Cancer and industrial activities in China

Xuemei Zhang

Shanghai Jiao Tong University

Haitao Yin (✉ htyin@sjtu.edu.cn)

Shanghai Jiao Tong University

Jay Shimshack

University of Virginia

Analysis

Keywords: pollution, cancer, industrial activity, China

Posted Date: August 11th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-709045/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Associations between pollution and life expectancy, infant mortality, and cardiorespiratory disease are documented in China. Yet, less is known about environmental drivers of Chinese cancers. Here, we systematically link polluting industrial activity to cancer incidence, cancer mortality, and cancer cluster designations. We investigate county-level associations between industrial production and age-adjusted incidence and mortality reported in official cancer registries. We then combine the locations of roughly 3 million enterprises with administrative data from roughly 600,000 villages and cancer cluster documentation from 380 villages. We show that county-level value-added from industry is associated with age-adjusted incidence and mortality for all cancers; bronchus, trachea, and lung cancers; stomach cancers; and esophageal cancers. We show that the odds that a village contains a documented cancer cluster increase 3-4 times if the village contains a pollution-intensive industrial facility. Leather, chemical, and dye enterprises appear to drive results. All else equal, smaller facilities increase the odds of cancer clusters.

Main Text

China's per capita income has increased roughly 25-fold since 1990. This rapid growth is associated with improvements in education and economic welfare for the average Chinese household. However, this growth is also linked to well-chronicled increases in pollution and pollution-induced morbidity and mortality (1, 2). An established literature explores associations between pollution and life expectancy, infant mortality, and cardiorespiratory disease in Western countries (3-6) and in China (7-13). A smaller literature links environmental conditions to Chinese cancer (14-16). Yet, cancer is a leading cause of death (17). Recent estimates suggest 2 to 3 million Chinese die of cancer each year and 3 to 4 million Chinese receive a new cancer diagnosis annually (17-20). Stomach, liver, and esophageal cancer incidence rates are more than 2 times global averages and lung cancer incidence is more than 1.5 times the global average (18). Cancer mortality is increasing rapidly relative to other causes of death in China, and the problem is especially acute in rural areas (17).

Even when pollution and cancer in China have been studied, little systematic evidence identifies sources. A growing literature documents the role of environmental determinants of cancer worldwide (*Text S1*). The International Agency for Research on Cancer (IARC) classifies more than one hundred agents to be carcinogenic to humans, for many of which industrial production represents a significant source. Manufacturing enterprises in China do indeed use or generate significant quantities of known carcinogenic particulates, organic and inorganic compounds, metalloids, and heavy metals. Common substances of toxic significance associated with industrial production include benzene, benzidine, formaldehyde, trichloroethylene, arsenic, cadmium, chromium, lead, mercury, and nickel (*Text S1*). Arsenic, benzene, cadmium, chromium, nickel, and other toxic substances are known to be a cause to lung and other respiratory cancers, and lead, arsenic, and other organic and inorganic substances are causally related to stomach, liver, and esophageal cancers (*Text S1*).

In spite of this lab-supported linkage between industrial production, emission of carcinogenic agents and cancer incidence, actual systematic evidence linking polluting industrial facilities to cancer incidence and mortality in China remains limited (16,21). Agricultural chemicals, heavy metals from sewage and irrigation, municipal solid and hazardous wastes, and transportation activities could be confounding pollution sources also driving variation in cancer incidence and mortality (21). Existing scholarship, given objectives like policy evaluation, often remains agnostic on sources or generally presumes that industrial pollution is a key mechanism driving pollution-induced changes in cancer incidence (15,21).

Here, we link polluting industrial activity to cancer across mainland China. As experimental or individual-level cohort data are unavailable at scale, we construct observational evidence using geo-located industrial activity and local cancer incidence, cancer mortality, and cancer cluster data. We first investigate cross-sectional associations between county-level industrial output and age-standardized cancer incidence and mortality constructed from official cancer registries. We explore determinants of county-level incidence and mortality from all cancer; trachea, bronchus, and lung cancer; stomach cancer; liver cancer; and esophageal age-adjusted cancer. We then conduct an observational study where ~600,000 villages represent units of analysis. We detail relationships between exact locations of industrial establishments and documented village-level cancer clusters. We match the physical locations of ~3,000,000 industrial facilities to health data on ~380 publicly acknowledged village-level cancer clusters with administrative records for ~600,000 Chinese villages. Village-level data allow for a fine-scale yet comprehensive population-level analysis; on average, each village represents around 2,000 people and an area smaller than 10km².

