How Do Household Energy Transitions Work?

Jill Baumgartner (Co-PI) Sam Harper (Co-PI)
On behalf of the Beijing Household Energy Transitions Team

2024-07-12

Table of contents

Abstract

Introduction

Since 2015, thousands of villages across Beijing and northern China have been treated by a Coal Ban and Heat Pump (CBHP) subsidy policy that banned household coal burning and subsidized the cost of replacement with 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 CBHP 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 CBHP policy 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 $PM_{2.5}$ were measured in waves 2, 3 and 4. The $PM_{2.5}$ filters were analyzed for mass 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 rollout of the CBHP policy. We used ‘extended’ two-way fixed effects models and marginal effects to quantify the effect of the policy on air pollution and health. We further evaluated whether villages treated by the policy in different years respond differently to the policy, and if any of 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 µg/m^3), and the main source contributors were local and transported dust, regional and domestic coal and biomass burning, and aerosols that form through secondary formation. By waves 2, 3, and 4 there were a cumulative total of 10, 17, and 20 (out of 50 total) villages exposed to the CBHP policy. 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 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 36 µg/m^3. Exposure to 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 (~7 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. The policy did not show evidence of meaningful effects 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 real-world household energy policy in Beijing, we observed high fidelity and compliance with the CBHP policy. 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 energy policies can provide population health benefits.

1 Introduction

China is deploying an ambitious policy to transition up to 70% of households in northern China from residential coal heating to electric or gas “clean” space heating, including a large-scale roll out across rural and peri-urban Beijing, referred to in this document as China’s Coal Ban and Heat Pump (CBHP) subsidy policy. 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 CBHP 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

Beijing 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 CBHP policy, 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, coal combustion from industrial, electricity, and residential heating sources was the single largest estimated contributor to population exposure 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 CBHP policy was considered a potentially important intervention to improve rural development, reduce local and regional $PM_{2.5}$, 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 widely promoted by the Chinese government (Dispersed Coal Management Research Group 2023). By 2021, over 36 million households in northern China were treated by the CBHP policy 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 cleaner household stoves is expected to reduce air pollution, their real-world effectiveness in achieving health-relevant air pollution reductions is unclear (Quansah et al. 2017). In particular, the indoor and local air quality benefits of large-scale household energy programs like the CBHP subsidy policy have been rarely empirically investigated, especially at a sub-city spatial resolution. 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 µg/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 evaluations of the CBHP policy reported small decreases in outdoor $PM_{2.5}$ (-7 to -2.4 µg/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). However, none of these studies included field-based measurements of air pollution or personal exposures, which are known to differ considerably from from modelled estimates based on assumptions of emissions reductions (Thompson et al. 2019), and few accounted for secular changes in air quality over time, limiting any 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) trials found no benefit of LPG stoves on blood pressure (Checkley et al. 2021; Ye et al. 2022) despite much larger 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 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 a coal ban (Dockery et al. 2013). A multi-city study of Chinese adults in cities where the CBHP policy 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 city-wide air quality or health-related policies (Wen et al. 2023).

Though household air pollution is a well-established health risk factor, which energy interventions can reduce air pollution exposures, improve health, and are scalable and sustainable remains a critical and unanswered question. 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% saying no and 45% saying yes (Lai et al. 2024).

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

With several exceptions (Alexander et al. 2018; Gould et al. 2023; McCracken et al. 2007; McCracken et al. 2011), decades of household energy intervention studies have shown limited or no health benefit, which demonstrates the complexity of evaluating interventions on exposures like 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 impacts (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), indoor temperature (Lewington et al. 2012), or behaviors including physical activity (Lindemann et al. 2017). Only recently were such 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 CBHP 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 intervention or policy? Does the policy lead to heating behavior changes that result in colder homes which may 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 CBHP policy'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 (pop. 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 begin in early November and tend to be cold, dry, and windy, with the lowest temperatures most 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 CBHP policy. Roughly half of the villages were expected to enter into the policy during our study (Figure **??**). 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.

