure-response estimates that reasonably align with our estimates of the policy impact on BP after conditioning on temperature and $PM_{2.5}$ in the mediation analysis.

We did not observe effects of the policy on measures of PP or cPP/SBP amplification. Pulse pressure is measured as the difference between SBP and DBP, and represents the pulsatile component of blood flow (Dart and Kingwell 2001). Thus, increases in PP can result from increases in SBP, decreases in DBP, or both. The lack of effect on PP in our study is attributed to the similar reductions in SBP and DBP from the policy. Similarly, PP/SBP amplification is measured as a ratio of peripheral to central pressures, and the decreases in central and brachial pressures with the

policy were also nearly identical in our study. Although the duration of our study was nearly twice as long as most previous household stove intervention studies conducted over two years or less, it is still possible that even longer-term reductions in BP are required to observe any structural changes in the caliber or elasticity of arterial walls that would subsequently be reflected in differences in PP or SBP/PP amplification (Dart and Kingwell 2001).

Our study also contributes to the limited evidence that transition from solid fuel to clean energy can reduce the self-report of symptoms consistent with chronic respiratory tract irritation. Exposure to the CHP reduced self-report of any chronic respiratory symptoms by roughly 7 percentage points, with most of these effects driven by reductions in self-reported chest trouble or difficulty breathing on several or most days of the week. These findings align with previous randomized trials in Guatemala and Mexico, where biomass chimney stove interventions lowered indoor carbon monoxide and reduced the self-reported prevalence of chronic respiratory symptoms, especially coughing and wheezing, in younger women after 12 and 18 months of intervention (Romieu et al. 2009; Smith-Sivertsen et al. 2009). In contrast to our findings, introduction of a solar cooker in Senegal provided no benefit to air pollution or self-reported respiratory symptoms (Beltramo and Levine 2013), and a recent trial of gas stoves in Peru similarly found no reduction in respiratory symptoms within the year of intervention despite very large reductions in $PM_{2.5}$ (Checkley et al. 2021).

We did not, however, find evidence that the reductions in chronic respiratory symptoms were mediated by changes in personal exposure to $PM_{2.5}$ or indoor temperature. This is not particularly unexpected since we did not observe an effect of the policy on personal exposure to $PM_{2.5}$ and any impacts of temperature on chronic respiratory symptoms would more likely arise from large, rapid changes in temperature (D'Amato et al. 2018) whereas we observed small, gradual changes in our study. Future work will consider mediation by seasonal indoor $PM_{2.5}$, which is a longer-term measure of 'usual' air pollution than 24-h personal exposure and was shown to be reduced by the policy in our study homes.

We found some evidence of heterogeneity in the health benefits of the policy by treatment cohort. Generally the policy showed strong reductions for the villages treated early and weak or null increases in BP for the last 3 villages exposed to the policy in 2021. We found less evidence for heterogeneity for self-reported respiratory symptoms, but did note potential evidence for increases in self-reported phlegm and wheezing attacks with the policy in the three villages treated in 2021. Notably this was also the treatment cohort with the smallest improvement in point temperature at the time of BP measurement and an increase in self-reported biomass use. Paradoxically, we observed a larger decrease in $PM_{2.5}$ and mixed solid fuel use in this group. It is possible that the composition of PM and mixed solid fuel was different in this cohort, with a greater contribution of biomass smoke, however we are unable to differentiate between biomass and coal in our 'mixed solid fuel' category. Notably, this group of villages were also treated during the COVID-19 pandemic, which could have impacted how the policy was introduced in unpredictable and difficult-to-measure

ways, which could have resulted in changes to other health risk factors that we did not evaluate in our study, e.g., changes in dietary intake.

We also found little evidence of an impact of the policy on blood biomarkers of inflammation and oxidative stress in the sub-sample of participants with blood collection in waves 1 and 2, but these were estimated with imprecision. Our results contrast with a natural experiment in urban Beijing that showed large regional and local air quality reductions during the 2008 Beijing Olympics and also observed benefits to airway inflammation (Huang et al. 2012) and blood markers of inflammation and oxidative stress in healthy urban Beijing residents during the Olympics compared with before and after (Rich et al. 2012). Our mediation analysis indicated that the blood pressure effects of the policy were mediated through both indoor temperature and air pollution, and the effects of the policy on SBP were mediated more through indoor temperature than air pollution. Although observational studies from rural northern China show impacts of exposure to temperature on inflammation and oxidative stress (Wang et al. 2020; Xu et al. 2019), it's possible that the relatively small increases in mean indoor temperature in treated households we observed were not sufficiently large to capture measurable changes in these biomarkers.

7.3 Impacts of the policy on air pollution and its sources

The primary aim of the CHP was to reduce air pollution emissions and improve regional air quality, and it specifically targeted coal-burning stoves in northern China. Our evaluation of the CHP indicates a substantial improvement in indoor air quality, with a reduction of roughly -20 $\mu g/m^3$ (95%CI: -38, -3) in wintertime indoor $PM_{2.5}$ levels. Still there is considerable room for indoor air quality improvement given that the indoor $PM_{2.5}$ levels in treated households in W4 (GM=49 $\mu g/m^3$, 95%CI: 42, 52) were still ~10 times higher than the annual WHO annual air quality guideline (5 $\mu g/m^3$) (World Health Organization 2021).

Similar to our indoor results, several recent randomized trials with high compliance (exclusive or near exclusive) in the use of LPG stoves in rural settings with low outdoor pollution in Peru, Ghana, Guatemala, India, and Rwanda (Checkley et al. 2021; Chillrud et al. 2021; Katz et al. 2020) (Johnson et al. 2022) found lower exposures to $PM_{2.5}$ (32-69%) in the intervention group compared with controls using traditional solid fuel stoves, but even in these relatively low pollution settings, post-intervention mean exposures (range: 24 to 52 $\mu g/m^3$) were still well-above the WHO's annual air quality guideline (5 $\mu g/m^3$). A trial in urban and peri-urban Nigeria, a high pollution setting more similar to our Beijing sites, did not observe an air pollution benefit of ethanol stoves but did observe improved birth and pregnancy outcomes and blood pressure (Alexander et al. 2018; Alexander et al. 2017).

Nonetheless, comparisons of the indoor $PM_{2.5}$ benefits of the CHP in our study with previous assessments of household energy interventions in China suggest that the CHP performed well. Homes with the NISP biomass chimney stoves had modestly lower indoor PM_4 than traditional open

fire stoves (223 versus 293 µg/m^3), though post-intervention air pollution levels were still an order of magnitude higher than the current health-motivated WHO (24-h) guideline (Sinton et al. 2004). The NISP's so-called "improved" coal heating stoves unexpectedly emitted higher concentrations of PM_4 and carbon monoxide than the traditional coal stoves (Edwards et al. 2004). In southwestern China (Sichuan), a difference-in-differences analysis of an government-supported household energy package pilot (semi-gasifier cookstove, water heater, pelletized biomass fuel) observed decreased indoor $PM_{2.5}$ (24–67%) in women treated by the energy package, but greater reductions (48–70%) were observed in untreated women, a result likely influenced by an unexpectedly large transition in gas cookstoves in untreated homes during the study period (Baumgartner et al. 2019).

The relatively high post-policy indoor air pollution levels and the limited benefit to personal exposures and outdoor $PM_{2.5}$ in treated villages in our study—despite excellent compliance with the policy—is likely due to three key factors. First, a quarter of our study households had at least one tobacco smoker, which is a large contributor to personal exposures to $PM_{2.5}$ in our study settings, especially during wintertime when people tend to spend more time indoors (Li et al. 2022). Second, although has Beijing rapidly and impressively reduced outdoor $PM_{2.5}$ over the last decade (annual mean of 89 µg/m^3 in 2013 decreased to 30 µg/m^3 in 2022) (Zhang et al. 2023), the wintertime outdoor $PM_{2.5}$ levels in the treated study villages remained high enough across the study waves (range of means: 26-38 µg/m^3 in treated villages) to limit the minimum exposure achievable with an indoor stove. The contribution of outdoor sources to personal exposures is further supported by our source apportionment analyses, which showed a clear contribution of regional sources (secondary sulfur, transported dust) to personal exposures. Finally, the continued use of biomass-burning kangs likely also contributed to indoor $PM_{2.5}$ and personal exposures. Kangs are a relatively simple and culturally entrenched combined cooking and space heating technique that have been used in China for over two thousand years (Zhuang et al. 2009). Kangs are mostly fueled by wood or other biomass that is freely and widely available in our study villages. The CHP did not ban biomass burning, and we observed persistent self-reported use of kangs after heat pump installation. Continued use of traditional solid fuel stoves alongside cleaner stoves and fuels (i.e., stove stacking) has long been a barrier to achieving large reductions in indoor and personal exposures after intervention (Shankar et al. 2020). A notable exception is the HAPIN trial which attained near exclusive use of LPG stoves and dramatic reductions in personal exposures to $PM_{2.5}$ (lowered by 66% compared with controls, 70 versus 24 µg/m^3) (Johnson et al. 2022), though the impressive air quality improvements were not accompanied by any health benefits across a range of neonatal, child, and maternal outcomes (Lai et al. 2024).

To comprehensively evaluate a large-scale policy like the CHP, our study's measurement approach required extensive long-term measurements in >1000 households in multiple waves using over 500 air pollution monitors that collected thousands of hours of measurements. The scale and duration of air pollution measurement achieved in this study would not have been possible without low-cost air pollution sensors that have proliferated in the past decade. Our use of low-cost sensors to capture long-term (5-6 months) indoor air quality data in rural settings places it at the forefront of applying cutting-edge technology to understand and mitigate household air pollution. This ap-

proach is somewhat unique for China as most studies, including those using lower-cost air pollution sensing networks (Chao et al. 2021; Mei et al. 2020), focus on urban air quality, driven by consideration for urban population demographic changes and industrial, power generation, and vehicular emissions (Shen et al. 2017). By focusing on rural indoor environments, our study addresses a crucial gap, offering insights into the effectiveness of a specific policy (CHP) on a micro-scale. Future evaluations of household energy interventions might also consider longitudinal measures of air pollution that track changes over longer periods to capture delayed effects. Estimating the causal effects of the CHP required a multifaceted approach to evaluation that incorporated a study design (difference-in-differences) and analytical methods (ETWFE, causal mediation) that are less common for evaluating air quality interventions. By incorporating a broader array of metrics and considering the systemic nature of air pollution and its health impacts, through this study, we sought to provide a more nuanced understanding of an intervention's effectiveness and the ways in which it may need to be augmented or restructured to achieve desired health outcomes.

7.4 Assumptions, strengths, and limitations

The validity of our DiD approach is subject to two key assumptions (Callaway and Sant'Anna 2021; Wooldridge 2021). First, no anticipation: we assume that anticipation of the CHP did not affect outcomes prior to policy implementation and did not differ between treated and untreated villages. We selected villages that were eligible for the policy but not currently treated. It was generally understood that the policy would first be implemented in the plains areas with more updated electric grids and then gradually expand into more remote and mountainous areas of Beijing, though most of our study villages were far from Beijing's urban core. In addition to these geographical parameters, some of our study villages were assigned to the policy whereas others applied to the local government, but they were generally unaware of if or when they would be treated at the time of enrollment. Second, parallel trends: our analysis assumes that in the absence of the policy the trends in air quality and health in treated and untreated villages would have remained the same over time. Because the parallel trends assumption is based on a counterfactual it cannot be empirically verified (similar to the assumption of no unmeasured confounding). However, given that we had two pre-intervention periods prior to the 2020 and 2021 cohorts being treated, we assessed the similarity of pre-intervention trends between W1 and W2 for the never treated group and the groups eventually treated in 2020 and 2021. Estimates and 95% CIs for the difference in pre-trends are given in the Appendix for BP outcomes (Figure A9), personal exposure to $PM_{2.5}$ and black carbon (Figure A10), and self-reported respiratory outcomes (Figure A11). We did not find strong evidence for systematic differences in the pre-policy trends for any outcome, but some estimates were imprecise and tests for pre-trends are generally considered to have low power (Roth 2022). We also adjusted for relevant time-varying confounders in estimating total effects and in mediation analyses, which aims to improve the credibility of parallel trends assumption.

We also note that, in general, the addition of covariates to our "basic" ETWFE models did not meaningfully change our estimates. Nevertheless, we cannot entirely rule out the possibility that

other programs or policies differentially affected air quality or health in treated and untreated villages, which could lead to over- or under-estimation of its effects. To investigate this possibility we surveyed village leaders about other rural development or health policies and programs in their villages throughout our four-year study period and did not identify any co-implemented programs that would differentially impact villages by treatment status and affect outcomes or mediators. Though a number of municipality- and district-level COVID-19-related preventative measures were implemented during the pandemic (e.g., lockdowns, travel restrictions), our study villages were not differentially exposed to any such measures. Finally, our mediation analysis assumes no residual time-varying confounding between our mediators (air pollution and temperature) and our health outcomes. Although we measured and evaluated a large number of time-varying risk factors for BP, we cannot entirely eliminate the possibility of potential residual confounding which could over-or under-estimate the mediating effects of indoor environmental factors.

Strengths of this comprehensive, field-based assessment of the CHP include our quasi-experimental design to evaluate a real-world clean energy intervention that would be near impossible to experimentally manipulate at the scale of our study. Our study design controlled for secular changes in health and we additionally collected data on and adjusted for important time-varying covariates. It's perhaps worth noting that control for secular trends was important in our context. For personal exposure, supplementary analyses that dropped the time fixed effects or compared treated vs. untreated villages with covariate adjustment would have suggested a substantial impact of the policy (Appendix Table A26). Our numerous sensitivity analyses showed the robustness of our findings to various analytic decisions. Most previous field-based household energy intervention studies were less than two-years in duration with a single post-treatment wave (Lai et al. 2024; Quansah et al. 2017), and our four-year study enabled longer-term evaluation of compliance with the coal ban and heat pump adoption/use and their impacts on air pollution and health. Despite the logistical challenges of COVID-19 pandemic shutdowns and related government restrictions that occurred throughout half of our study period, we were able to continue the study and successfully retain participants in all 50 study villages over four years. Our large sample size of 1000 participants in 50 villages across multiple study waves enabled us to evaluate both the total effects of the policy and separately for different treatment cohorts. By comparison, the few previous field-based assessments of household energy interventions (trials and pre-post designs with controls) and blood pressure ranged in size from 44 to 324 participants (Kumar et al. 2021; Lai et al. 2024; Onakomaiya et al. 2019), with exception of HAPIN trial that enrolled ~3000 pregnant women (Ye et al. 2022).

This study also has several limitations to consider when interpreting our results. First, the COVID-19 pandemic began in the middle of our study. Although there were no recorded COVID-19 infections in our study villages during the study period, Beijing's municipality-wide preventative measures impacted all study villages at the same time (e.g., travel restrictions, closure of public spaces, lockdowns, and quarantines). Roughly half of our treated villages entered into the policy during the pandemic, which likely had some influence on its roll out. We observed the largest benefits in BP and several respiratory outcomes in villages treated before the pandemic compared with those treated after it started. However, we cannot differentiate between treatment cohort

effects attributable treatment during the COVID-19 pandemic versus other factors that different between treatment cohorts (i.e., geographic location, access to biomass, fuel prices).

Second, the CHP roll-out began in 2016 but we did not begin enrolling villages into our study until 2018. Thus, many of our study villages are farther from the urban core and generally of lower socioeconomic status than many villages treated in the first three years of the policy. Previous studies of the CHP suggest that treated villages of all socioeconomic levels benefited from less-polluted and warmer indoor environments, but that the benefits were larger in wealthier villages that were more likely to use the heat pumps more often and set to a higher indoor temperature (Barrington-Leigh et al. 2019; Meng et al. 2023). Further, most of our study villages had relatively easy access to (free) biomass fuel, and may be more likely to use biomass-burning kangs compared with villages near the urban core where biomass fuel is less readily available. Thus our results may not be generalizable to all of rural and peri-urban Beijing, especially to the more urbanized, wealthier villages treated between 2016 and 2018, and may underestimate the impacts of the policy on indoor environmental factors that were important cardio-respiratory health mediators in our study.

Third, like any field-based study, we had a number of constraints with data collection. We were unable to measure indoor air quality in W1 due to logistical and budget constraints, and thus cannot directly estimate the effects of the policy on indoor $PM_{2.5}$ for the 10 villages treated in 2019, which is also the treatment cohort that experienced the greatest health benefits. Similarly, we were unable to collect blood samples in the last wave because all of our measurements were conducted in participant homes rather than clinics to avoid group contact during the pandemic. In addition, our study logistics required visiting 50 villages over a period of just several months. Thus, we were unable to return to villages if a previously enrolled participant was not at home at the time that staff visited the village. In such instances, we either randomly selected either another eligible participant in the same home or we randomly selected another household with eligible participants from the village roster and our study participants differed slightly across waves. Though this is unlikely to impact our findings since our village-level study and analysis is robust to participation of a random sample of participants in each wave, and there were few notable differences in key demographic characteristics or health behaviors between participants who contributed to a different number of waves or between participants across each of the three waves that included individual-level measurements.