Given observational epidemiologic data rather than experimental data, we pay careful attention to minimizing common biases. In the village-level analyses, we minimize omitted variable bias (confounding) and measurement error (classification errors) with control variables, fixed effect methods, instrumental variable, and other techniques (22-27). Control variables include economic, political, and geographic measures. Fixed effect approaches exploit variation across villages within counties to compare exposure villages only to control villages within the same county. Instrumental variable and bivariate probit approaches enhance causal attribution by exploiting variation in an exogenous proxy correlated with exposure (industrial facility locations) but plausibly otherwise uncorrelated with cancer (22-27). We explore the possibility of differential misclassification of outcomes, where with (without) polluting industrial facilities might be more (less) likely to be classified as “cancer clusters” or “cancer villages,” holding actual cancer incidence and mortality constant.

Broad associations between industrial production and cancer

We first document county-level relationships between value-added from industrial production (in hundred million CNY) and all-cause cancer mortality and incidence reported in official cancer registries. Counties in the 1st-4th quartiles (denoted Q1-Q4) of value-added from industrial production exhibit the following observed cancer rates: cumulative male mortality rates to age 74 - Q1, 16.1%; Q2, 18.5%; Q3, 18.5%; and

Q4, 19.8%; cumulative male incidence rates to age 74 - Q1, 23.1%; Q2, 27.0%; Q3, 26.5%; and Q4, 29.0%; cumulative female mortality rates to age 74 - Q1, 8.5%; Q2, 9.4%; Q3, 9.4%; and Q4, 9.6%; cumulative female incidence rates to age 74 - Q1, 15.3%; Q2, 17.7%; Q3, 17.4%; and Q4, 18.6%. Tests of equivalence reject (at the 5% level) a null of no difference in mortality and incidence between Q1 and Q4 for all matched pairs. Counties with lower value-added from industrial production experience statistically lower all-cause cancer incidence and mortality relative to counties with higher value-added from industrial production.

Fig. 1 and **Fig. 2** document more complete county-level associations between industrial value-added and National Central Cancer Registry (NCCR) incidence and mortality. In the figures, confidence intervals are imprecisely estimated in the tails due to limited data. This fact notwithstanding, the top panels of **Fig. 1** show significant positive associations between county-level value-added from industry and age-adjusted all cancer incidence and mortality for males. The bottom panels of **Fig. 1** show significant positive associations between county-level value-added from industry and age-adjusted all cancer incidence for females, but associations between industrial production and cancer mortality for females are less significant. **Fig. 2** shows that age-adjusted bronchus, trachea, and lung; stomach; and esophageal cancer mortality for males co-move with value-added from industry. We detect no clear relationships between industrial production and age-adjusted liver cancer mortality for males. All qualitative patterns are similar in age-adjusted incidence for males, age-adjusted mortality for females, and age-adjusted incidence for females (*Fig. S1, Fig. S2, Fig. S3*).

We next explore associations between local industrial production and the location of village-level cancer clusters (or “cancer villages”). Here, we construct cancer cluster data from official Chinese media sources. We characterize 380 villages in 212 counties as documented cancer clusters (28-31). Incidence and mortality specifics are unavailable for many of these cancer villages, and details may be reported with error. Where available, reported incidence and mortality rates in cancer villages are roughly 3 to 5 times larger than national averages (*Text S2*).

Fig. 3 documents cancer clusters’ locations and illustrates the strong association between industrial production and cancer villages. Counties in Q2, Q3, Q4 of the share of employment in industrial activities have 1.7, 1.9, and 2.8 times more reported spatial cancer clusters than counties in Q1. Counties in Q2, Q3, Q4 of value-added from industrial activity have 4.0, 6.3, and 11.4 times more reported spatial cancer clusters than counties in Q1. Even if we restrict the sample to prefectures that contain at least one recognized village-level cancer cluster, counties with cancer clusters have average GDP and value-added from industrial activities that are 63% ($p < 0.01$) and 79% ($p < 0.01$) higher than relatively similar counties without cancer clusters (*Table S1*).