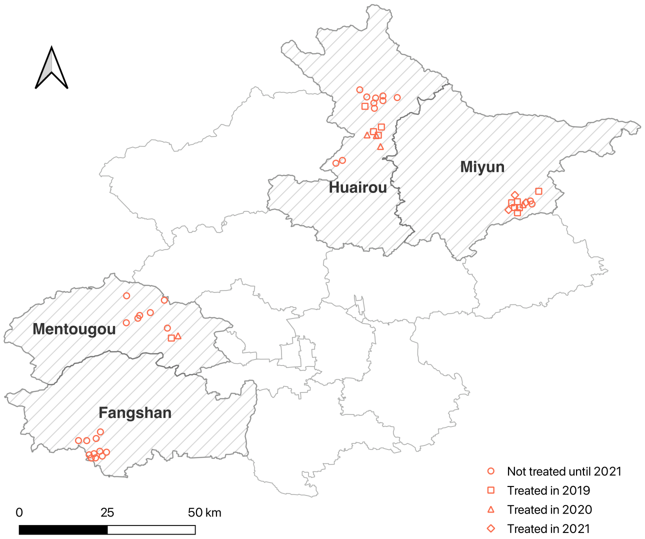

Figure 1: Map of village implementation of CBHP policy

We recruited approximately 20 households in each village and randomly selected one eligible person from each household to participate. 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. If previous participants were not at home or refused to participate, staff first tried to randomly recruit an eligible participant from the same household. If there was not another eligible or willing participant in the household, we randomly selected and recruited a participant from a new household using the village roster. 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 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 clinics to avoid group contact, and blood samples were not collected. Outdoor (community) air pollution was measured throughout the study period.

4.3.1 Air Pollution

Outdoor air pollution

In each village, two sensors for particulate matter air pollution were set up to measure outdoor (community) $PM_{2.5}$ at different locations in each village. One sensor was placed near the center of the village, and the other was placed no less than 500m away from the centrally-located sensor. Sensors were placed at least 1.5m above the ground and not in a location within sight of a visible point source of $PM_{2.5}$.

We collected filter-based community $PM_{2.5}$ samples to calibrate the sensor-based $PM_{2.5}$ measurements as well as to conduct 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). 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. For community measurements, a UPAS was co-located with each $PM_{2.5}$ sensor in each village in rotation. Every week, the used filters were removed and replaced with a new filter. In total, we successfully collected 126, 371, and 289 filter-based, community outdoor $PM_{2.5}$ samples in w1, w2, and w4, respectively. Field blank 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. UPASs 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. $PM_{2.5}$ 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. In w2, 23 quartz-based outdoor $PM_{2.5}$ samples and 3 field blanks were collected. In w4, 11 quartz-based outdoor $PM_{2.5}$ samples and 3 field blanks were collected.

For $PM_{2.5}$ sensor calibration and quality control, all PM sensors were co-located with a reference-grade $PM_{2.5}$ instrument (Model 5030 Synchronized Hybrid Ambient Realtime Particulate (SHARP) Monitor, Thermo Fisher Scientific, United States) on the rooftop of a building at Peking University campus and/or the Tapered Element Oscillating Microbalance (TEOM, Thermo Scientific™ 1405 TEOM™) at the Chinese Academy of Sciences University campus for 7 to 10 days before and after each field wave (Figure **??**). Sensor-measured $PM_{2.5}$ concentrations were highly correlated with those measured by the reference instruments (Spearman correlation coefficients (rho) >0.75 in each pre- and post-calibration).