Fourth, respiratory symptoms were self-reported and thus our estimated effects on respiratory symptoms must be interpreted with caution. Participants in our study were aware of their treatment status, and knowledge of being treated by a policy aimed at reducing local air pollution may have affected their reporting of the perceived health benefits of intervention (Peel et al. 2015). Previous trials of improved biomass stoves, for example, noted a tendency of participants to report favorable response to the stove regardless of its physiologic efficacy (Burwen and Levine 2012; Smith-Sivertsen et al. 2009). Such reporting bias could have inaccurately increased the estimated effect of the policy, though we did not observe consistent effects across all respiratory outcomes or treatment cohorts,

which one might expect if treated participants were inclined to give more favorable responses. Our study also benefited from co-measurement of BP, an objective measure that is less likely to be biased by participant or staff awareness of treatment status because our staff consistently followed strict quality control guidelines for measurement across all study villages and homes, regardless of treatment status.

Finally, some remarks concerning power and uncertainty. We do not distinguish our results based on traditional notions of "statistical significance" (Wasserstein et al. 2019), but given the large number of outcomes and wide range of reported estimates and uncertainties, questions about the power of our design are relevant. Using observed effect sizes to ask questions about power (sometimes called 'post-hoc power calculations') is illogical given the direct relationship between power and observed p-values or confidence limits (Hoenig and Heisey 2001). Instead we aim to put our estimates in context using a retrospective design analysis (Gelman and Carlin 2014). Retrospective design analyses use plausible values for hypothetical effect sizes to ask about the strength of evidence a replicated study with our design and level of precision would be likely to provide. The main quantities for a design analysis include the probability that the replicated estimate would be "statistically significant" (power), the probability that the replicated estimate would have the wrong sign ("S-bias"), and the ratio of the replicated estimate divided by the true effect size (the "exaggeration ratio" or "M-bias"). We report these values for a range of conservative, but plausible, effect sizes below.

For blood pressure we used a value of a 2.5 mmHg decrease in systolic or a 2 mmHg decrease in diastolic (Baumgartner et al. 2011; Steenland et al. 2018). For self-reported respiratory outcomes we assumed hypothetical effects on the order of a 10% decline for each outcome (translated into absolute percentage point declines), and for inflammatory markers we hypothesized true effects of a roughly 5% decline (Pope III et al. 2004; Tang et al. 2020), again translated into absolute terms.

The full results for all health outcomes are shown in the Appendix (Table A27). As an example, consider our estimate of the 7.5 percentage point decrease in any respiratory symptom (95% CI 2.3, 12.7). Assuming that the true effect of the policy was a reduction of 5 percentage points, a replicated estimate with our design features and precision would have 47% power, virtually no chance of reporting the wrong sign (S-bias = 0), and in expectation the estimated effect will be only 1.4 times too high. Similarly, if the true effect of the policy on brachial systolic BP was a small reduction of 2.5 mmHg, a replication of our study design would have roughly 73% power, a 2% chance of having the wrong sign, and an average exaggeration ratio of just 0.5, meaning it would be unlikely to exaggerate the true effect.

The greater precision for blood pressure and respiratory outcomes lends greater confidence to these estimated impacts. On the other hand, our estimates of the impact on inflammatory markers contain greater uncertainty (larger SEs), largely due to logistical constraints that prevented data collection in the last wave. For example, our adjusted ATT for the effect of the CHP on IL-6 was an increase of 0.8 pg/mL but our data are compatible with effects as small as a 0.3 pg/mL decrease and as large as a 2.0 pg/mL increase. Table A27 shows that if the true effect of the CHP on IL-6 is a modest 0.2 pg/mL decline, a replication using our design and level of precision will have roughly

6% power, a 39% chance of having the wrong sign, and the estimated effect will on average be nearly 2.5 times too high. Because we chose generally conservative values for hypothetical effects, design analyses with larger hypothetical true effects would tend to show greater power, lower likelihood of reporting the wrong sign, and less exaggerated effects.

Because of the inherent uncertainty surrounding hypothesized "true" effects, it often makes sense to conduct retrospective design analyses for a range of effect sizes. We show estimates for BP reductions of 0 to 4 mmHg in Appendix Figure A12. The HEI-funded objectives of this study were designed based on pre-study power calculations for blood pressure reductions of roughly 0.25 standard deviations (roughly 4 mmHg for older rural population with an average SD of 16 mmHg). Appendix Figure A12 shows that if the true effect of the policy were similar to our pre-study estimates of 0.25 SD difference in systolic BP, a study with our level of precision would have >90% power, virtually no chance of reporting an estimate with the wrong sign, and an average exaggeration ratio of just 1.04. This suggests that our design was well-powered to detect clinically meaningful effects (Rahimi et al. 2021) on blood pressure.

8 Implications of Findings

In this comprehensive field-based assessment of the rural Clean Heating Policy in Beijing, we observed high fidelity and compliance with the policy in our study villages and households where nearly all households in treated villages stopped using coal and shifted to electric-powered heaters. Exposure to the policy reduced blood pressure and self-reported chronic respiratory symptoms, and these health benefits were mediated by reductions in indoor $PM_{2.5}$ and improvements in home temperature. We did not observe the same benefits of the policy on outdoor air quality or personal exposures, likely because the relatively high contribution of other regional and local air pollution sources to outdoor and personal exposures may have masked the benefits from a single source reduction. We also did not observe benefits of the policy on different measures of inflammation and oxidative stress in the sub-sample of participants with biomarker assessment, even though we observed respiratory symptoms and BP benefits of the policy in a sensitivity analysis limited to the same participants. Still, our overall findings indicate that this ambitious policy achieved its goals in dramatically reducing residential coal burning and improving indoor environmental quality, which provided modest benefits to health.

Our results showing an indoor environment and cardio-respiratory health benefit of a real-world, large-scale clean energy policy are timely, as they are synchronous with ongoing and planned clean energy policies in China and other countries in a global effort to "ensure access to affordable, reliable, sustainable, and modern energy for all" (Sustainable Development Goal-7) and directly respond to a recent call-to-action from global cardiovascular societies that emphasized the urgent need for interventional studies that inform targeted pollution-reducing strategies to reduce cardiovascular disease (Brauer et al. 2021).

9 Data Availability Statement

The de-identified data, code, documentation, and study resources including the standard operating procedures for all study measurements are openly available on the Open Science Foundation (OSF) platform.

10 Acknowledgements

In addition to the HEI funding that supported this work, we also acknowledge support from the Canadian Institutes for Health Research (CIHR #159477) and the Social Sciences and Humanities Research Council (SSHRC #430-2017-00998 and #435-2016-0531). HEI funding supported the addition of indoor temperature and indoor air quality measurements starting in wave 2 to wave 4 and also fully supported the data collection campaigns in waves 3 and 4. None of these funders had any role in study design, data collection and analysis, decision to publish, or preparation of this report.

We would like to thank and acknowledge the over 1400 study participants and the over 50 field staff members who assisted with data collection and laboratory analysis. This work would not have been possible without the dedicated efforts and contributions of investigators and trainees who contributed to the development of ideas, data collection, and results that are reported here and in publications resulting from this work. Here, we wish to acknowledge Koren Mann, Arijit Nandi, Robert Platt, Kennedy Hirst, Enkhuun Byambadorj, Kaibing Xue, and Martha Lee. We also acknowledge the efforts of project coordinators in Canada and China: Xinwei Liu, Jing Shang, Xiaoxia Hu, Jian Ma, Leona Siaw, Neha Ahmed, and Laojie Li.

11 References

Ahmed T, Dutkiewicz VA, Shareef A, Tuncel G, Tuncel S, Husain L. 2009. Measurement of black carbon (BC) by an optical method and a thermal-optical method: Intercomparison for four sites. Atmospheric Environment 43:6305–6311; doi:10.1016/j.atmosenv.2009.09.031.

Alexander DA, Northcross A, Karrison T, Morhasson-Bello O, Wilson N, Atalabi OM, et al. 2018. Pregnancy outcomes and ethanol cook stove intervention: A randomized-controlled trial in Ibadan, Nigeria. Environment International 111:152–163; doi:10.1016/j.envint.2017.11.021.

Alexander D, Northcross A, Wilson N, Dutta A, Pandya R, Ibigbami T, et al. 2017. Randomized Controlled Ethanol Cookstove Intervention and Blood Pressure in Pregnant Nige-

rian Women. American Journal of Respiratory and Critical Care Medicine 195:1629–1639; doi:10.1164/rccm.201606-1177OC.

An L, Hong B, Cui X, Geng Y, Ma X. 2021. Outdoor thermal comfort during winter in China's cold regions: A comparative study. Science of The Total Environment 768:144464; doi:10.1016/j.scitotenv.2020.144464.

Anderson GB, Peng RD, Ferreri JM. 2016. Weathermetrics: Functions to Convert Between Weather Metrics.

Archer-Nicholls S, Carter E, Kumar R, Xiao Q, Liu Y, Frostad J, et al. 2016. The Regional Impacts of Cooking and Heating Emissions on Ambient Air Quality and Disease Burden in China. Environmental Science & Technology 50:9416–9423; doi:10.1021/acs.est.6b02533.

Arel-Bundock V. 2024. Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests.

Baker AC, Larcker DF, Wang CCY. 2022. How much should we trust staggered difference-in-differences estimates? Journal of Financial Economics 144:370–395; doi:10.1016/j.jfineco.2022.01.004.

Barrington-Leigh C, Baumgartner J, Carter E, Robinson BE, Tao S, Zhang Y. 2019. An evaluation of air quality, home heating and well-being under Beijing's programme to eliminate household coal use. Nature Energy 4:416–423; doi:10.1038/s41560-019-0386-2.

Baumgartner J, Carter E, Schauer JJ, Ezzati M, Daskalopoulou SS, Valois M-F, et al. 2018. Household air pollution and measures of blood pressure, arterial stiffness and central haemodynamics. Heart 104:1515–1521; doi:10.1136/heartjnl-2017-312595.

Baumgartner J, Clark S, Carter E, Lai A, Zhang Y, Shan M, et al. 2019. Effectiveness of a Household Energy Package in Improving Indoor Air Quality and Reducing Personal Exposures in Rural China. Environmental Science & Technology 53:9306–9316; doi:10.1021/acs.est.9b02061.

Baumgartner J, Schauer JJ, Ezzati M, Lu L, Cheng C, Patz JA, et al. 2011. Indoor Air Pollution and Blood Pressure in Adult Women Living in Rural China. Environmental Health Perspectives 119:1390–1395; doi:10.1289/ehp.1003371.

Beltramo T, Levine DI. 2013. The effect of solar ovens on fuel use, emissions and health: Results from a randomised controlled trial. Journal of Development Effectiveness 5:178–207; doi:10.1080/19439342.2013.775177.

Brauer M, Casadei B, Harrington RA, Kovacs R, Sliwa K, the WHF Air Pollution Expert

Group. 2021. Taking a Stand Against Air Pollution—The Impact on Cardiovascular Disease: A Joint Opinion From the World Heart Federation, American College of Cardiology, American Heart Association, and the European Society of Cardiology. Circulation 143; doi:10.1161/CIRCULATIONAHA.120.052666.

Burwen J, Levine DI. 2012. A rapid assessment randomized-controlled trial of improved cookstoves in rural Ghana. Energy for Sustainable Development 16:328–338; doi:10.1016/j.esd.2012.04.001.

Callaway B. 2020. Difference-in-Differences for Policy Evaluation. In: *Handbook of Labor, Human Resources and Population Economics* (K.F. Zimmermann, ed). Springer International Publishing:Cham. 1–61.

Callaway B, Sant'Anna PHC. 2021. Difference-in-Differences with multiple time periods. Journal of Econometrics 225:200–230; doi:10.1016/j.jeconom.2020.12.001.

Cameron AC, Miller DL. 2015. A practitioner's guide to cluster-robust inference. Journal of Human Resources 50: 317–372.

Card D, Krueger AB. 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. American Economic Review 84: 772–93.

Chao C-Y, Zhang H, Hammer M, Zhan Y, Kenney D, Martin RV, et al. 2021. Integrating Fixed Monitoring Systems with Low-Cost Sensors to Create High-Resolution Air Quality Maps for the Northern China Plain Region. ACS Earth and Space Chemistry 5:3022–3035; doi:10.1021/acsearthspacechem.1c00174.

Checkley W, Williams KN, Kephart JL, Fandiño-Del-Rio M, Steenland NK, Gonzales GF, et al. 2021. Effects of a Household Air Pollution Intervention with Liquefied Petroleum Gas on Cardiopulmonary Outcomes in Peru. A Randomized Controlled Trial. American Journal of Respiratory and Critical Care Medicine 203:1386–1397; doi:10.1164/rccm.202006-2319OC.

Chillrud SN, Ae-Ngibise KA, Gould CF, Owusu-Agyei S, Mujtaba M, Manu G, et al. 2021. The effect of clean cooking interventions on mother and child personal exposure to air pollution: Results from the Ghana Randomized Air Pollution and Health Study (GRAPHS). Journal of Exposure Science & Environmental Epidemiology 31:683–698; doi:10.1038/s41370-021-00309-5.

Clark S, Carter E, Shan M, Ni K, Niu H, Tseng JTW, et al. 2017. Adoption and use of a semi-gasifier cooking and water heating stove and fuel intervention in the Tibetan Plateau, China. Environmental Research Letters 12:075004; doi:10.1088/1748-9326/aa751e.

Costello BT, Schultz MG, Black JA, Sharman JE. 2015. Evaluation of a Brachial Cuff and Suprasys-

tolic Waveform Algorithm Method to Noninvasively Derive Central Blood Pressure. American Journal of Hypertension 28:480–486; doi:10.1093/ajh/hpu163.

D'Amato M, Molino A, Calabrese G, Cecchi L, Annesi-Maesano I, D'Amato G. 2018. The impact of cold on the respiratory tract and its consequences to respiratory health. Clinical and Translational Allergy 8:20; doi:10.1186/s13601-018-0208-9.

Dai Q, Liu B, Bi X, Wu J, Liang D, Zhang Y, et al. 2020. Dispersion Normalized PMF Provides Insights into the Significant Changes in Source Contributions to PM $_{2.5}$ after the COVID-19 Outbreak. Environmental Science & Technology 54:9917–9927; doi:10.1021/acs.est.0c02776.

Danesh J, Kaptoge S, Mann AG, Sarwar N, Wood A, Angleman SB, et al. 2008. Long-term interleukin-6 levels and subsequent risk of coronary heart disease: Two new prospective studies and a systematic review. PLoS medicine 5:e78; doi:10.1371/journal.pmed.0050078.

Dart AM, Kingwell BA. 2001. Pulse pressure—a review of mechanisms and clinical relevance. Journal of the American College of Cardiology 37:975–984; doi:10.1016/S0735-1097(01)01108-1.

Dispersed Coal Management Research Group . 2023. China Dispersed Coal Governance Report.

Dockery DW, Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, et al. 2013. Effect of air pollution control on mortality and hospital admissions in Ireland. Research Report (Health Effects Institute) 3–109.

Dominici F, Greenstone M, Sunstein CR. 2014. Science and regulation. Particulate matter matters. Science (New York, NY) 344:257–9; doi:10.1126/science.1247348.

Dong G-H, Qian Z(Min), Xaverius PK, Trevathan E, Maalouf S, Parker J, et al. 2013. Association Between Long-Term Air Pollution and Increased Blood Pressure and Hypertension in China. Hypertension 61:578–584; doi:10.1161/HYPERTENSIONAHA.111.00003.

Duan X, Jiang Y, Wang B, Zhao X, Shen G, Cao S, et al. 2014. Household fuel use for cooking and heating in China: Results from the first Chinese Environmental Exposure-Related Human Activity Patterns Survey (CEERHAPS). Applied Energy 136:692–703; doi:10.1016/j.apenergy.2014.09.066.

Edwards RD, Smith KR, Zhang J, Ma Y. 2004. Implications of changes in household stoves and fuel use in China. Energy Policy 32:395–411; doi:10.1016/S0301-4215(02)00309-9.

Emerging Risk Factors Collaboration. 2012. C-reactive protein, fibrinogen, and cardiovascular

disease prediction. New England Journal of Medicine 367: 1310–1320.

Ezzati M, Baumgartner JC. 2017. Household energy and health: Where next for research and practice? Lancet (London, England) 389:130–132; doi:10.1016/S0140-6736(16)32506-5.

Food and Drug Administration. 2018. Bioanalytical Method Validation Guidance for Industry.

Gao J, Wang K, Wang Y, Liu S, Zhu C, Hao J, et al. 2018. Temporal-spatial characteristics and source apportionment of PM2.5 as well as its associated chemical species in the Beijing-Tianjin-Hebei region of China. Environmental Pollution 233:714–724; doi:10.1016/j.envpol.2017.10.123.

GBD MAPS Working Group. 2016. Burden of disease attributable to coal-burning and other air pollution sources in China.

Gelman A, Carlin J. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspectives on Psychological Science 9:641–651; doi:10.1177/1745691614551642.

Goin DE, Riddell CA. 2023. Comparing Two-way Fixed Effects and New Estimators for Difference-in-Differences: A Simulation Study and Empirical Example. Epidemiology 34:535; doi:10.1097/EDE.0000000000001611.

Goodman-Bacon A. 2021. Difference-in-differences with variation in treatment timing. Journal of Econometrics 225:254–277; doi:10.1016/j.jeconom.2021.03.014.

Gould CF, Bejarano ML, Kioumourtzoglou M-A, Lee AG, Pillarisetti A, Schlesinger SB, et al. 2023. Widespread Clean Cooking Fuel Scale-Up and under-5 Lower Respiratory Infection Mortality: An Ecological Analysis in Ecuador, 1990–2019. Environmental Health Perspectives 131:037017; doi:10.1289/EHP11016.