Village-level regression analyses

Counties with high and low cancer incidence may differ substantially in other ways so the above associations are not necessarily causal. Socio-economic characteristics associated with economic development are not necessarily related to higher cancer incidence and mortality in official cancer

registries (32, 33, *Fig. S4*). Counties with documented “cancer villages” are statistically no more likely to be near major rivers or provincial borders (*Fig. S5*) and are statistically no more likely to experience worse baseline health (*Table S2*). Nevertheless, counties with cancer villages are more populous, less agricultural, more educated. They contain wealthier households, have fewer minorities per capita, and are disproportionately located in eastern China (*Table S2, Fig. S5*).

Thus, we explore village-level regression analyses. As discussed in *Methods*, we take a variety of approaches to minimize omitted variable bias (confounding), measurement error (classification errors), and other statistical concerns. We discuss assumptions necessary to interpret village-level relationships as plausibly causal evidence linking industrial facilities to cancer in China. We explore sensitivity. Since villages with and without cancer clusters in the full sample differ on several village-level economic, geographic, and political measures (*Table S3, top panel*), we also analyze both a full sample of 599,822 villages and a restricted sample of 73,157 villages located only in counties with at least one cancer village. Restricted sample analyses make cleaner “apples to apples” comparisons of “case” and “control” villages that differ less on observables other than exposure measures of interest (*Table S3, bottom panel*).

Standard case-control analyses at the village-level show that the odds that a village contains a spatial cancer cluster are increasing with the presence of polluting industrial facilities (*Table S4*). In a full sample analysis (599,822 villages), the odds of a village containing a spatial cancer cluster increase 3.88 times (95% CI: 2.93 – 5.14) if the village contains one or more polluting industrial facilities. In a restricted sample of villages only in counties with at least one cancer cluster, the odds of a village containing a spatial cancer cluster increase 2.82 times (95% CI: 2.17 – 3.65) if the village contains one or more polluting industrial facilities. Results are insensitive to including village-level economic or geo-political controls (*Table S4*).

Table 1 presents results from regression analyses that more completely address confounding and other statistical concerns. **Table 1** suggests three qualitative take-home messages. First, even with regression approaches designed to enhance the plausibility of causal attribution, the probability that a village contains a spatial cancer cluster is strongly increasing with the presence of polluting industrial facilities. Second, estimates are reasonably robust across a host of empirical methods. Third, results are reasonably stable when adding (or omitting) economic, geographic, and political control variables.

Quantitative marginal effects in **Table 1** communicate the incremental probability that a village contains a documented cancer cluster if the village contains a polluting industrial facility (relative to a control village in the same county without a polluting industrial facility). Empirical magnitudes range from 0.007 to 0.059, with common estimates falling between 0.007 to 0.014. Thus, villages with one or more polluting industrial facilities are expected to have a 0.007 – 0.014 higher probability of being designated as a “cancer village.” Given that the baseline probability that a non-exposed village is classified as a cancer village is around 0.004, these marginal effect magnitudes suggest that the odds of containing a cancer cluster increase 3 to 4.5 times when a village is exposed to a polluting establishment.

These results are similar to results from standard case-control methods that may isolate less plausibly causal relationships.

Fig. 4 summarizes results from village-level analyses where key explanatory variables are disaggregated by sector to explore heterogeneity (*Table S8*). The most robust finding is that statistically significant associations between spatial cancer clusters and polluting industries may be largely driven by chemical, dye, and leather/tanning facilities. Interpreting the coefficients summarized by the bottom bars in each row of **Fig. 4** as odds ratios, we show that the odds that a village contains a spatial cancer cluster increase 2.62 times ($p < 0.01$, 95% CI: 1.64 – 4.17) if the village contains one or more chemical facilities, 1.87 times ($p < 0.01$, 95% CI: 1.07 – 3.27) if the village contains one or more dye facilities, and 7.44 times ($p < 0.01$, 95% CI: 3.48 – 15.87) if the village contains one or more leather facilities.

Variation in relationships

We consider the relative effects of total polluting industry size and average polluting facility size (*Table S9*). We find that villages with larger total polluting industrial sectors are associated with greater odds of spatial cancer clusters. However, given a fixed total polluting industry size, larger numbers of smaller facilities are associated with greater cancer cluster odds than smaller numbers of larger facilities. Conditional on a village containing at least one polluting industrial facility, the odds of the village containing a spatial cancer cluster increase between 1.36 times ($p < 0.01$) and 1.28 times ($p < 0.01$) if the total size of the polluting sector doubles. Conditional on a fixed total polluting sector size, each additional polluting facility is associated with an increase in the odds of a village containing a spatial cancer cluster of between 1.023 times ($p < 0.07$) and 1.029 times ($p < 0.02$). We also consider spatial heterogeneity across north vs. south regions and across rural vs. urban areas (*Table S10*). We find no statistical evidence that village-level relationships between industrial facilities and cancer cluster designations differ statistically by region or land use. We do not find robust statistical evidence for more general spatial heterogeneity in the cancer registry data (*Table S11, 34*).