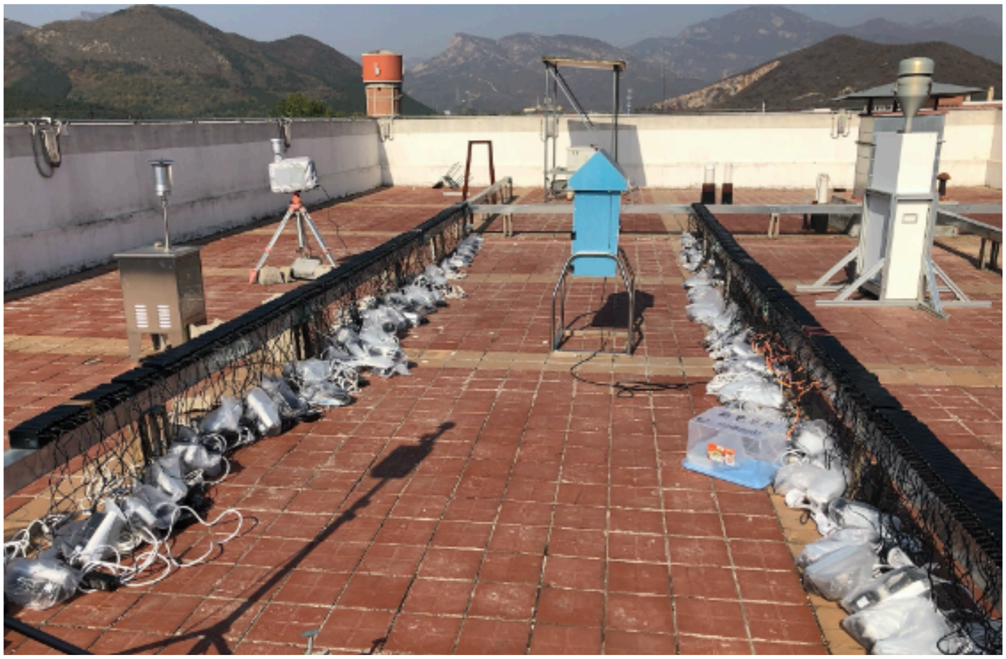

Figure 2: Calibration of real-time sensors against a reference monitor at University of the Chinese Academy of Sciences.

Indoor $PM_{2.5}$

In the second, third, and fourth data collection waves we randomly selected six households from the 20 recruited in each village to measure indoor concentrations of $PM_{2.5}$. In w4, we aimed to monitor indoor $PM_{2.5}$ in the same households where we measured indoor $PM_{2.5}$ in w2. If a household dropped out of the project or declined indoor $PM_{2.5}$ monitoring, we then recruited another household already enrolled in this study to measure indoor $PM_{2.5}$. In total, indoor measurements were conducted in 300 households in both w2 and w4 and 246 households in w3 (Table **??**).

Time-resolved indoor $PM_{2.5}$ was measured using the same commercially available sensor (PMS7003 Plantower, Zefan, Inc.) used for outdoor sensor-based $PM_{2.5}$ and recorded every 1 min. The sensor was placed on a table in a room where participants reported spending most of their time when awake. Indoor $PM_{2.5}$ sensors were deployed between late November and mid January within field waves, depending on the village and household visit schedule. Measurement continued from the time of deployment until sensors were recollected from homes in late April to capture the full heating season.

We randomly selected three households with $PM_{2.5}$ sensors 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 indoor $PM_{2.5}$ sensor measurements. Filter-based $PM_{2.5}$ samples were collected using Ultrasonic Personal Aerosol Samplers (UPAS,

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

Access Sensor Technologies) 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 149 and 148 indoor PM$_{2.5}$ filter samples in w2 and w4, respectively.

As with the community 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 elemental carbon (EC) and organic carbon (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 discontinuation of manufacturing of the Zefluor filter media. 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

To measure personal exposure we used two types of samplers: Personal Exposure Monitors (PEMs, Apex Pro; Casella, UK) and Ultrasonic Personal Aerosol Samplers (UPAS, Access Sensor Technologies, Fort Collins, CO, USA). 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 a small waistpack (for the PEM and sampling pump) or an arm band or 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.

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

All filters were placed in individually labeled cases, sealed in plastic bags, and then transported to a field laboratory and immediately stored in a -20°C freezer. Following completion of the field sampling campaign, the samples and blanks were transported to Colorado State University, where they were stored in a -20°C freezer prior to gravimetric and chemical analysis.

All 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 (L'Orange et al. 2021). Before weighing 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 1µg resolution in triplicate or more, until the differences among the last three weights were less than 3 g. The average of three readings was used to determine filter mass, which was then blank-corrected using the median value of blank filters (3µg for UPAS-collected filters [53% of 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.

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

We established linear regression models between the filter-based $PM_{2.5}$ mass concentrations (i.e., the 'gold standard' reference) and the sensor-based $PM_{2.5}$ concentrations averaged over the same sampling period as the filter-based samples. The slopes of the models were used as the adjustment factors for the sensor-based $PM_{2.5}$ concentrations. Separate regression models were conducted for indoor and outdoor sensors and for each data collection wave given the sensitivity of the sensors to relative humidity, temperature, and particle sources, which may differ for indoor versus outdoor conditions and across waves In w3, where only sensor-based measurements were conducted for indoor $PM_{2.5}$, we applied an adjustment factor developed from a linear regression model that incorporated data from both w2 and w4.

The PM sensors were also evaluated before and after each data collection wave to identify any sensors that needed further repair or replacement. The $PM_{2.5}$ sensors underwent a calibration process that began with synchronization to real-time $PM_{2.5}$ monitors at Peking University (PKU) campus.