Hoenig JM, Heisey DM. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. The American Statistician 55:19–24; doi:10.1198/000313001300339897.

Huang W, Wang G, Lu S-E, Kipen H, Wang Y, Hu M, et al. 2012. Inflammatory and Oxidative Stress Responses of Healthy Young Adults to Changes in Air Quality during the Beijing Olympics. American Journal of Respiratory and Critical Care Medicine 186:1150–1159; doi:10.1164/rccm.201205-0850OC.

Johnson M, Pillarisetti A, Piedrahita R, Balakrishnan K, Peel JL, Steenland K, et al. 2022. Exposure Contrasts of Pregnant Women during the Household Air Pollution Intervention Network Randomized Controlled Trial. Environmental Health Perspectives 130:097005; doi:10.1289/EHP10295.

Johnston FH, Hanigan IC, Henderson SB, Morgan GG. 2013. Evaluation of interventions to reduce air pollution from biomass smoke on mortality in Launceston, Australia: Retrospective analysis of daily mortality, 1994-2007. BMJ 346:e8446–e8446; doi:10.1136/bmj.e8446.

Kanagasabai T, Xie W, Yan L, Zhao L, Carter E, Guo D, et al. 2022. Household Air Pollution and Blood Pressure, Vascular Damage, and Subclinical Indicators of Cardiovascular Disease in Older Chinese Adults. American Journal of Hypertension 35:121–131; doi:10.1093/ajh/hpab141.

Kashtan YS, Nicholson M, Finnegan C, Ouyang Z, Lebel ED, Michanowicz DR, et al. 2023. Gas and Propane Combustion from Stoves Emits Benzene and Increases Indoor Air Pollution. Environmental Science & Technology 57:9653–9663; doi:10.1021/acs.est.2c09289.

Katz J, Tielsch JM, Khatry SK, Shrestha L, Breysse P, Zeger SL, et al. 2020. Impact of Improved Biomass and Liquid Petroleum Gas Stoves on Birth Outcomes in Rural Nepal: Results of 2 Randomized Trials. Global Health: Science and Practice 8:372–382; doi:10.9745/GHSP-D-20-00011.

Keele L, Tingley D, Yamamoto T. 2015. Identifying mechanisms behind policy interventions via causal mediation analysis. Journal of Policy Analysis and Management 34: 937–963.

Khuzestani RB, Schauer JJ, Wei Y, Zhang Y, Zhang Y. 2017. A non-destructive optical color space sensing system to quantify elemental and organic carbon in atmospheric particulate matter on Teflon and quartz filters. Atmospheric Environment 149:84–94; doi:10.1016/j.atmosenv.2016.11.002.

Kipen H, Rich D, Huang W, Zhu T, Wang G, Hu M, et al. 2010. Measurement of inflammation and oxidative stress following drastic changes in air pollution during the Beijing Olympics: A panel study approach. Annals of the New York Academy of Sciences 1203:160–167; doi:10.1111/j.1749-6632.2010.05638.x.

Kumar N, Phillip E, Cooper H, Davis M, Langevin J, Clifford M, et al. 2021. Do improved biomass cookstove interventions improve indoor air quality and blood pressure? A systematic review and meta-analysis. Environmental Pollution 290:117997; doi:10.1016/j.envpol.2021.117997.

Lai. 2019. Relative contributions of household solid fuel use and outdoor air pollution to chemical components of personal PM2.5 exposures. Indoor Air-international Journal of Indoor Air Quality and Climate.

Lai PS, Lam NL, Gallery B, Lee AG, Adair-Rohani H, Alexander D, et al. 2024. Household Air Pollution Interventions to Improve Health in Low- and Middle-Income Countries: An Official American Thoracic Society Research Statement. American Journal of Respiratory and Critical

Care Medicine 209:909–927; doi:10.1164/rccm.202402-0398ST.

Lee M, Carter E, Yan L, Chan Q, Elliott P, Ezzati M, et al. 2021. Determinants of personal exposure to PM2.5 and black carbon in Chinese adults: A repeated-measures study in villages using solid fuel energy. Environment International 146:106297; doi:10.1016/j.envint.2020.106297.

Lewington S, LiMing L, Sherliker P, Yu G, Millwood I, Zheng B, et al. 2012. Seasonal variation in blood pressure and its relationship with outdoor temperature in 10 diverse regions of China: The China Kadoorie Biobank. Journal of hypertension 30: 1383.

Li X, Baumgartner J, Harper S, Zhang X, Sternbach T, Barrington-Leigh C, et al. 2022. Field measurements of indoor and community air quality in rural Beijing before, during, and after the COVID-19 lockdown. Indoor Air 32:e13095; doi:10.1111/ina.13095.

Lindemann U, Stotz A, Beyer N, Oksa J, Skelton DA, Becker C, et al. 2017. Effect of indoor temperature on physical performance in older adults during days with normal temperature and heat waves. International journal of environmental research and public health 14; doi:10.3390/ijerph14020186.

Liu B, Wu J, Zhang J, Wang L, Yang J, Liang D, et al. 2017. Characterization and source apportionment of PM2.5 based on error estimation from EPA PMF 5.0 model at a medium city in China. Environmental Pollution 222:10–22; doi:10.1016/j.envpol.2017.01.005.

Lowe A, Harrison W, El-Aklouk E, Ruygrok P, Al-Jumaily AM. 2009. Non-invasive model-based estimation of aortic pulse pressure using suprasystolic brachial pressure waveforms. Journal of Biomechanics 42:2111–2115; doi:10.1016/j.jbiomech.2009.05.029.

Lv Y, Zhu R, Xie J, Yoshino H. 2022. Indoor environment and the blood pressure of elderly in the cold region of China. Indoor and Built Environment 31:2482–2498; doi:10.1177/1420326X221109510.

Manning WG, Mullahy J. 2001. Estimating log models: To transform or not to transform? Journal of Health Economics 20:461–494; doi:10.1016/S0167-6296(01)00086-8.

McCracken JP, Smith KR, Díaz A, Mittleman MA, Schwartz J. 2007. Chimney Stove Intervention to Reduce Long-term Wood Smoke Exposure Lowers Blood Pressure among Guatemalan Women. Environmental Health Perspectives 115:996–1001; doi:10.1289/ehp.9888.

McCracken J, Smith KR, Stone P, Díaz A, Arana B, Schwartz J. 2011. Intervention to Lower Household Wood Smoke Exposure in Guatemala Reduces ST-Segment Depression on Electrocardiograms. Environmental Health Perspectives 119:1562–1568; doi:10.1289/ehp.1002834.

Mei H, Han P, Wang Y, Zeng N, Liu D, Cai Q, et al. 2020. Field Evaluation of Low-Cost Particulate Matter Sensors in Beijing. Sensors 20:4381; doi:10.3390/s20164381.

Meng W, Zhu L, Liang Z, Xu H, Zhang W, Li J, et al. 2023. Significant but Inequitable Cost-Effective Benefits of a Clean Heating Campaign in Northern China. Environmental Science & Technology 57:8467–8475; doi:10.1021/acs.est.2c07492.

Naimi AI, Kaufman JS, MacLehose RF. 2014. Mediation misgivings: Ambiguous clinical and public health interpretations of natural direct and indirect effects. International journal of epidemiology 43:1656–61; doi:10.1093/ije/dyu107.

Ni K, Carter E, Schauer JJ, Ezzati M, Zhang Y, Niu H, et al. 2016. Seasonal variation in outdoor, indoor, and personal air pollution exposures of women using wood stoves in the Tibetan Plateau: Baseline assessment for an energy intervention study. Environment International 94:449–457; doi:10.1016/j.envint.2016.05.029.

Niu J, Chen X, Sun S. 2024. China's Coal Ban policy: Clearing skies, challenging growth. Journal of Environmental Management 349:119420; doi:10.1016/j.jenvman.2023.119420.

Olson MR, Graham E, Hamad S, Uchupalanun P, Ramanathan N, Schauer JJ. 2016. Quantification of elemental and organic carbon in atmospheric particulate matter using color space sensing—hue, saturation, and value (HSV) coordinates. Science of The Total Environment 548–549:252–259; doi:10.1016/j.scitotenv.2016.01.032.

Onakomaiya D, Gyamfi J, Iwelunmor J, Opeyemi J, Oluwasanmi M, Obiezu-Umeh C, et al. 2019. Implementation of clean cookstove interventions and its effects on blood pressure in low-income and middle-income countries: Systematic review. BMJ Open 9:e026517; doi:10.1136/bmjopen-2018-026517.

Pearl J. 2000. *Causality: Models, reasoning, and inference*. Cambridge University Press:Cambridge, U.K. ; New York.

Pearson TA, Mensah GA, Alexander RW, Anderson JL, Cannon RO 3rd, Criqui M, et al. 2003. Markers of inflammation and cardiovascular disease: Application to clinical and public health practice: A statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association. Circulation 107: 499–511.

Peel JL, Baumgartner J, Wellenius GA, Clark ML, Smith KR. 2015. Are Randomized Trials Necessary to Advance Epidemiologic Research on Household Air Pollution? Current Epidemiology Reports 2:263–270; doi:10.1007/s40471-015-0054-4.

Pope III CA, Hansen ML, Long RW, Nielsen KR, Eatough NL, Wilson WE, et al. 2004. Ambient particulate air pollution, heart rate variability, and blood markers of inflammation in a panel of elderly subjects. Environmental health perspectives 112:339–45; doi:10.1289/ehp.6588.

Quansah R, Semple S, Ochieng CA, Juvekar S, Armah FA, Luginaah I, et al. 2017. Effectiveness of interventions to reduce household air pollution and/or improve health in homes using solid fuel in low-and-middle income countries: A systematic review and meta-analysis. Environment International 103:73–90; doi:10.1016/j.envint.2017.03.010.

Rahimi K, Bidel Z, Nazarzadeh M, Copland E, Canoy D, Ramakrishnan R, et al. 2021. Pharmacological blood pressure lowering for primary and secondary prevention of cardiovascular disease across different levels of blood pressure: An individual participant-level data meta-analysis. The Lancet 397:1625–1636; doi:10.1016/S0140-6736(21)00590-0.

Rehfuess EA, Puzzolo E, Stanistreet D, Pope D, Bruce NG. 2014. Enablers and Barriers to Large-Scale Uptake of Improved Solid Fuel Stoves: A Systematic Review. Environmental Health Perspectives 122:120–130; doi:10.1289/ehp.1306639.

Rich DQ, Kipen HM, Huang W, Wang G, Wang Y, Zhu P, et al. 2012. Association Between Changes in Air Pollution Levels During the Beijing Olympics and Biomarkers of Inflammation and Thrombosis in Healthy Young Adults. JAMA 307; doi:10.1001/jama.2012.3488.

Ridker PM. 2001. High-sensitivity C-reactive protein: Potential adjunct for global risk assessment in the primary prevention of cardiovascular disease. Circulation 103: 1813–8.

Ridker PM, Hennekens CH, Buring JE, Rifai N. 2000. C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. The New England journal of medicine 342:836–43; doi:10.1056/NEJM200003233421202.

Romieu I, Riojas-Rodríguez H, Marrón-Mares AT, Schilmann A, Perez-Padilla R, Masera O. 2009. Improved Biomass Stove Intervention in Rural Mexico: Impact on the Respiratory Health of Women. American Journal of Respiratory and Critical Care Medicine 180:649–656; doi:10.1164/rccm.200810-1556OC.

Rosenthal J, Quinn A, Grieshop AP, Pillarisetti A, Glass RI. 2018. Clean cooking and the SDGs: Integrated analytical approaches to guide energy interventions for health and environment goals. Energy for sustainable development : the journal of the International Energy Initiative 42:152–159; doi:10.1016/j.esd.2017.11.003.

Roth J. 2022. Pretest with Caution: Event-Study Estimates after Testing for Parallel Trends. American Economic Review: Insights 4:305–322; doi:10.1257/aeri.20210236.

RTI International. 2009. Standard Operating Procedure for the X-Ray Fluorescence Analysis of Particulate Matter Deposits on Teflon Filters: PM Xrf Analysis.

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. 1st ed. Wiley.

Rückerl R, Greven S, Ljungman P, Aalto P, Antoniades C, Bellander T, et al. 2007. Air pollution and inflammation (interleukin-6, C-reactive protein, fibrinogen) in myocardial infarction survivors. Environmental health perspectives 115:1072–80; doi:10.1289/ehp.10021.

Ruiz-Mercado I, Canuz E, Walker JL, Smith KR. 2013. Quantitative metrics of stove adoption using Stove Use Monitors (SUMs). Biomass and Bioenergy 57:136–148; doi:10.1016/j.biombioe.2013.07.002.

Scott AJ, Scarrott C. 2011. Impacts of residential heating intervention measures on air quality and progress towards targets in Christchurch and Timaru, New Zealand. Atmospheric Environment 45:2972–2980; doi:10.1016/j.atmosenv.2010.09.008.

Secrest MH, Schauer JJ, Carter EM, Lai AM, Wang Y, Shan M, et al. 2016. The oxidative potential of PM2.5 exposures from indoor and outdoor sources in rural China. The Science of the Total Environment 571:1477–1489; doi:10.1016/j.scitotenv.2016.06.231.

Shakya KM, Peltier RE. 2015. Non-sulfate sulfur in fine aerosols across the United States: Insight for organosulfate prevalence. Atmospheric environment (Oxford, England : 1994) 100:159–166; doi:10.1016/j.atmosenv.2014.10.058.

Shang J, Zhang Y, Schauer JJ, Tian J, Hua J, Han T, et al. 2020. Associations between source-resolved PM2.5 and airway inflammation at urban and rural locations in Beijing. Environment International 139:105635; doi:10.1016/j.envint.2020.105635.

Shankar AV, Quinn AK, Dickinson KL, Williams KN, Masera O, Charron D, et al. 2020. Everybody stacks: Lessons from household energy case studies to inform design principles for clean energy transitions. Energy Policy 141:111468; doi:10.1016/j.enpol.2020.111468.

Shen H, Tao S, Chen Y, Ciais P, Güneralp B, Ru M, et al. 2017. Urbanization-induced population migration has reduced ambient PM $_{2.5}$ concentrations in China. Science Advances 3:e1700300; doi:10.1126/sciadv.1700300.

Sinton JE, Smith KR, Peabody JW, Yaping L, Xiliang Z, Edwards R, et al. 2004. An assessment of programs to promote improved household stoves in China. Energy for Sustainable Development 8:33–52; doi:10.1016/S0973-0826(08)60465-2.

Smith-Sivertsen T, Díaz E, Pope D, Lie RT, Díaz A, McCracken J, et al. 2009. Effect of Re-

ducing Indoor Air Pollution on Women's Respiratory Symptoms and Lung Function: The RESPIRE Randomized Trial, Guatemala. American Journal of Epidemiology 170:211–220; doi:10.1093/aje/kwp100.

Snider G, Carter E, Clark S, Tseng J(TzuW, Yang X, Ezzati M, et al. 2018. Impacts of stove use patterns and outdoor air quality on household air pollution and cardiovascular mortality in southwestern China. Environment International 117:116–124; doi:10.1016/j.envint.2018.04.048.

Song C, Liu B, Cheng K, Cole MA, Dai Q, Elliott RJR, et al. 2023. Attribution of Air Quality Benefits to Clean Winter Heating Policies in China: Combining Machine Learning with Causal Inference. Environmental Science & Technology 57:17707–17717; doi:10.1021/acs.est.2c06800.

Steenland K, Pillarisetti A, Kirby M, Peel J, Clark M, Checkley W, et al. 2018. Modeling the potential health benefits of lower household air pollution after a hypothetical liquified petroleum gas (LPG) cookstove intervention. Environment International 111:71–79; doi:10.1016/j.envint.2017.11.018.

Sternbach TJ, Harper S, Li X, Zhang X, Carter E, Zhang Y, et al. 2022. Effects of indoor and outdoor temperatures on blood pressure and central hemodynamics in a wintertime longitudinal study of Chinese adults. Journal of Hypertension 40:1950–1959; doi:10.1097/HJH.0000000000003198.

Sullivan AP, Holden AS, Patterson LA, McMeeking GR, Kreidenweis SM, Malm WC, et al. 2008. A method for smoke marker measurements and its potential application for determining the contribution of biomass burning from wildfires and prescribed fires to ambient PM2.5 organic carbon. Journal of Geophysical Research: Atmospheres 113; doi:10.1029/2008JD010216.

Sun L, Abraham S. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics 225:175–199; doi:10.1016/j.jeconom.2020.09.006.

Tan X, Chen G, Chen K. 2023. Clean heating and air pollution: Evidence from Northern China. Energy Reports 9:303–313; doi:10.1016/j.egyr.2022.11.166.

Tang H, Cheng Z, Li N, Mao S, Ma R, He H, et al. 2020. The short- and long-term associations of particulate matter with inflammation and blood coagulation markers: A meta-analysis. Environmental Pollution 267:115630; doi:10.1016/j.envpol.2020.115630.

Tao J, Zhang L, Cao J, Zhang R. 2017. A review of current knowledge concerning PM2.5 chemical composition, aerosol optical properties and their relationships across China. Atmospheric Chemistry and Physics 17:9485–9518; doi:10.5194/acp-17-9485-2017.

Thompson RJ, Li J, Weyant CL, Edwards R, Lan Q, Rothman N, et al. 2019. Field Emission Measurements of Solid Fuel Stoves in Yunnan, China Demonstrate Dominant Causes of Uncertainty in Household Emission Inventories. Environmental Science & Technology 53:3323–3330; doi:10.1021/acs.est.8b07040.