Discussion

This study provides novel evidence that cancer incidence, mortality, and cluster designations are associated with industrial activity in China on a comprehensive scale. Results have potential implications (35-37). First, findings may inform the benefits and costs of public health investigations and policy interventions in China. Results suggest that benefits of industrial pollution policies may be understated if based solely on health outcomes such as infant mortality and cardio-respiratory illness that are studied in the extant literature. Second, findings suggest that pollution impacts in China may have highly localized components. Results shed light on social trade-offs between regional or national pollution policies that presume relatively homogeneous pollution damages versus policies that are more locally targeted to specific conditions. Abating pollution where and when health damages are potentially higher may generate a large public health ‘bang per buck’.

More generally, this work informs a multidisciplinary discourse on where and how cancer may cluster in space (35-37). The analysis highlights patterns that may suggest triage strategies for public health and medical care investigations. Villages with large numbers of small industrial facilities may merit thorough investigation by disease control and medical care experts. Villages with active leather, chemical, and dye enterprises may benefit from careful study by disease experts. Contaminants common in these sectors may warrant further investigation in the lab and in the field. Finally, results inform targeted risk communication strategies that may enhance local understanding, trigger specific abatement efforts by local facilities, or spur household-level avoidance behaviors.

We note potential threats to causal interpretation and emphasize that interpreting our results as causal requires assumptions documented in the *Methods* section. We take multiple approaches to address confounding, each with its own strengths and weaknesses (22-27). We pay special attention to concerns about differences in smoking and diet across villages. These behavioral factors are known to be a significant cause to cancer. We alleviate this concern via the followings: (i) we use a fixed effect at the county-level, and therefore rely on within-county variation for identification. The estimates are not biased as long as residents in different villages of the same county does not differ much in their smoking or diet; (ii) we investigate the China Family Panel Studies data and find no correlation between smoking or diet and the presence of a polluting facility. Therefore, omitting smoking or diet is unlikely to introduce endogeneity and bias the estimates (*Table S12*)

We believe selection bias is relatively minor, as county-level analyses rely on cancer registries explicitly constructed to be nationally representative (20) and village-level analyses examine a near census of villages outside of sparsely populated far western and northwestern China (*Text S2*). We reach similar conclusions with village, town, and county-level analysis. Thus, multilevel investigation suggests that an ecological fallacy or the related modifiable areal unit problem are unlikely to drive our results (38-39). We advocate for follow-up on individual data. Our statistical methods minimize several non-differential misclassification of exposure concerns. We do acknowledge that very small facilities, with annual sales under 5 million RMB, are underrepresented in the sample dataset. As such, relevant village-level results should be interpreted as the effect of medium and large size facilities on the probability of designation as a cancer cluster. Although the results are not biased by the omission of the smallest industrial establishments, external validity for the smallest polluting industrial establishments is not assured.

Differential misclassification of outcomes is possible in the village-level analysis. The concern is that villages with (without) polluting industrial facilities might be more (less) likely to be classified as “cancer clusters” or “cancer villages,” holding actual cancer incidence and mortality constant. Absent methods that eliminate the concern, we explore differential misclassification with quantitative sensitivity analyses (40). Explorations suggest that we would need roughly 180% more “unexposed cases” (designated cancer villages in locations without industrial facilities) via false negatives for the truth to be ‘no relationship between the location of polluting industrial facilities and cancer’ (*Text S3*). Explorations suggest more than 64% of “exposed cases” (designated cancer villages in locations with industrial

facilities) would have to be false positives for truth to be ‘no relationship between the location of polluting industrial facilities and village-level cancer’ (*Text S3*). Although we are unable to conclusively rule out differential misclassification of outcomes, misclassification would have to be large to explain findings.