This pre- and post-wave calibration included a week-long session using the Beta Attenuation Monitor (BAM) alongside daily 24-hour filter samples. During this time, approximately 240 sensors were placed on the rooftop of the College of Urban and Environmental Sciences building, each recording data every minute. A similar approach was taken at the University of Chinese Academy of Sciences (UCAS) campus, where around 400 PM sensors were installed on the rooftop of the Environmental Monitoring Site of the College of Resources and Environment, with data logging at one-minute intervals. Daily collections of 24-hour PTFE and quartz filter samples accompanied the sensors' measurements to ensure accuracy. The calibration process was repeated post-fieldwork to account for any potential shifts or discrepancies in sensor performance. This approach aimed to maintain consistent and accurate measurements from the PM sensors throughout the study.

Chemical analysis of PM mass

We analyzed the chemical composition of community and personal exposure $PM_{2.5}$ samples to quantify the individual components and species. $PM_{2.5}$ component concentrations were determined by dividing the quantified component mass by the sampled air volume, after correcting for field blanks collected in the corresponding wave.

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 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 (Ca2+, Mg2+, Na+, NH4+, 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 (SO42−, NO3−, Cl−) (Sullivan et al., 2008).

Organic (OC) and elemental carbon (EC) on PTFE filters were 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, and these properties are used to develop the 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 the Teflon (PTFE) and quartz filters collected 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 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 determined from quartz filters collected indoors and outdoors (as described above). $PM_{2.5}$ samples collected on quartz filters were analyzed for OC and EC using a Sunset Laboratory OC/EC Lab instrument (Sunset Laboratories, Inc., MODEL, USA) according to 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 mass. For the thermal desorption component, the sample is oxidized twice, according to a strict temperature regime. The first oxidation stage thermally removes OC in a mobile phase of pure helium gas to be converted from carbon dioxide (CO2) to methane (CH4) 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 by OC pyrolysis during the first heating 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.

Following gravimetric analysis, all PTFE filters were also analyzed 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 sampling in the field. 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.

For statistical analysis, we estimated the effect of the policy on personal exposures to $PM_{2.5}$ and

BC using the results from filter-based measurements collected over 24-h periods. We measured indoor and outdoor $PM_{2.5}$ for up to 6 months in our study households, and thus we calculated both the 24-h mean values (to coincide with the same 24-h period that personal exposure samples were collected) and the wintertime seasonal mean values (with winter 'season' defined as January 15 to March 15) of $PM_{2.5}$.

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 excluded from 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-fuelled 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 mo - 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 staff and adapted prior to implementation. Prior to each wave in this study, the questionnaire and 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 assess whether the questions were understandable and 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 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 logistic 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 validated against invasive cBP measurements in previous 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. 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 each standard 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 were measured in duplicate and samples were measured once due to high precision in a pilot study (Food and Drug Administration 2018).

4.3.8 Anthropometric measurements.

Body weight, height, and waist circumference were measured at the clinic visit in the first two waves and in participant homes in the last wave. 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 stepped 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 tape was 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 rollout 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 CBHP policy 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, impacts of the COVID-19 pandemic). By comparing the *changes* in outcomes among treated villages to the *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.