Tuck MK, Chan DW, Chia D, Godwin AK, Grizzle WE, Krueger KE, et al. 2009. Standard Operating Procedures for Serum and Plasma Collection: Early Detection Research Network Consensus Statement *Standard Operating Procedure Integration Working Group*. Journal of Proteome Research 8:113–117; doi:10.1021/pr800545q.

van Buuren S, Groothuis-Oudshoorn K. 2011. **Mice** : Multivariate Imputation by Chained Equations in *R*. Journal of Statistical Software 45; doi:10.18637/jss.v045.i03.

Van Donkelaar A, Hammer MS, Bindle L, Brauer M, Brook JR, Garay MJ, et al. 2021. Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. Environmental Science & Technology 55:15287–15300; doi:10.1021/acs.est.1c05309.

VanderWeele TJ. 2015. *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press:New York.

Volckens J, Quinn C, Leith D, Mehaffy J, Henry CS, Miller-Lionberg D. 2017. Development and evaluation of an ultrasonic personal aerosol sampler. Indoor air 27:409–416; doi:10.1111/ina.12318.

Wang Q, Zhao Q, Wang G, Wang B, Zhang Y, Zhang J, et al. 2020. The association between ambient temperature and clinical visits for inflammation-related diseases in rural areas in China. Environmental Pollution 261:114128; doi:10.1016/j.envpol.2020.114128.

Wang Y, Zhang Y, Schauer JJ, De Foy B, Guo B, Zhang Y. 2016. Relative impact of emissions controls and meteorology on air pollution mitigation associated with the Asia-Pacific Economic Cooperation (APEC) conference in Beijing, China. Science of The Total Environment 571:1467–1476; doi:10.1016/j.scitotenv.2016.06.215.

Wasserstein RL, Schirm AL, Lazar NA. 2019. Moving to a World Beyond "p < 0.05." The American Statistician.

Wen H, Nie P, Liu M, Peng R, Guo T, Wang C, et al. 2023. Multi-health effects of clean residential heating: Evidences from rural China's coal-to-gas/electricity project. Energy for Sustainable Development 73:66–75; doi:10.1016/j.esd.2023.01.013.

Wooldridge JM. 2021. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-

in-Differences Estimators.; doi:10.2139/ssrn.3906345.

World Health Organization. 2021. WHO Global Air Quality Guidelines: Particulate Matter PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide.

Xu H, Brook RD, Wang T, Song X, Feng B, Yi T, et al. 2019. Short-term effects of ambient air pollution and outdoor temperature on biomarkers of myocardial damage, inflammation and oxidative stress in healthy adults. Environmental Epidemiology 3:e078; doi:10.1097/EE9.0000000000000078.

Xu W, Collet J-P, Shapiro S, Lin Y, Yang T, Wang C, et al. 2009. Validation and clinical interpretation of the St George's Respiratory Questionnaire among COPD patients, China. The International Journal of Tuberculosis and Lung Disease: The Official Journal of the International Union Against Tuberculosis and Lung Disease 13: 181–189.

Yan L, Carter E, Fu Y, Guo D, Huang P, Xie G, et al. 2020. Study protocol: The INTERMAP China Prospective (ICP) study. Wellcome Open Research 4:154; doi:10.12688/wellcomeopenres.15470.2.

Yang K. 2021. (Power industry provides inexhaustible power for national rejuvenation). (China Energy News Network).

Yap P-S, Garcia C. 2015. Effectiveness of Residential Wood-Burning Regulation on Decreasing Particulate Matter Levels and Hospitalizations in the San Joaquin Valley Air Basin. American Journal of Public Health 105:772–778; doi:10.2105/AJPH.2014.302360.

Ye W, Steenland K, Quinn A, Liao J, Balakrishnan K, Rosa G, et al. 2022. Effects of a Liquefied Petroleum Gas Stove Intervention on Gestational Blood Pressure: Intention-to-Treat and Exposure-Response Findings From the HAPIN Trial. Hypertension 79:1887–1898; doi:10.1161/HYPERTENSIONAHA.122.19362.

Young OR, Guttman D, Qi Y, Bachus K, Belis D, Cheng H, et al. 2015. Institutionalized governance processes: Comparing environmental problem solving in China and the United States. Global Environmental Change 31:163–173; doi:10.1016/j.gloenvcha.2015.01.010.

Yu C, Kang J, Teng J, Long H, Fu Y. 2021. Does coal-to-gas policy reduce air pollution? Evidence from a quasi-natural experiment in China. Science of The Total Environment 773:144645; doi:10.1016/j.scitotenv.2020.144645.

Yun X, Shen G, Shen H, Meng W, Chen Y, Xu H, et al. 2020. Residential solid fuel emissions contribute significantly to air pollution and associated health impacts in China. Science Advances 6:eaba7621; doi:10.1126/sciadv.aba7621.

Zhang J(Jim), Smith KR. 2007. Household Air Pollution from Coal and Biomass Fuels in China: Measurements, Health Impacts, and Interventions. Environmental Health Perspectives 115:848–855; doi:10.1289/ehp.9479.

Zhang Q, Zheng Y, Tong D, Shao M, Wang S, Zhang Y, et al. 2019. Drivers of improved $PM_{2.5}$ air quality in China from 2013 to 2017. Proceedings of the National Academy of Sciences 116:24463–24469; doi:10.1073/pnas.1907956116.

Zhang W, Xu H, Yu X, Li J, Zhang Y, Dai R, et al. 2023. Rigorous Regional Air Quality Standards for Substantial Health Benefits. Earth's Future 11:e2023EF003860; doi:10.1029/2023EF003860.

Zhuang Z, Li Y, Chen B, Guo J. 2009. Chinese kang as a domestic heating system in rural northern China—A review. Energy and Buildings 41:111–119; doi:10.1016/j.enbuild.2008.07.013.

Zigler CM, Kim C, Choirat C, Hansen JB, Wang Y, Hund L, et al. 2016. *Causal inference methods for estimating long-term health effects of air quality regulations. Research report 187.* Health Effects Institute / Health Effects Institute:Boston, MA.

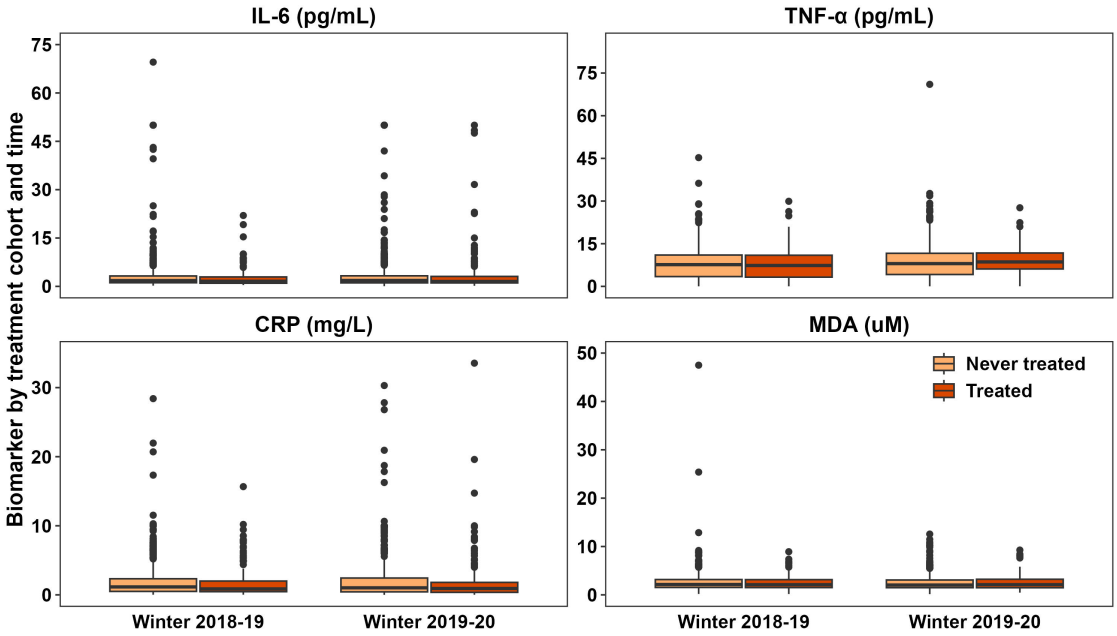
Zíková N, Wang Y, Yang F, Li X, Tian M, Hopke PK. 2016. On the source contribution to Beijing PM2.5 concentrations. Atmospheric Environment 134:84–95; doi:10.1016/j.atmosenv.2016.03.047.

A Appendices

A.1 Biomarker descriptives

Below we show boxplots for the blood inflammatory and oxidative stress markers.

Figure A1: Boxplots for markers of systemic inflammation including C-reactive protein (CRP), interleukin-6 (IL-6), tumour necrosis factor alpha (TNF-α) and malondialdehyde (MDA)



A.2 Missing data

A.2.1 Missingness across imputed variables

The reasons for missing data differ by variable:

- 'Valid response' indicates that either the variable is populated or that 'missing' is an appropriate response. For example, participants without diagnosed hypertension were not eligible to answer the survey question on whether they were taking BP-lowering medication. Participants without physician-diagnosed hypertension should be missing this variable (i.e., missing is a valid response).
- 'Non-response' indicates a missing value that should have been populated but was not. For example, we measured weight, height, and waist circumference at the clinic visit rather than during household visits in W1 and W2. The clinic visits took place on a different day than household visits and thus some participants who completed the home visits were away from the village on the day of the clinic visit. Thus, participants with missing "waist circumference", "height" or "weight" are considered 'non-responses'. As a second example, if an indoor $PM_{2.5}$ monitor was placed in the household but did not collect any data, this counts as a 'non-response' since the measurement was attempted but not completed.
- 'Not sampled' indicates that a value is missing because our protocol was to measure some variables in a sub-sample of participants and the participant was never sampled. For example, indoor $PM_{2.5}$ was measured in a 30% randomly selected sub-sample of households and thus the 70% of households in each village were 'not sampled' for an indoor $PM_{2.5}$ measurement.

Table A1: Missing and valid values of variables used in multiple imputation.

Variable	Valid (%)	Non-response (%)	Not sampled (%)
Study wave	3082 (100)	0 (0)	0 (0)
District	3082 (100)	0 (0)	0 (0)
Village	3082 (100)	0 (0)	0 (0)
Household ID	3082 (100)	0 (0)	0 (0)
Participant ID	3082 (100)	0 (0)	0 (0)
Village-level policy treatment status	3082 (100)	0 (0)	0 (0)
Systolic central BP	3081 (100)	1 (0)	0 (0)
Diastolic central BP	3081 (100)	1 (0)	0 (0)
Systolic brachial BP	3082 (100)	0 (0)	0 (0)
Diastolic brachial BP	3082 (100)	0 (0)	0 (0)
Participant sex	3081 (100)	1 (0)	0 (0)
Participant age	3078 (99.9)	4 (0.1)	0 (0)
Exposure to tobacco smoke	3081 (100)	1 (0)	0 (0)
Frequency of alcohol consumption in the past 12 months	3081 (100)	1 (0)	0 (0)
Physician diagnosis of high BP	3081 (100)	1 (0)	0 (0)
Whether participant took medication for high BP	1527 (49.5)	39 (1.3)	1516 (49.2)
Waist circumference	2568 (83.3)	514 (16.7)	0 (0)
Weight	2614 (84.8)	468 (15.2)	0 (0)
Height	2610 (84.7)	472 (15.3)	0 (0)
Self-reported diagnosis of diabetes	3082 (100)	0 (0)	0 (0)
Self-reported diagnosis of chronic kidney disease	3082 (100)	0 (0)	0 (0)
Household wealth index	2945 (95.6)	137 (4.4)	0 (0)
Whether BP was measured in AM or PM	3082 (100)	0 (0)	0 (0)
Right arm circumference	3073 (99.7)	9 (0.3)	0 (0)
Frequency of farming activities in past 6 months	3081 (100)	1 (0)	0 (0)
Frequency of exercise in the past 6 months	3081 (100)	1 (0)	0 (0)
If participant snores while sleeping	3081 (100)	1 (0)	0 (0)
If participant quits breathing while sleeping	3081 (100)	1 (0)	0 (0)
Marital status	3054 (99.1)	28 (0.9)	0 (0)
Self-reported diagnosis for coronary heart disease or myocardial infarction	3081 (100)	1 (0)	0 (0)
Self-reported diagnosis for stroke or transient ischemic attack (TIA)	3081 (100)	1 (0)	0 (0)
Heating season (Jan 15 to Mar 15) mean indoor PM2.5	494 (16)	122 (4)	2466 (80)
Indoor temperature	3074 (99.7)	8 (0.3)	0 (0)
Self-reported health status	3081 (100)	1 (0)	0 (0)
Blood pressure cuff size	3078 (99.9)	4 (0.1)	0 (0)
Participant's highest level of education	3059 (99.3)	23 (0.7)	0 (0)
Participant's current occupation	3053 (99.1)	29 (0.9)	0 (0)
Time-varying binary indicator for enrollment in policy	3082 (100)	0 (0)	0 (0)
Year of first treatment with policy	3082 (100)	0 (0)	0 (0)

A.2.2 Missing data by enrollment cohort and outcome

The number (N) and percent (Pct.) of missing observations are displayed in the table below by the enrollment cohort and outcome. 'Missing' indicates that the variable should have been populated for an observation but was not. 'Valid' indicates that the variable is populated. 'Not sampled' indicates that a value is missing because our protocol was to measure some variables in a subsample of participants and the participant was never sampled for the variable.

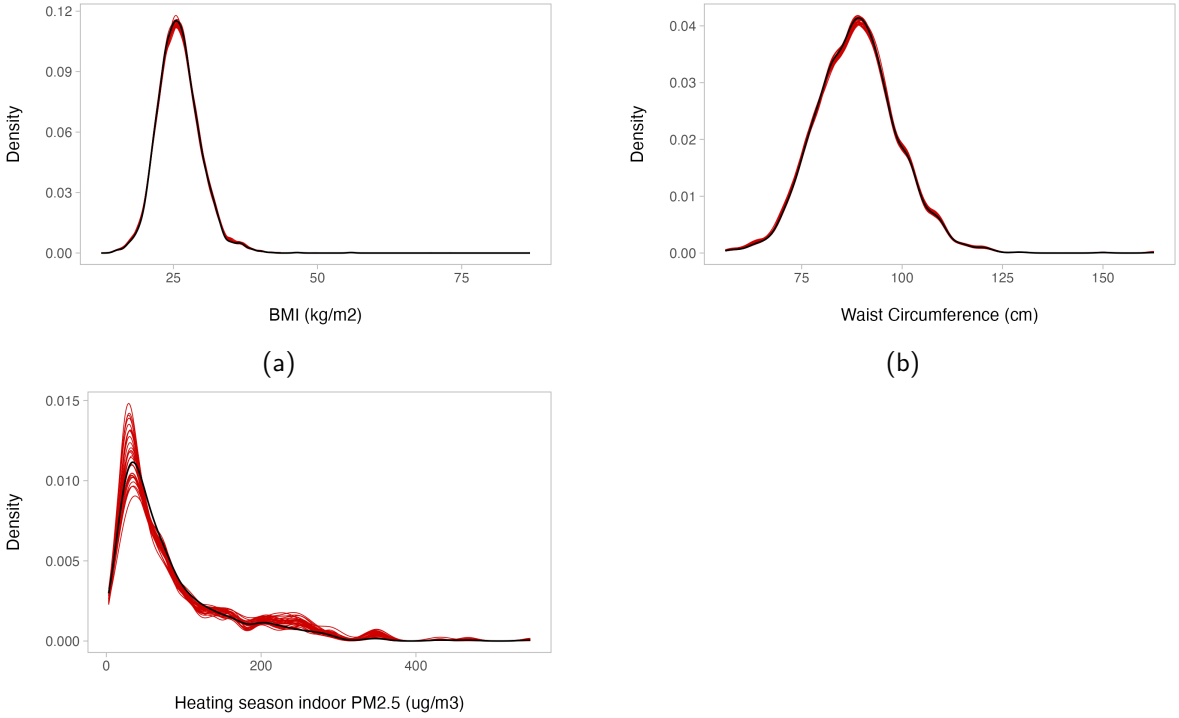
Table A2: Missing values by enrollment cohort and outcome