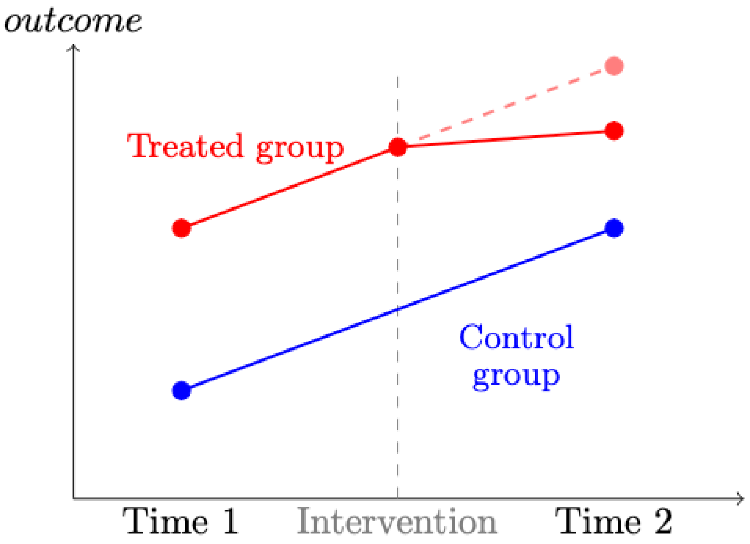

Figure 3: Stylized example of difference-in-differences

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 **??**). 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, 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 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 rollouts 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 CBHP intervention 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. Taking as an example the directed acyclic graph (DAG) in Figure **??**, with T as the policy, X as a set of pre-treatment covariates, M_1 as $\mathrm{PM}_{2.5}$, M_2 as indoor temperature, and Y as systolic blood pressure, we can define the controlled direct effect (CDE) as the effect of the CBHP policy on systolic blood pressure if we fix the values of $\mathrm{PM}_{2.5}$ 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 $\mathrm{PM}_{2.5}$ and indoor temperature at uniform levels of average background exposure, or some other hypothetical level.

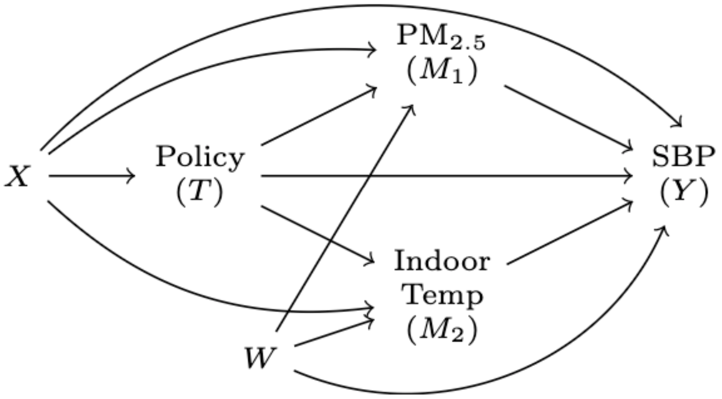

Figure 4: Hypothetical Directed Acyclic Graph showing direct and indirect effects with outcome (Y), pre-treatment covariates (X), policy (T), multiple mediators (M_1, M_2), as well as covariates for the mediators (W).

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

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

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 cohorts of villages that were first exposed to the policy at the same time q (e.g., in 2019, 2020, or 2021), fs_t are time fixed effects corresponding to different winter data collection waves (2018-19, 2019-20, or 2021-22), and τ_{rt} are the cohort-time ATTs in the context of a linear model. For binary or count outcomes the cohort-time ATTs are derived by estimating marginal effects from non-linear models (Arel-Bundock 2024). For all models we cluster 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 **??**. 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 **??** as follows:

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

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 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 $\text{PM}_{2.5}$. Based on the setup in Equation **??** the CDE is estimated as: $\delta + \eta_{rt} MT$. In the absence of interaction between the exposure and the mediator (i.e., $\eta_{rt} = 0$) the CDE will simply be the estimated treatment effects $\sum_{r=q}^{T} \sum_{s=r}^{T} \tau_{rt}$, 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 a total effect of the CBHP policy.

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). The latter are particularly helpful in the context of the documented declines in $PM_{2.5}$ in China attributable to sources other than the CBHP policy (Van Donkelaar et al. 2021; Zhang et al. 2019)

For models estimating the effect of the policy on 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 CBHP policy. In the multivariable models, we also adjusted for strong predictors of the outcome that were not affected by treatment, and thus not confounders, to improve model precision. The covariates included in each of the models are provided in the tables.

For air pollution outcomes, we considered the following covariates: village population and total number of households in the village; temperature, relative humidity, wind direction, wind speed, boundary layer height; home area and home area heated; home insulation; smoking status of participant and whether or not they lived with a smoker; whether or not the household reported using wood (i.e., biomass) for household energy activities, and if so, self-reported quantity of wood. Potential non-linearity between continuous covariates and our study outcomes were evaluated using natural cubic splines with different degrees of freedom. Ultimately, the following covariates were included in the final DiD models for outdoor, indoor, and personal exposures to air pollution, based on whether measurable changes in the covariate over time were observed. For the final adjusted DiD model for personal exposure source contributions due to mixed combustion of solid fuels (hereafter 'mixed combustion'), we adjusted for: temperature (represented by a spline with 2 degrees of freedom); participant smoking status; and whether or not the household reported using biomass fuel. For the final 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