		Never enrolled (N=1880)		Enrolled 2019 (N=642)		Enrolled 2020 (N=446)		Enrolled 2021 (N=173)	
		N	Pct.	N	Pct.	N	Pct.	N	Pct.
Respiratory symptoms:									
Any symptoms	Missing	23	1.2	4	0.6	5	1.1	1	0.6
	Valid response	1857	98.8	638	99.4	441	98.9	172	99.4
Cough	Missing	23	1.2	4	0.6	5	1.1	1	0.6
	Valid response	1857	98.8	638	99.4	441	98.9	172	99.4
Phlegm	Missing	23	1.2	4	0.6	5	1.1	1	0.6
	Valid response	1857	98.8	638	99.4	441	98.9	172	99.4
Wheezing	Missing	23	1.2	4	0.6	5	1.1	1	0.6
	Valid response	1857	98.8	638	99.4	441	98.9	172	99.4
Shortness of breath	Missing	24	1.3	4	0.6	5	1.1	2	1.2
	Valid response	1856	98.7	638	99.4	441	98.9	171	98.8
Chest trouble	Missing	24	1.3	4	0.6	5	1.1	1	0.6
	Valid response	1856	98.7	638	99.4	441	98.9	172	99.4
Blood pressure:									
Brachial SBP	Missing	30	1.6	11	1.7	17	3.8	1	0.6
	Valid response	1850	98.4	631	98.3	429	96.2	172	99.4
Central SBP	Missing	30	1.6	12	1.9	17	3.8	1	0.6
	Valid response	1850	98.4	630	98.1	429	96.2	172	99.4
Brachial DBP	Missing	30	1.6	11	1.7	17	3.8	1	0.6
	Valid response	1850	98.4	631	98.3	429	96.2	172	99.4
Central DBP	Missing	30	1.6	12	1.9	17	3.8	1	0.6
	Valid response	1850	98.4	630	98.1	429	96.2	172	99.4
Biomarkers:									
IL6	Missing	8	0.6	4	0.9	1	0.3	0	0
TNF	Missing	8	0.6	4	0.9	1	0.3	0	0
CRP	Missing	8	0.6	4	0.9	1	0.3	0	0
MDA	Missing	11	0.9	4	0.9	1	0.3	0	0
FeNO	Missing	15	0.8	0	0.0	0	0.0	0	0.0
	Not sampled	1341	71.3	448	69.8	383	85.9	111	64.2
	Valid response	524	27.9	194	30.2	63	14.1	62	35.8
Environmental outcomes:									
Personal PM	Missing	13	0.7	4	0.6	6	1.3	0	0.0
	Not sampled	984	52.3	343	53.4	239	53.6	84	48.6
	Valid response	883	47.0	295	46.0	201	45.1	89	51.4
Indoor PM	Missing	60	3.2	11	1.7	16	3.6	4	2.3
	Not sampled	1502	79.9	518	80.7	360	80.7	135	78.0
	Valid response	318	16.9	113	17.6	70	15.7	34	19.7
Indoor temperature	Missing	33	1.8	10	1.6	16	3.6	1	0.6
	Valid response	1847	98.2	632	98.4	430	96.4	172	99.4

A.2.3 Imputation results

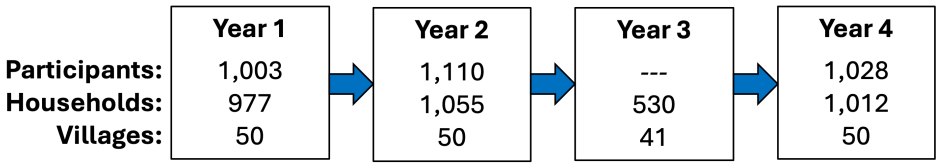
The figures below show density plots for the values of body mass index, waist circumference, and indoor $PM_{2.5}$ from the multiple imputation models. The red lines show the values for each of the 30 imputed datasets, and the black line shows the value for the observed data.

Figure A2: Kernal density plots showing distribution of multiply-imputed values for body mass index (kg/m^2), waist circumference (cm), and indoor $PM_{2.5}$ ($\mu g/m^3$) (red lines) and observed values (heavy black line).



A.3 Participant flow diagram

Figure A3: Flow chart of BHET study participation at the participant, household, and village levels across study years.



A.4 Sample sizes

Table A3: Sample sizes for health and environmental measurements for participants (P), households (HH), and villages (V).

	All waves			Wave 1			Wave 2			Wave 3			Wave 4		
Outcome	P	HH	V	P	HH	V	P	HH	V	P	HH	V	P	HH	V
Total	1438	1236	50	1003	977	50	1110	1055	50	.	530	41	1028	1012	50
Health Measures															
BP	1423	.	.	975	.	.	1103	.	.	.	.	.	1004	.	.
Respiratory symptoms	1429	.	.	991	.	.	1107	.	.	.	.	.	1010	.	.
FeNO	511	.	.	268	.	.	323	.	.	.	.	.	252	.	.
Inflammatory biomarkers															
IL6	1064	.	.	732	.	.	874	.	.	.	.	.	.	.	.
TNF	1064	.	.	732	.	.	874	.	.	.	.	.	.	.	.
CRP	1064	.	.	732	.	.	874	.	.	.	.	.	.	.	.
MDA	1064	.	.	729	.	.	874	.	.	.	.	.	.	.	.
Personal air pollution															
Filter-derived PM2.5	761	.	.	489	.	.	485	.	.	.	.	.	494	.	.
Filter-derived BC	755	.	.	489	.	.	478	.	.	.	.	.	476	.	.
Indoor air pollution															
Sensor-derived PM2.5 [a]	.	330	.	.	.	.	.	268	.	.	246	.	.	245	.
Filter-derived PM2.5	.	177	.	.	.	.	.	147	.	.	.	.	.	148	.
Filter-derived BC	.	176	.	.	.	.	.	146	.	.	.	.	.	138	.
Outdoor air pollution															
Sensor-derived PM2.5 [a]	.	.	50	.	.	40	.	.	50	.	.	50	.	.	50
Filter-derived PM2.5	.	.	50	.	.	44	.	.	50	.	.	50	.	.	50
Filter-derived BC	.	.	50	.	.	44	.	.	50	.	.	50	.	.	50
Indoor temperature															
'Point' temperature [b]	.	1228	.	.	956	.	.	1050	.	.	.	.	.	1001	.
Long-term temperature	.	753	.	.	366	.	.	557	.	.	454	.	.	458	.

Note: P = Participants, HH = Households, V = Villages

[a] Sample size for seasonal measurements.

[b] Measured in the 5 minutes before blood pressure.

A.5 District-level statistics

Table A4: Descriptive characteristics by district.

	Fangshan (n=11)		**Huairo (n=18)**		**Miyun (n=12)**		**Mentougou (n=9)**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number of households	699.1	514.4	163.1	111.6	274.4	198.8	204.3	73.20
Per capita income (RMB, 1000s)	7.2	1.4	20.0	2.8	17.3	2.8	11.3	2.30
Distance to Beijing center (km)	67.2	2.3	88.6	8.9	83.0	3.7	45.0	5.70
Altitude (m)	146.0	36.9	353.9	121.3	283.1	85.9	312.4	133.90
Winter briquette quantity (tonnes)	3.7	1.2	4.0	1.6	2.8	1.2	2.4	1.00
Winter wood quantity (kilograms)	1180.0	1192.0	2411.0	3845.0	2309.0	1997.0	1493.0	3144.00
Participant age (years)	59.0	9.1	60.0	8.8	60.0	8.9	62.0	9.70
Education (0-None or primary school; 1-secondary +)	0.4	0.5	0.3	0.5	0.3	0.4	0.3	0.46
Participant weight (kg)	69.0	11.0	66.0	11.0	65.0	10.0	68.0	11.00
Participant height (cm)	161.0	8.6	161.0	8.2	159.0	8.4	161.0	8.60

Note: Number of villages given in parenthesis. SD = Standard deviation

A.6 Policy uptake

Table A5 shows results from applying our extended two-way fixed effects models (in separate analyses) to coal and biomass consumption.

Table A5: Policy impacts on self-reported fuel use (kg)

		Coal		Biomass	
Cohort	Time	ATT[a]	(95% CI)	ATT[b]	(95% CI)
Average ATT					
All	All	-2361	(-2677, -2044)	-487	(-805, -168)
Cohort-Time ATTs					
2019	2019	-2631	(-2913, -2348)	-653	(-991, -315)
2019	2021	-2416	(-2847, -1984)	-633	(-1201, -64)
2020	2021	-2018	(-2474, -1562)	-350	(-701, 0)
2021	2021	-1961	(-2895, -1027)	338	(-30, 705)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval

[a] Joint test that all ATTs are equal: F(3, 2886)= 1.856, p= 0.135

[b] Joint test that all ATTs are equal: F(3, 2886)= 5.545, p= 0.001

A.7 Heterogeneity in treatment effects

A.7.1 Personal exposure

Table A6 shows limited evidence that the ATTs across cohorts and time demonstrate meaningful heterogeneity.

Table A6: Heterogenous treatment effects: Personal exposures

		PM2.5		Black carbon	
Cohort	Time	ATT[a]	(95% CI)	ATT[b]	(95% CI)
Average ATT					
All	All	0.2	(-19.6, 19.9)	-0.4	(-1.5, 0.6)
Cohort-Time ATTs					
2019	2019	7.4	(-24.4, 39.2)	-0.9	(-1.9, 0.2)
2019	2021	-13.0	(-36.6, 10.6)	-0.4	(-2.0, 1.2)
2020	2021	15.7	(-34.8, 66.3)	-0.1	(-1.8, 1.5)
2021	2021	-13.7	(-36.6, 9.3)	0.1	(-1.4, 1.6)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval

[a] Joint test that all cohort-time ATTs are equal: F(3, 1260)= 0.501, p= 0.682

[b] Joint test that all cohort-time ATTs are equal: F(3, 1151)= 0.965, p= 0.408

A.7.2 Indoor $PM_{2.5}$

Table Table A7 shows estimates for cohort-time ATTs for 24-hr and seasonal indoor $PM_{2.5}$.

Table A7: Heterogenous treatment effects for Indoor $PM_{2.5}$.

		24-hr		Seasonal	
Cohort	Time	ATT[a]	(95% CI)	ATT[b]	(95% CI)
Average ATT					
All	All	-20.0	(-45.6, 5.5)	-20.3	(-37.5, -3.0)
Cohort-Time ATTs					
2020	2021	-13.3	(-45.9, 19.3)	-17.6	(-39.6, 4.4)
2021	2021	-35.9	(-58.9, -13.0)	-25.9	(-42.8, -9.0)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval

[a] Joint test that all ATTs are equal: F(1, 393)= 0.057, p= 0.811

[b] Joint test that all ATTs are equal: F(1, 360)= 0.675, p= 0.412

A.7.3 Indoor temperature

Table A8: Heterogeneous treatment effects for indoor temperature

	Point temp (°C)				Mean temp (°C)				Min temp (°C)			
Cohort Time	ATT	(95%CI)	p[a]	Obs	ATT	(95%CI)	p[a]	Obs	ATT	(95% CI)	p[a]	Obs
All times												
2019 2019	1.66	(0.5, 2.8)			0.33	(-0.8, 1.5)			1.96	(0.5, 3.4)		
2019 2021	2.17	(0.5, 3.9)			0.80	(0.0, 1.6)			5.04	(2.3, 7.8)		
2020 2021	2.39	(0.7, 4.1)			0.66	(-0.5, 1.8)			7.27	(4.6, 9.9)		
2021 2021	0.60	(-1.2, 2.4)	0.37	2999	1.60	(0.5, 2.7)	0.45	1350	2.37	(0.2, 4.5)	0	1350
Daytime												
2019 2019					0.36	(-0.8, 1.5)						
2019 2021					0.91	(0.1, 1.7)						
2020 2021					0.95	(-0.2, 2.1)						
2021 2021					1.67	(0.5, 2.8)	0.48	1346				
Daytime heating												
2019 2019					0.92	(-0.2, 2.0)			1.94	(0.5, 3.4)		
2019 2021					2.02	(0.9, 3.2)			5.46	(2.7, 8.2)		
2020 2021					2.63	(1.5, 3.7)			6.69	(4.2, 9.2)		
2021 2021					2.72	(0.8, 4.6)	0	1350	2.53	(0.4, 4.7)	0	1350
Heating season												
2019 2019					0.95	(-0.2, 2.1)						
2019 2021					2.18	(0.9, 3.4)						
2020 2021					2.97	(1.9, 4.0)						
2021 2021					2.80	(0.9, 4.7)	0	1346				

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval

[a] P-value for omnibus test of heterogeneity across cohort time groups.

A.7.4 Blood pressure outcomes

Table A9 shows ATTs by treatment cohort and time, as well as the results of joint tests of heterogeneity across ATTs.

Table A9: Heterogeneous treatment effects for the total effect of the CHP on blood pressure.

		Adjusted DiD		Heterogeneity tests	
Cohort	Time	ATT[a]	(95% CI)	F-Statistic[b]	p-value
Brachial SBP					
2019	2019	-2.4	(-5.2, 0.5)		
2019	2021	-1.5	(-4.0, 1.0)		
2020	2021	-1.3	(-5.0, 2.5)		
2021	2021	2.4	(-0.5, 5.3)	2.3	0.080
Central SBP					
2019	2019	-2.0	(-4.7, 0.6)		
2019	2021	-2.0	(-4.5, 0.5)		
2020	2021	-1.8	(-5.1, 1.5)		
2021	2021	2.1	(-1.1, 5.3)	1.9	0.140
Brachial DBP					
2019	2019	-2.7	(-4.7, -0.7)		
2019	2021	-2.4	(-4.0, -0.7)		
2020	2021	0.2	(-1.5, 1.9)		
2021	2021	0.8	(-0.5, 2.0)	6.8	0.000
Central DBP					
2019	2019	-2.7	(-4.6, -0.8)		
2019	2021	-2.5	(-4.2, -0.9)		
2020	2021	0.1	(-1.7, 1.9)		
2021	2021	1.1	(-0.1, 2.2)	10.0	0.000

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval, DiD = Difference-in-Differences.

[a] Adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication.

[b] F-statistics and p-values for joint tests of equality across cohort and time ATTs

A.7.5 Mediation analyses for blood pressure

Table A10 shows the cohort-time treatment effects for the mediation model for blood pressure.

Table A10: Heterogeneous treatment effects for blood pressure mediation model.

		Adjusted Total Effect		CDE Mediated By:					
				Indoor PM		Indoor Temp		PM + Temp	
Cohort	Time	ATT[a]	(95% CI)	ATT[b]	(95% CI)	ATT[b]	(95% CI)	ATT[b]	(95% CI)
Brachial SBP									
2019	2019	-2.16	(-5.0, 0.7)	-1.67	(-4.7, 1.3)	-1.17	(-4.1, 1.8)	-0.63	(-3.7, 2.5)
2019	2021	-1.54	(-4.0, 1.0)	-0.86	(-3.6, 1.9)	-0.13	(-2.7, 2.5)	0.56	(-2.4, 3.5)
2020	2021	-1.45	(-5.2, 2.3)	-0.68	(-4.6, 3.2)	-0.01	(-3.6, 3.6)	0.83	(-3.2, 4.9)
2021	2021	2.28	(-0.5, 5.1)	2.48	(-0.8, 5.8)	1.55	(-2.1, 5.2)	1.76	(-2.2, 5.7)
Central SBP									
2019	2019	-1.81	(-4.4, 0.8)	-1.21	(-4.1, 1.7)	-0.83	(-3.6, 1.9)	-0.19	(-3.2, 2.8)
2019	2021	-1.82	(-4.3, 0.7)	-1.05	(-3.9, 1.8)	-0.47	(-2.8, 1.9)	0.31	(-2.5, 3.1)
2020	2021	-1.85	(-5.2, 1.5)	-1.12	(-4.7, 2.5)	-0.52	(-3.9, 2.8)	0.28	(-3.6, 4.2)
2021	2021	2.15	(-1.0, 5.3)	2.32	(-1.2, 5.9)	1.37	(-2.1, 4.8)	1.57	(-2.2, 5.3)
Brachial DBP									
2019	2019	-2.51	(-4.5, -0.6)	-2.03	(-4.2, 0.2)	-2.01	(-3.9, -0.1)	-1.47	(-3.6, 0.7)
2019	2021	-2.34	(-3.9, -0.7)	-1.71	(-3.8, 0.4)	-1.74	(-3.2, -0.3)	-1.08	(-3.0, 0.9)
2020	2021	0.08	(-1.7, 1.8)	0.23	(-1.6, 2.1)	0.92	(-1.0, 2.8)	1.12	(-1.0, 3.2)
2021	2021	0.73	(-0.5, 1.9)	1.08	(-0.7, 2.8)	0.11	(-1.3, 1.5)	0.47	(-1.4, 2.3)
Central DBP									
2019	2019	-2.54	(-4.4, -0.7)	-1.97	(-4.2, 0.2)	-2.26	(-4.1, -0.4)	-1.62	(-3.8, 0.5)
2019	2021	-2.45	(-4.0, -0.9)	-1.72	(-3.8, 0.4)	-2.05	(-3.5, -0.6)	-1.29	(-3.3, 0.7)
2020	2021	0.08	(-1.7, 1.9)	0.23	(-1.7, 2.2)	0.94	(-1.0, 2.9)	1.14	(-1.0, 3.3)
2021	2021	1.12	(0.0, 2.2)	1.51	(-0.1, 3.2)	0.48	(-0.8, 1.8)	0.88	(-0.9, 2.7)

Note: Results combined across 30 multiply-imputed datasets. ATT = Average Treatment Effect on the Treated, CDE = Controlled Direct Effect, DBP = Diastolic blood pressure, SBP = Systolic blood pressure. Median p-values for heterogeneity tests for multiple mediation models: bSBP = 0.78, cSBP = 0.85, bDBP = 0.11, cDBP = 0.04.

[a] Adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication.

[b] Mediators were set to the mean value for untreated participants at baseline.

A.7.6 Self-reported respiratory outcomes

Appendix tables A11, A12, A13, A14, A15, A16 below show Average Treatment Effect on the Treated (ATTs) by treatment cohort and time. ATTs are derived from estimating marginal effects from extended two-way fixed effects models with additional adjustment for age, sex, and smoking status.

Table A11: Heterogenous treatment effects for self-reported respiratory outcomes: Any respiratory symptom.

Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	-7.5	(-12.7, -2.3)
Cohort-Time ATTs			
2019	2019	-11.3	(-18.4, -4.2)
2019	2021	-9.3	(-16.7, -1.9)
2020	2021	0.9	(-10.8, 12.6)
2021	2021	-6.7	(-12.7, -0.7)

Note: Joint test that all ATTs are equal: F(3, 3050)= 0.998, p= 0.393.

Table A12: Heterogenous treatment effects for self-reported respiratory outcomes: Coughing.

Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	-2.7	(-7.1, 1.7)
Cohort-Time ATTs			
2019	2019	-4.9	(-10.5, 0.7)
2019	2021	-0.8	(-8.1, 6.5)
2020	2021	-2.2	(-8.8, 4.3)
2021	2021	-2.3	(-10.0, 5.4)

Note: Joint test that all ATTs are equal: F(3, 3050)= 0.482, p= 0.695.

Table A13: Heterogenous treatment effects for self-reported respiratory outcomes: Phlegm

Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	-1.6	(-5.6, 2.4)
Cohort-Time ATTs			
2019	2019	-5.3	(-13.1, 2.6)
2019	2021	-2.9	(-8.8, 3.0)
2020	2021	3.1	(-2.9, 9.1)
2021	2021	6.1	(0.8, 11.5)

Note: Joint test that all ATTs are equal: F(3, 3050)= 3.415, p= 0.017.

Table A14: Heterogenous treatment effects for self-reported respiratory outcomes: Wheezing attacks.

Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	1.0	(-1.9, 3.9)
Cohort-Time ATTs			
2019	2019	-0.9	(-3.5, 1.8)
2019	2021	2.3	(-1.7, 6.4)
2020	2021	-1.0	(-6.8, 4.7)
2021	2021	8.6	(-2.3, 19.5)

Note: Joint test that all ATTs are equal: F(3, 3050)= 2.474, p= 0.06.

Table A15: Heterogenous treatment effects for self-reported respiratory outcomes: Trouble breathing

Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	-3.4	(-9.2, 2.4)
Cohort-Time ATTs			
2019	2019	-5.0	(-12.6, 2.6)
2019	2021	-5.1	(-13.1, 2.9)
2020	2021	2.4	(-6.4, 11.1)
2021	2021	-5.1	(-18.7, 8.5)

Note: Joint test that all ATTs are equal: F(3, 3050)= 0.916, p= 0.432.

Table A16: Heterogenous treatment effects for self-reported respiratory outcomes: Chest trouble

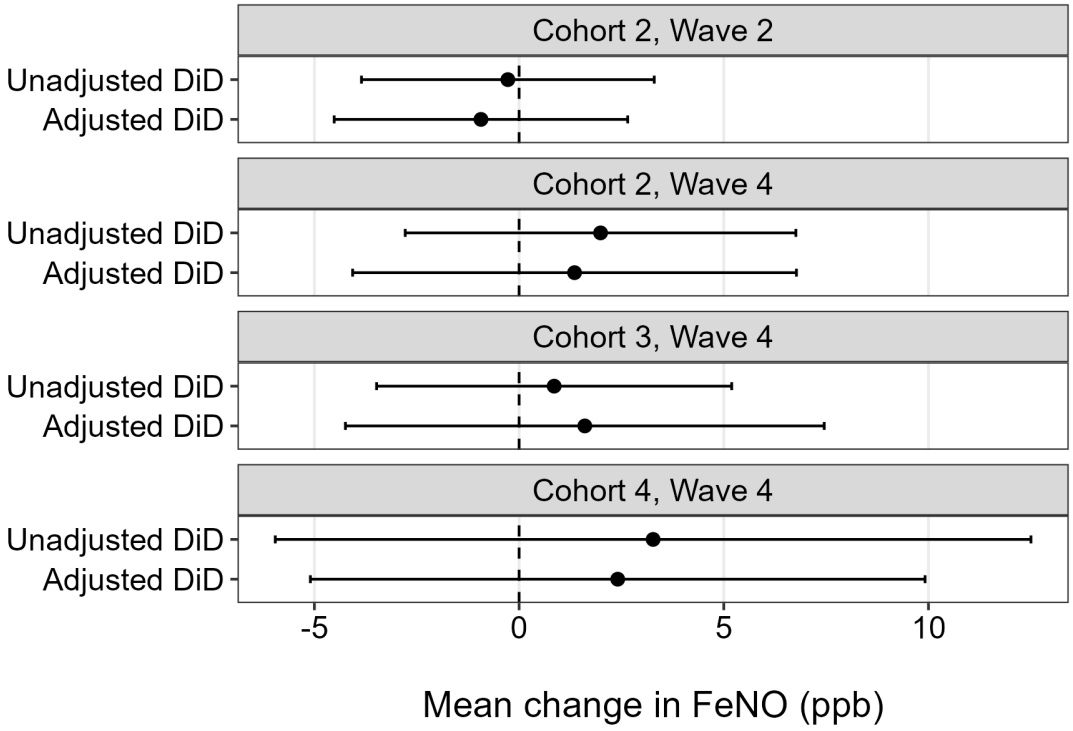
Cohort	Time	ATT	(95% CI)
Average ATT			
All	All	-3.4	(-8.1, 1.3)
Cohort-Time ATTs			
2019	2019	-2.7	(-8.9, 3.5)
2019	2021	-2.7	(-10.2, 4.7)
2020	2021	-2.0	(-8.3, 4.2)
2021	2021	-11.3	(-17.5, -5.1)

Note: Joint test that all ATTs are equal: F(3, 3050)= 3.176, p= 0.023.

A.7.7 FeNO

Figure A4 shows the cohort-time treatment effects for FeNO for both basic and covariate-adjusted DiD models.

Figure A4: Heterogenous treatment effects for FeNO.



A.7.8 Outdoor and personal mixed combustion

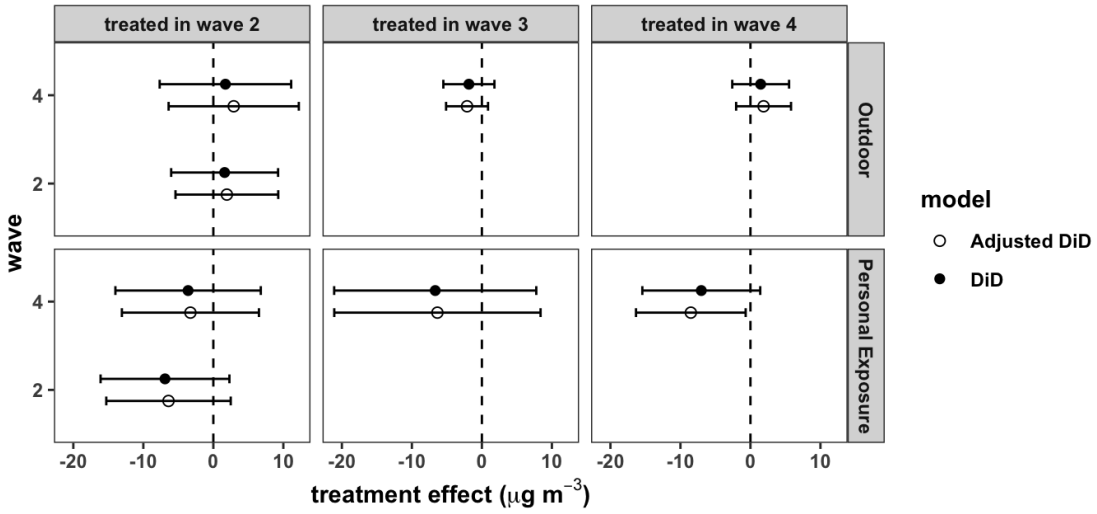


Figure A5: Adjusted and unadjusted treatment effect for outdoor and personal exposure (μg/m^3) to the mixed combustion source by treatment year.

A.8 Impact of adjustment for district on air pollution estimates

Table A17: Impact of adding district-level fixed effects to models for personal and indoor air pollution.

	Personal PM2.5		**Black carbon**		**24-hr indoor**		**Seasonal indoor**	
	(a)	(b)	(a)	(b)	(a)	(b)	(a)	(b)
ATT	0.2	0.0	-0.4	-0.3	-20.0	-21.8	-20.3	-22.0
	(10.1)	(9.4)	(0.5)	(0.5)	(12.4)	(13.3)	(7.8)	(8.6)
Observations	1,270	1,270	1,161	1,161	399	399	366	366
Year fixed effects	X	X	X	X	X	X	X	X
Cohort fixed effects	X	X	X	X	X	X	X	X
District fixed effects		X		X		X		X

Note: (a) excludes and (b) includes district fixed effects. All models adjusted for household size, smoking, outdoor temperature, and outdoor dewpoint. Standard errors (in parenthesis) clustered by village.

A.9 Impact of sample composition on brachial and central blood pressure results.

Table A18: Effects of excluding later-enrolled participants on estimates of the CHP on systolic and diastolic blood pressure.

		Adjusted DiD		
		Individuals	ATT[a]	(95% CI)
All participants				
Systolic BP (mmHg)	Brachial	1423	-1.4	(-3.3, 0.5)
	Central	1423	-1.6	(-3.4, 0.3)
Diastolic BP (mmHg)	Brachial	1423	-1.6	(-3.0, -0.3)
	Central	1423	-1.7	(-3.0, -0.3)
Excluding participants enrolled after W1				
Systolic BP (mmHg)	Brachial	992	-1.6	(-3.3, -0.0)
	Central	992	-1.6	(-3.1, -0.1)
Diastolic BP (mmHg)	Brachial	992	-1.7	(-2.9, -0.4)
	Central	992	-1.7	(-2.9, -0.5)

Note: ATT = Average Treatment Effect on the Treated, BP = blood pressure, CI = confidence interval, DiD = Difference-in-Differences, ETWFE = Extended Two-Way Fixed Effects.

[a] Marginal effect from ETWFE models adjusted for age, sex, waist circumference, smoking, alcohol consumption, and use of blood pressure medication. Results combined across 30 multiply-imputed datasets.

A.10 Impact of including Season 3 data

Table A19 shows differences in the ATTs for the impact of seasonal indoor $PM_{2.5}$ when wave 3 data (collected in 41 villages during COVID-19) are included versus excluded.

Table A19: Effects of the CHP policy on indoor seasonal $PM_{2.5}$ based on whether Wave 3 data are included vs. excluded.

		With Season 3 data			Without Season 3 data		
Cohort	Time	Obs	ATT	(95% CI)	Obs	ATT	(95% CI)
Average ATT (μg/m3)							
All	All	546	-33.5	(-55.0, -12.0)	389	-27.9	(-49.8, -6.1)
Cohort-Time ATTs (μg/m3)							
2020	2020	546	-36.1	(-60.8, -11.4)			
2020	2021	546	-28.8	(-59.1, 1.4)	389	-25.8	(-55.4, 3.8)
2021	2021	546	-36.3	(-50.5, -22.1)	389	-32.0	(-49.6, -14.4)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval.

Table A20 shows the impact of including Wave 3 data on the estimates of the impact of the policy on 24-hr and seasonal outdoor $PM_{2.5}$.

Table A20: Effects of the CHP policy on outdoor 24-hr and seasonal $PM_{2.5}$ based on whether Wave 3 data are included vs. excluded.

	With Season 3 data			Without Season 3 data		
	Obs	ATT	(95% CI)	Obs	ATT	(95% CI)
24-hr PM2.5	11174	-1.5	(-6.5, 3.6)	11174	-2.1	(-10.0, 5.8)
Seasonal PM2.5	139	0.8	(-4.5, 6.0)	139	0.5	(-4.8, 5.9)

Note: ATT = Average Treatment Effect on the Treated, CI = confidence interval.

A.11 Impact of sample composition on FeNO results

Table A21 shows differences in the ATTs for the impact of the CBHP policy on FeNO depending on whether the estimation sample includes all individuals or is limited to those with repeated measures across campaigns.

Table A21: Effects of the CHP on FeNO (ppb) based on the number of individuals with repeated measurements.

	All participants		Participants with >1 measure		Participants with 3 measures	
	ATT	(95% CI)	ATT	(95% CI)	ATT	(95% CI)
DiD	0.9	(-1.6, 3.3)	-0.4	(-2.9, 2.0)	0.3	(-2.8, 3.3)
Adjusted DiD	0.3	(-2.2, 2.8)	-0.6	(-3.2, 2.0)	0.3	(-2.9, 3.4)
Observations	793		541		272	

Note: ATT = Average Treatment Effect on the Treated, CI = Confidence Interval, DiD = Difference-in-Differences

A.12 Alternative PMF analyses

A.12.1 Disaggregated analyses

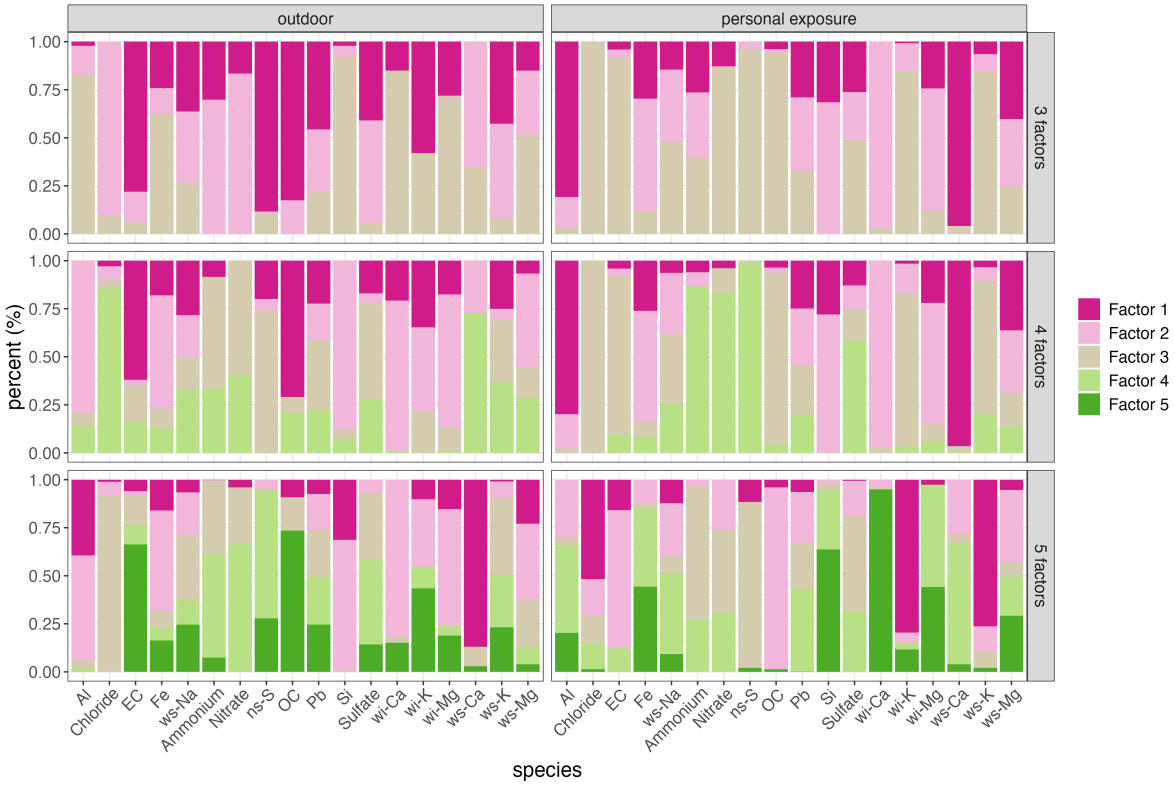
Figure A6 shows results for the 3-, 4-, and 5-factor solutions for both personal and outdoor exposure samples.

For the 3-factor solution, Factor 1, represented in dark pink, is interpreted as transported dust. This conclusion is supported by moderate loadings of mineral dust-related species such as silicon (Si) and water-soluble calcium (Ca). Calcium and aluminum are typically associated with crustal materials, especially in arid or semi-arid regions like the Gobi Desert and the Taklamakan Desert, which are known sources of transported dust to northern China, including Beijing. Factor 2, represented in light pink, likely represents local dust, with significant contributions from metals typically associated with crustal materials. Factor 3, represented in beige, is interpreted as a mixture of combustion-related sources, with contributions from elemental and organic carbon, as well as evidence of secondary particulate matter formation. This includes carbon species indicative of combustion, contributions from Pb and water-soluble potassium (K), and secondary formation markers such as sulfate (SO_4^{2-}).

In the 4-factor solution, Factor 1 (dark pink) remains identified as transported dust. Factor 2 (light pink) continues to represent local dust. Factor 3 (beige) now represents mixed combustion, capturing contributions from both biomass and coal-related sources. A new Factor 4 (light green) is interpreted as secondary sulfur, given its strong association with sulfate and ammonium (NH_4^+), both markers for secondary particulate matter formation.

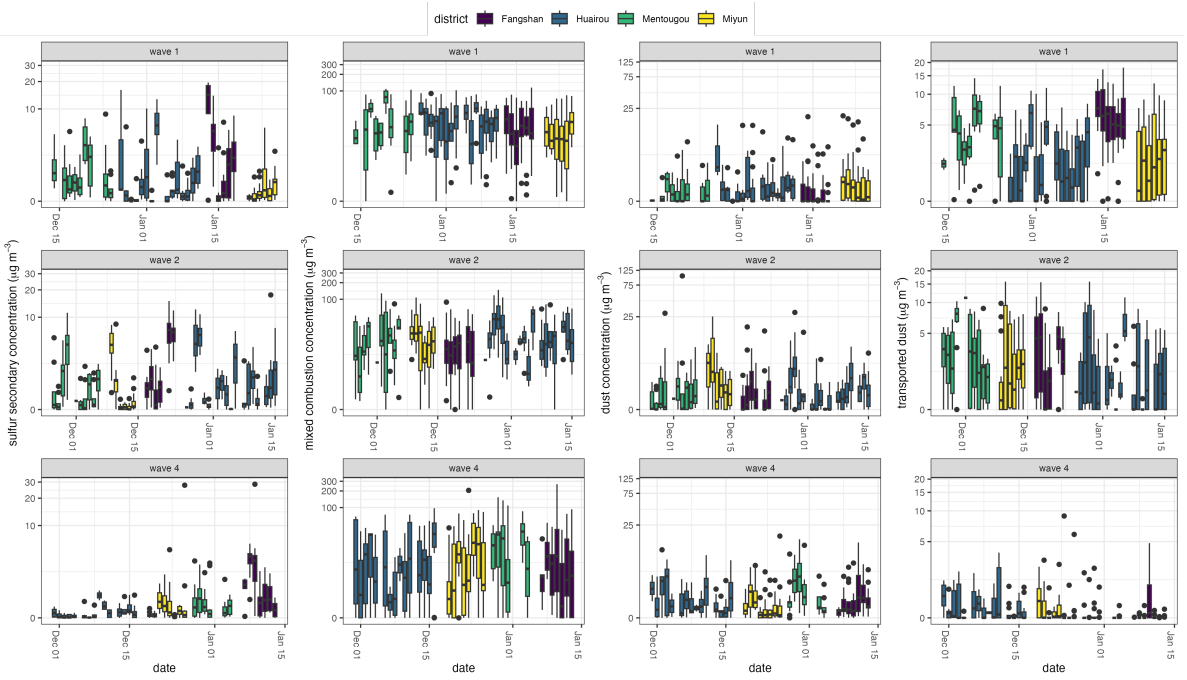
In the 5-factor solution, Factor 1 (dark pink) continues to represent transported dust. Factor 2 (light pink) shifts its primary attribution to secondary sulfur. Factor 3 (beige) is interpreted as mixed combustion, now dominated by biomass burning contributions. Factor 4 (light green) likely represents coal combustion, based on species such as arsenic (As) and selenium (Se), which are typically associated with coal burning. Finally, Factor 5 (dark green) represents local dust, primarily associated with crustal elements.

Figure A6: 3-, 4-, and 5- factor solutions separately for outdoor and personal exposure samples.



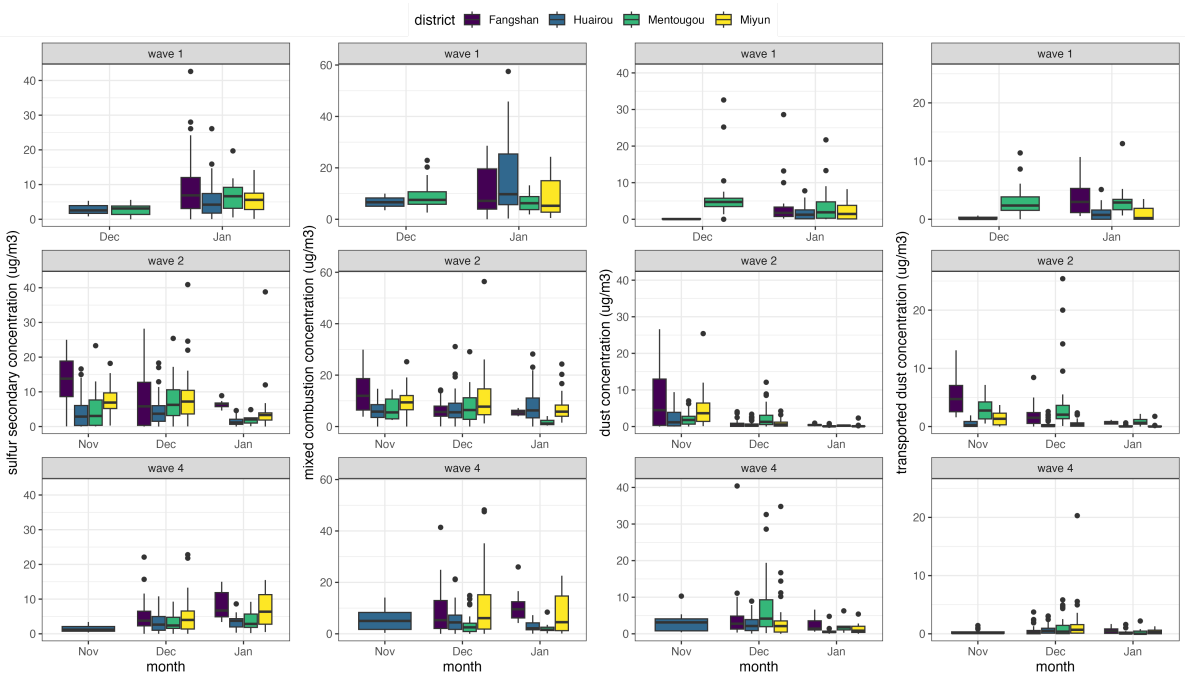
A.12.2 PMF results disaggregated by day

Figure A7: Mass concentrations (ug/m3) of contributions to PM2.5 mass by each of the four named sources identified in the source analysis. From left to right, the source contributions represented are 'sulfur secondary', 'mixed combustion', 'dust', and 'transported dust'. Source contribution mass concentrations are shown by day and color-coded by district, with purple for Fangshan, blue for Huairou, green for Mentougou, and yellow for Miyun.



A.12.3 PMF results disaggregated by month

Figure A8: Mass concentrations (ug/m3) of contributions to PM2.5 mass by each of the four named sources identified in the source analysis. From left to right, the source contributions represented are 'sulfur secondary', 'mixed combustion', 'dust', and 'transported dust'. Source contribution mass concentrations are shown by month (November, December, January) and color-coded by district, with purple for Fangshan, blue for Huairou, green for Mentougou, and yellow for Miyun.



A.12.4 Personal exposure sample diagnostics

Table A22 shows diagnostics for the PMF analysis for outdoor exposure samples.

Table A22: PMF error estimation diagnostics for personal exposure samples.

	Potential Factor Solution			
Diagnostic	3	4	5	6
Qexp	17301	16126	14951	13776
Qtrue	1225294	101524	84622	70122
Qrobust	117329	96215	80219	66865
Qr/Qexp	6.78	5.97	5.37	4.85
Q/Qexp >6	ns-S, NH3, NO3, SO4, ws-Na, Al, Cl, Pb	ns-S, ws-Na, Al, Cl, Pb	NO3, ws-Na, Al, Cl	NO3, ws-Na, Al, ws-K
DISP % dQ	<0.01%	<0.01%	<0.01%	<0.01%
DISP Swaps	0	0	0	0
Bootstrap mapping < 100%	mixed combustion - 95%	sulfur secondary - 85%	coal - 80%	chloride - 45%

A.12.5 Outdoor exposure sample diagnostics

Table A23 shows diagnostics for the PMF analysis for outdoor exposure samples.

Table A23: PMF error estimation diagnostics for outdoor samples.

	Potential Factor Solution			
Diagnostic	3	4	5	6
Qexp	10716	9980	9244	8508
Qtrue	43280	34603	28467	21947
Qrobust	41521	33209	26432	21347
Qr/Qexp	3.87	3.33	2.86	2.51
Q/Qexp >6	ws-Ca, Cl	Cl	Pb	None
DISP % dQ	<0.01%	<0.01%	<0.01%	<0.01%
DISP Swaps	0	0	0	0
Bootstrap mapping < 100%	None	sulfur secondary - 55%	None	coal - 90%, EC + OC - 90%, chloride - 90%

A.12.6 $PM_{2.5}$ constituents for outdoor samples

Table A24: Mean (95% confidence interval) species concentrations (x100 μg/m^3) for outdoor samples included in PMF.

	Wave 1		Wave 2		Wave 4	
Species	Mean	95% CI	Mean	95% CI	Mean	95% CI
Al	110	(108-113)	66.7	(65.5-67.8)	60.9	(59.8-62.1)
Fe	85.2	(82.9-87.5)	57.9	(57-58.9)	64.1	(63-65.2)
Pb	6.26	(5.82-6.71)	4.32	(4.06-4.57)	7.69	(7.27-8.11)
Si	139	(136-142)	76.7	(75.4-77.9)	128	(126-129)
Chloride	56.1	(54.5-57.8)	32.2	(31.5-32.9)	15.6	(14.9-16.4)
EC	123	(121-125)	112	(111-113)	103	(101-104)
ws-Na	11.4	(10.8-12)	9.28	(8.88-9.68)	9.23	(8.76-9.7)
Ammonium	136	(133-138)	152	(151-154)	78.3	(77.3-79.4)
Nitrate	225	(222-227)	259	(257-261)	146	(145-148)
ns-S	87	(85.3-88.7)	70	(69.2-70.9)	75.9	(74.9-76.9)
OC	1080	(1070-1080)	901	(898-904)	649	(645-653)
Sulfate	223	(220-225)	206	(205-208)	132	(131-134)
wi-Ca	79.4	(77.2-81.7)	40.7	(39.8-41.6)	79	(77.6-80.4)
wi-K	81.4	(79.5-83.3)	33.7	(33-34.3)	37.8	(37-38.6)
wi-Mg	39.8	(38.6-41)	21.2	(20.6-21.7)	38.6	(37.7-39.4)
ws-Ca	33.5	(32.5-34.5)	23.5	(22.8-24.2)	15.1	(14.3-15.9)
ws-K	52.4	(50.8-54)	25.4	(24.9-25.9)	17.9	(17.4-18.4)
ws-Mg	6.98	(6.41-7.55)	3.19	(2.97-3.41)	2.96	(2.72-3.2)

Note: CI = Confidence interval.

A.12.7 $PM_{2.5}$ constituents for personal samples

Table A25: Mean (95% confidence interval) species concentrations (x100 μg/m^3) for personal samples included in PMF.

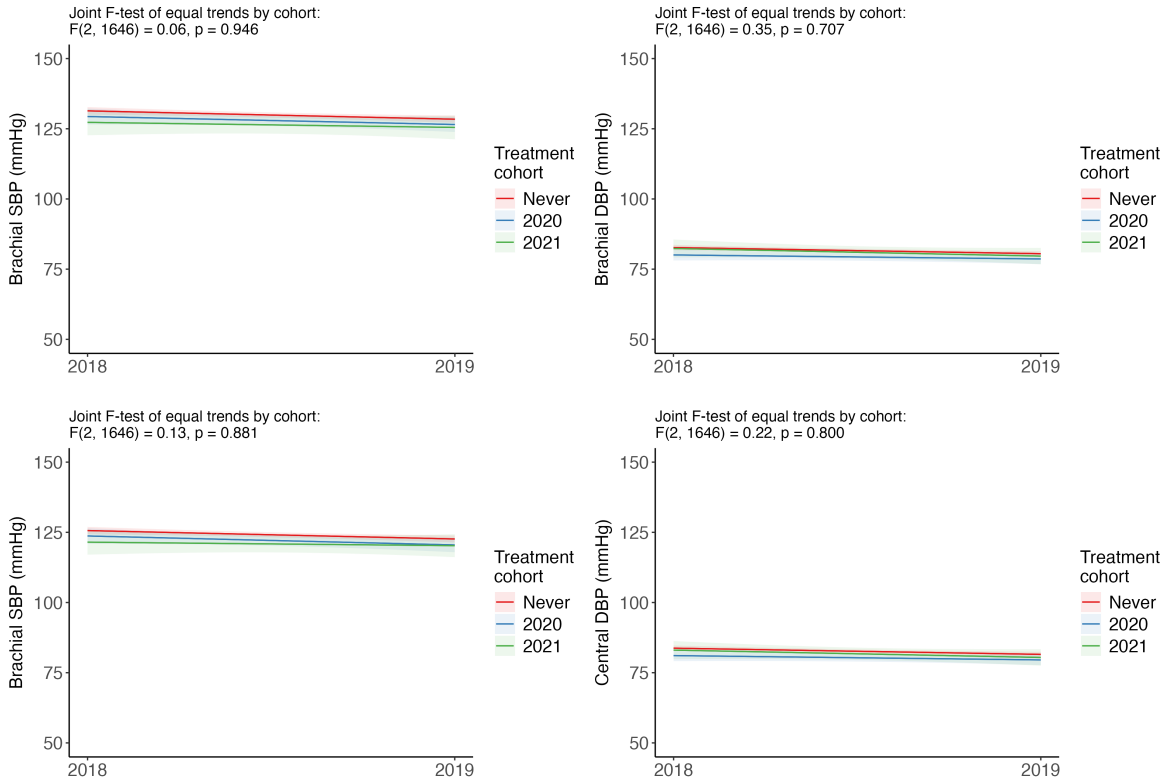
	Wave 1		Wave 2		Wave 4	
Species	Mean	95% CI	Mean	95% CI	Mean	95% CI
Al	203	(202-204)	213	(212-214)	27.4	(26.7-28.1)
Fe	40	(39.4-40.6)	57.1	(55.9-58.4)	33.6	(32.9-34.3)
Pb	15.8	(15.4-16.1)	13.7	(13.5-14)	12.8	(12.4-13.1)
Si	77.2	(75.9-78.4)	93.5	(92-95)	52.1	(51.2-53)
Chloride	70.2	(69-71.5)	43.8	(42.7-45)	34.6	(33.6-35.5)
EC	241	(239-243)	219	(218-221)	194	(192-196)
ws-Na	8.75	(8.36-9.14)	12.3	(11.8-12.7)	10.8	(10.3-11.2)
Ammonium	60.5	(59.5-61.5)	87.8	(86.4-89.1)	31.1	(30.2-31.9)
Nitrate	151	(149-152)	235	(232-237)	59.1	(58.1-60.1)
ns-S	46.3	(45.4-47.2)	44.1	(43.2-45)	27.5	(26.7-28.3)
OC	3740	(3730-3740)	3430	(3420-3440)	2680	(2680-2690)
Sulfate	129	(127-130)	126	(125-128)	77.6	(76.5-78.7)
wi-Ca	46.8	(45.7-47.9)	81.6	(79.8-83.5)	48.4	(47.6-49.2)
wi-K	88.2	(87-89.4)	72.8	(71.6-74)	67.7	(66.5-68.9)
wi-Mg	25.6	(25-26.2)	30.4	(29.7-31.1)	26.5	(26-27)
ws-Ca	57.1	(56.4-57.9)	42.8	(41.9-43.8)	15.7	(15.1-16.2)
ws-K	49	(48-50)	43.8	(42.9-44.7)	28.7	(27.9-29.5)
ws-Mg	4	(3.75-4.24)	3.77	(3.49-4.04)	2.06	(1.81-2.3)

Note: CI = Confidence interval.

A.13 Pre-trends

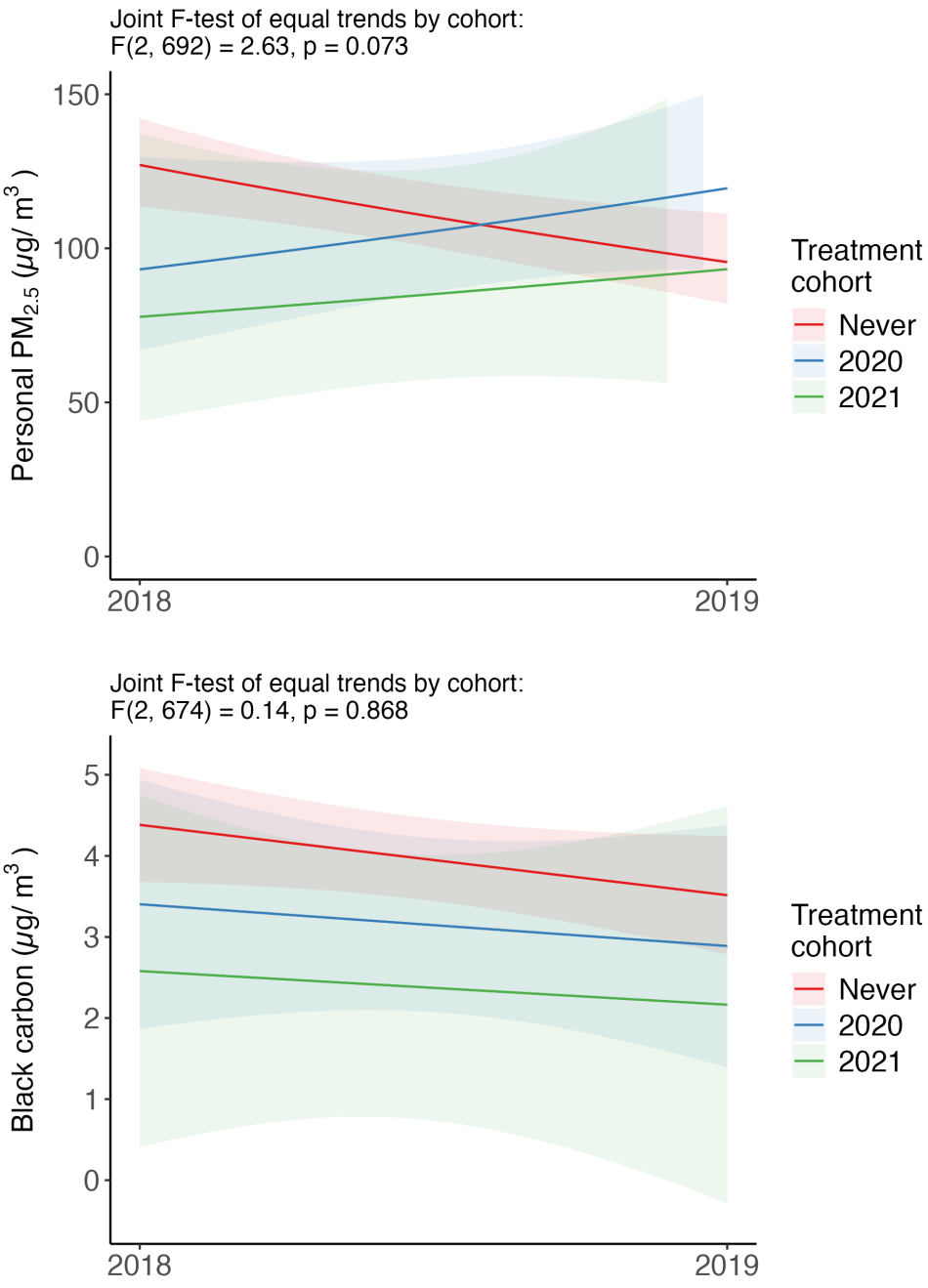
A.13.1 Blood pressure

Figure A9: Comparison of pre-interventions trends in blood pressure between waves 1 and 2 for never treated and villages treated later.



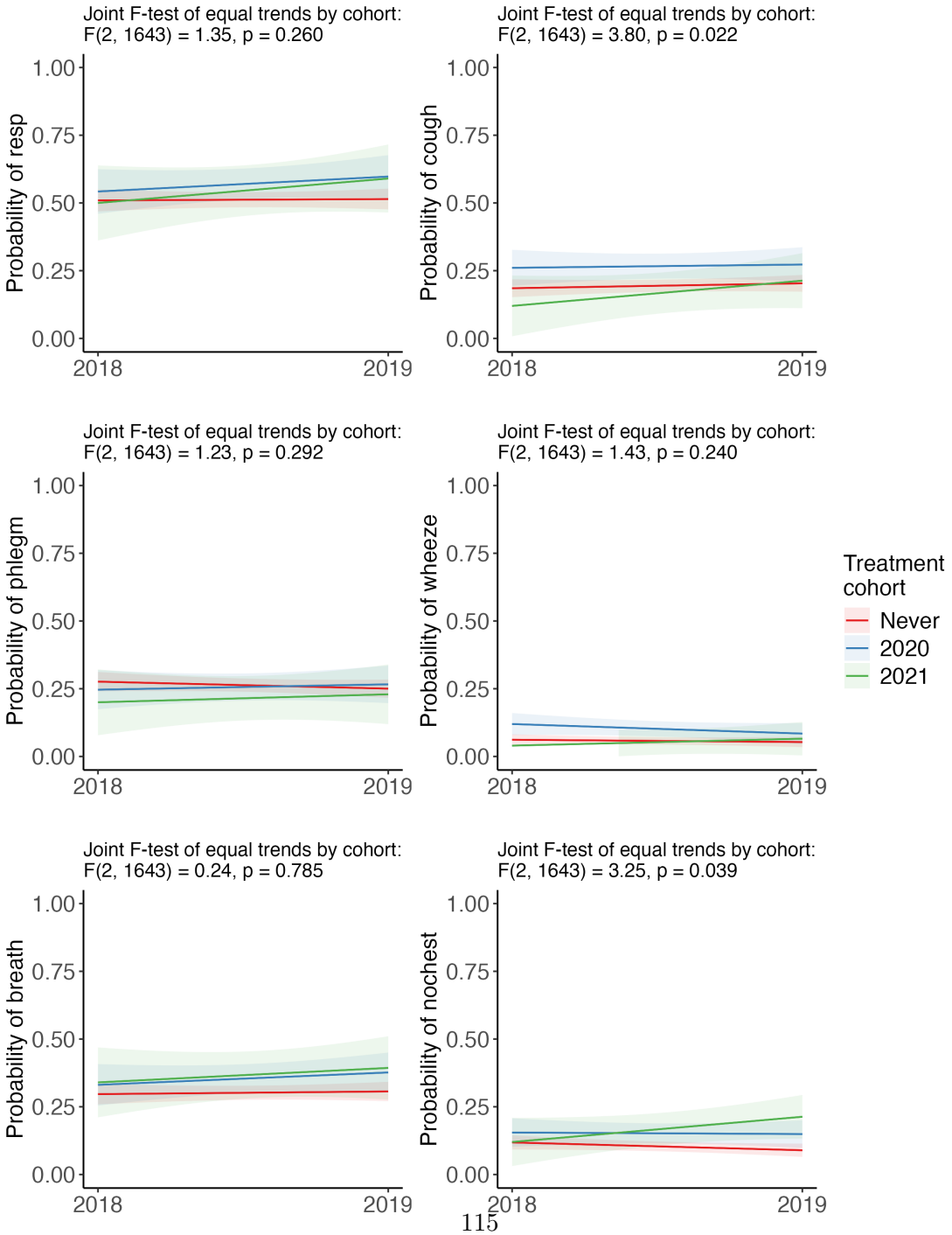
A.13.2 Personal exposure and black carbon

Figure A10: Comparison of pre-intervention trends in personal exposure and black carbon between waves 1 and 2 for never treated and villages treated later.



A.13.3 Self-reported respiratory outcomes

Figure A11: Comparison of pre-intervention trends in self-reported respiratory outcomes between waves 1 and 2 for never treated and villages treated later.



A.14 Impact of group and time fixed effects

Table A26 shows the impact of adding different sets of cohort and time fixed effects to the model for personal exposure.

Table A26: Effects of the CHP on personal exposure ($\mu g/m^3$) with variations in fixed effects for treatment group and time.

	Adjusted DiD	**No time FE**	**No group FE**	**No FE**
ATT	0.2	-27.5	-4.9	-22.0
	[-19.6, 19.9]	[-49.4, -5.5]	[-24.0, 14.3]	[-40.1, -3.8]
Observations	1,270	1,270	1,270	1,270
Year fixed effects	X		X	
Cohort fixed effects	X	X		

Note: DiD = Difference-in-Differences. FE = Fixed effects. All models adjusted for household size, smoking, outdoor temperature, and outdoor humidity. Standard errors clustered by village and 95% confidence intervals shown in brackets.

A.15 Retrospective design analysis

A.15.1 All outcomes

Table A27 shows the results of a design-based analysis (Gelman and Carlin 2014) for different hypothetical effect sizes, conditional on our design and sample size.

Table A27: Design-based analysis for various hypothetical effect sizes.

		Observed Results		Hypothetical Design Analysis			
		Estimate	SE	Effect	Power (%)	S-bias[a]	M-bias[b]
Blood pressure (mmHg)							
Systolic BP (mmHg)	Brachial	-1.4	1.0	-2.5	73.2	0.02	0.5
	Central	-1.6	0.9	-2.5	75.4	0.02	0.5
Diastolic BP (mmHg)	Brachial	-1.6	0.7	-2.0	80.0	0.00	1.1
	Central	-1.7	0.7	-2.0	82.7	0.00	1.1
Pulse Pressure	Brachial	0.2	0.6	0.5	12.9	0.23	1.0
	Central	0.1	0.6	0.5	14.5	0.22	1.0
BP Amplification x100	Pulse pressure	-0.0	0.6	0.1	5.3	0.44	4.5
	Systolic BP	0.1	0.2	0.1	10.0	0.28	1.2
Respiratory outcomes							
Self-reported (pp)	Any symptom	-7.5	2.7	-5.0	47.0	0.00	1.4
	Coughing	-2.7	2.2	-2.0	14.5	0.21	1.0
	Phlegm	-1.6	2.0	-3.0	31.2	0.10	0.7
	Wheezing attacks	1.0	1.5	-1.0	10.4	0.27	1.2
	Trouble breathing	-3.4	3.0	-3.0	17.4	0.18	0.9
	Chest trouble	-3.4	2.4	-1.0	7.0	0.36	1.8
Measured	FeNO (ppb)	0.3	1.3	-0.5	6.8	0.36	1.9
Inflammatory markers (%)							
Measured	IL6 (pg/mL)	0.8	0.6	-0.2	6.0	0.39	2.5
	TNF-alpha (pg/mL)	0.8	0.5	-0.4	15.0	0.21	0.9
	CRP (mg/L)	0.1	0.3	-0.1	6.5	0.37	2.1
	MDA (μM)	0.2	0.2	-0.2	13.2	0.23	1.0

Note: BP = Blood Pressure, pp = percentage points, ppb = parts per billion, SE = Standard Error

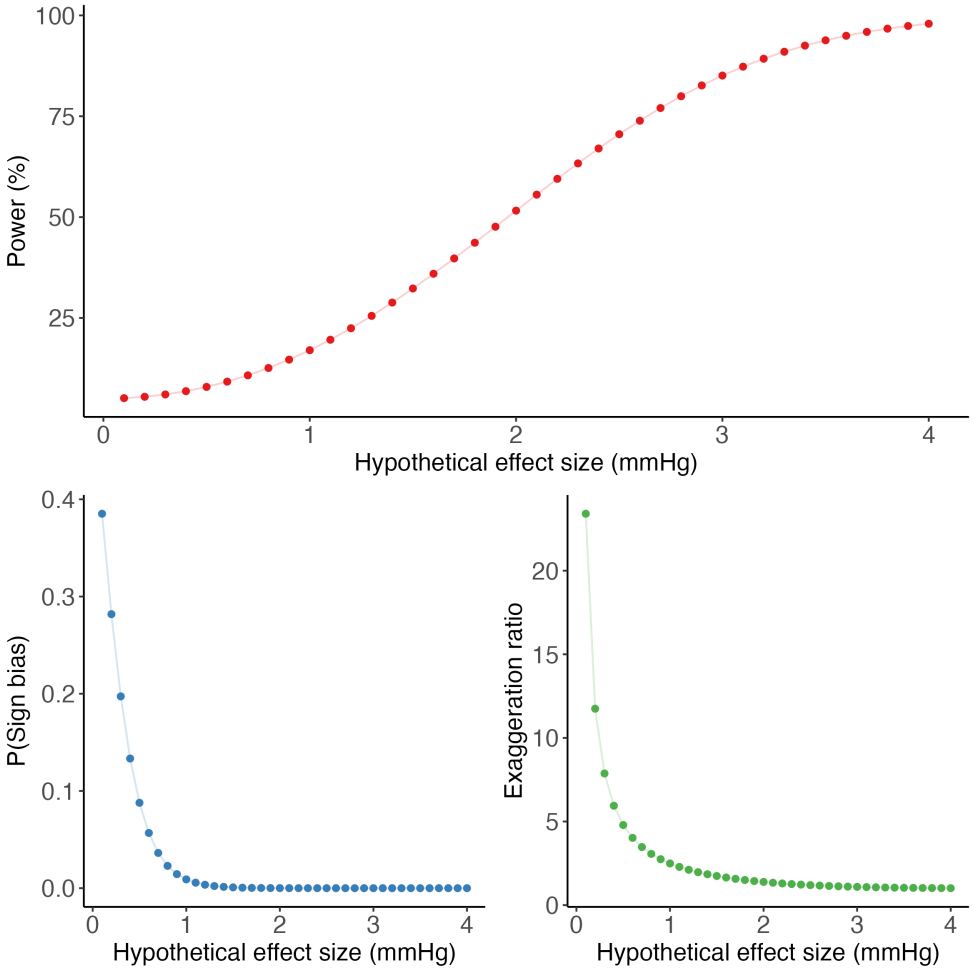
[a] Assuming the true effect size and our study design and precision, the probability that an observed estimate from a replication will have the wrong sign.

[b] Assuming the true effect size and our study design and precision, the ratio by which an observed estimate from a replication will exaggerate the true effect.

A.15.2 Blood pressure

Figure A12 show the range of estimates for power, sign-bias, and effect exaggeration for several hypothetical effect sizes for the impact of the program on blood pressure, conditional on our study design and sample size (standard error of 1 mmHg).

Figure A12: Power, sign-bias, and exaggeration ratios for various hypothetical effects of the policy on blood pressure, given our study design.



Abbreviations and other terms

ANMB	Absolute Normalized Mean Bias
ATT	Average Treatment Effect on the Treated
BAM	Beta Attenuation Monitor
BC	Black carbon
BP	Blood pressure
CI	Confidence Interval
CIE	International Commission on Illumination
CHP	Clean Heating Policy
cDBP	Central diastolic blood pressure
CRP	C-reactive protein
cSBP	Central systolic blood pressure
DAG	Directed acyclic graph
DiD	Difference-in-Differences
DISP	Displacement of Factor Elements
EC	Elemental carbon
EDXRF	Evo energy-dispersive X-ray fluorescence
ETWFE	Extended Two-Way Fixed Effects
FEM	Federal equivalent method
FID	Flame ionization detector
FeNO	Fractional exhaled nitric oxide
HAPIN	Household Air Pollution Intervention Network
HPLC	High-performance liquid chromatography
IL-6	Interleukin-6
MDA	Malondialdehyde
NISP	National Improved Stove Program
NIST	National Institute of Standards and Technology
ns-S	Non-Sulfate Sulfur
OC	Organic Carbon
OD	Optic densities
PKU	Peking University
$PM_{2.5}$	Particulate matter less than 2.5 microns in aerodynamic diameter
RMSE	Root mean square error
SRM	Standard reference material
TNF-α	Tumour necrosis factor alpha
UCAS	University of Chinese Academy of Sciences
UPAS	Ultrasonic Personal Aerosol Samplers
W1, W2, W3, W4	Wave 1, Wave 2, Wave 3, Wave 4
wi	Water Insoluble Species

About the authors

Jill Baumgartner, PhD is a professor jointly appointed in the Department of Equity, Ethics, and Policy and the Department of Epidemiology, Biostatistics & Occupational Health at McGill University. Her work evaluates the health impacts of air pollution, household energy use, and climate change..

Sam Harper, PhD is a professor in the Department of Epidemiology, Biostatistics & Occupational Health at McGill University. His research focuses on evaluating the impacts of social and economic policies on health.

Chris Barrington-Leigh, PhD is an associate professor jointly appointed in the Department of Equity, Ethics, and Policy and the Bieler School of Environment at McGill University. His research evaluates

Collin Brehmer is a PhD student in the Department of Civil and Environmental Engineering at Colorado State University.

Ellison M. Carter, PhD is an associate professor appointed in the Department of Civil and Environmental Engineering at Colorado State University. Her research combines interests and expertise in air quality, exposure science, and chemistry and aims to answer questions relevant to energy policy and their impacts on air pollution exposures and human health.

Xiaoying Li, PhD is a Research Scientist in the Department of Mechanical Engineering at Colorado State University and formerly a postdoctoral fellow at McGill University. Her research interests include investigating the interactions of indoor, outdoor air pollution, and personal exposures to air pollution and evaluating the impacts of clean energy interventions on air quality and human health.

Brian E. Robinson, PhD is an associate professor in the Department of Geography at McGill University who studies how people's livelihoods are influenced by ecosystem services and resource use, particularly in developing regions. His interdisciplinary research explores the interactions between livelihoods, the environment, and the institutions that govern resource management.

Guofeng Shen, PhD is an assistant professor of environmental science at Peking University. His research interests and experiences are in sustainable household energy and environment, largely focusing on fates, impacts and controls of hazardous pollutants produced from indoor solid fuel use that is an important indicator of sustainable development.

Talia J. Sternbach is a PhD student in the Department of Epidemiology, Biostatistics & Occupational Health at McGill University.

Shu Tao, PhD, is a professor of environmental science at Peking University who focuses on measuring and modeling clean energy transition, air pollution emissions, population exposures, and their estimated health impacts in China. His research integrates environmental science, atmospheric modeling, field measurements, and exposure assessment to better understand the sources, distribution, and health effects of air pollutants.

Kaibing Xue is a PhD student in atmospheric science at the University of the Chinese Academy of Sciences

Wenlu Yuan is a PhD student in the Department of Epidemiology, Biostatistics & Occupational Health at McGill University.

Xiang Zhang is a PhD student in the Department of Geography at McGill University.

Yuanxun Zhang, PhD, is a professor of atmospheric sciences at the University of the Chinese Academy of Sciences and directs the Yanshan Earth Critical Zone at the National Observation and Research Station in China. He has expertise in atmospheric chemistry and measurement of air pollution and its chemical composition. His research focuses on understanding the chemical processes and mechanisms driving the formation of air pollutants and their impacts on climate and human health.

Other publications

Li X, Baumgartner J, Barrington-Leigh C, Harper S, Robinson B, Shen G, et al. 2022a. Socioeconomic and Demographic Associations with Wintertime Air Pollution Exposures at Household, Community, and District Scales in Rural Beijing, China. Environ Sci Technol 56:8308–8318; doi:10.1021/acs.est.1c07402.

Li X, Baumgartner J, Harper S, Zhang X, Sternbach T, Barrington-Leigh C, et al. 2022b. Field measurements of indoor and community air quality in rural Beijing before, during, and after the COVID-19 lockdown. Indoor Air 32:e13095; doi:10.1111/ina.13095.

Sternbach TJ, Harper S, Li X, Zhang X, Carter E, Zhang Y, et al. 2022. Effects of indoor and outdoor temperatures on blood pressure and central hemodynamics in a wintertime longitudinal study of Chinese adults. J Hypertension 40:1950–1959; doi:10.1097/HJH.0000000000003198.