We are unable to provide direct evidence on timing, biomedical mechanisms, and exposure pathways. We acknowledge that the complex etiology and long latency of cancers poses challenges for linking observed cancer incidence, mortality, and cluster designations to environmental causes (35-37). We note that latency may vary across site-specific neoplasms. Our cross-sectional statistical relationships are best interpreted as reflecting longer-term epidemiological relationships. We find robust evidence that trachea, bronchus, and lung cancers; stomach cancers; and esophageal cancers are linked to polluting industrial activity in China. These cancer sites are commonly related in laboratory, in vitro, and epidemiologic studies to contaminants found in industrial pollution (18). We find particularly significant relationships for the leather, chemical, and dye industries. These sectors use or generate significant quantities of known carcinogenic heavy metals, organic chemicals, inorganic chemicals, and other substances of toxic significance (*Text S1*). Industrial pollution may influence cancer incidence, mortality, and cluster designations via ambient exposure or through occupational exposure. This study’s comprehensive population-based results are roughly consistent with smaller-scale investigations of occupational risk factors and cancers (41), but also seem to apply to diverse populations and settings. A goal of this work is to highlight promising directions for future research into mechanisms, exposure pathways, and specific contaminants.

Methods

Administrative data. We construct an administrative dataset from China National Bureau of Statistics (NBS) 2010 data. We observe the universe of the 2,457 county-level divisions (henceforth “counties”) outside of the five large, remote, and less populated western / northwestern provinces of Tibet, Xinjiang, Inner Mongolia, Qinghai, and Gansu. We obtain geographic information system (GIS) and county-level socio-demographic data from the 2010 Population Census of China, the PRC National Bureau of Surveying and Mapping, and the China Data Center at the University of Michigan. We merge these data to county-level data from cancer registries.

Administrative data include 641,022 village-level divisions (henceforth “villages”). The typical village has a land area of 10 km² and a population of roughly 2,000 (*Text S2*). We retain the 599,822 villages with identifiers and geocodes that facilitate matches to industrial facility locations and other data. We compared sample villages (93.6%) to villages with missing or unmatched identifiers (6.4%) and the practical differences are small (*Text S2*). For each of the final 599,822 sample villages, we observe the socio-demographic characteristics of its county. We observe latitude and longitude for the village centroid, which we use to construct village-level geopolitical variables including distance to nearest province border, county border, river, coastline, large city (urban hukou population > 1 million), and major railway. We merge these data to village-level industrial enterprise data and village-level cancer data.

Industrial facility and economic activity data. We obtained establishment data from the NBS 2008 Second Economic Census. The dataset contains village of location, industry, and size information for roughly 8,864,000 establishments in China engaged in “secondary” (industrial) or “tertiary” (service) business activities (*Text S2*). Establishment data include all facilities owned by the state, and all facilities owned by domestic or foreign private owners with annual sales exceeding 5 million CNY, or about 700,000 USD (*Text S2*). Comparisons with published statistics suggest sample enterprises represent over 91% of total 2008 revenue. Results will not necessarily apply to private facilities with annual sales below 5 million CNY.

Census establishments include manufacturing facilities and utilities, wholesale and retail facilities, educational institutions, health and social welfare institutions, and many others. We break manufacturing facilities and utilities into “polluting” industrial facilities and “non-polluting” industrial facilities using 4-digit sector codes following China Ministry of Environmental Protection (MEP) conventions (42) (*Text S2*). Nearly all facilities in the chemical, dye, fiber, leather, pharmaceutical, cement, coking, electricity, food, iron and steel, metals, paper, rubber, and vegetable oil industries are classified as “polluting” (*Text S2*).

For each of the 599,822 sample villages, we merged in facility information to create several village-level business activity measures: presence (or number) of polluting industrial facilities, presence (or number) of non-polluting industrial facilities, employees at polluting industrial facilities, output at polluting industrial facilities, retail activity, employees in educational sector, and employees in health and social welfare sector. For our main analysis, we merged facilities to villages using NBS’s 12-digit village committee codes (*Text S2*).

Cancer data. County-level analyses use cancer incidence and mortality data from the China Ministry of Health’s 2013 National Central Cancer Registry (NCCR). We follow convention and define incidence by the probability of new cancer diagnosis and mortality by deaths attributed to cancer. NCCR mortality data at each registry are collected from hospital and health station medical records, and new diagnoses are reported by hospitals, health stations, and individual doctors to local disease control centers who in turn report to NCCR. Underlying disease surveillance points were chosen using clustered random sampling with the goal of approximating a nationally representative sample (16,20). We chose the 2013 dataset because registry population and geographic coverage has been growing rapidly since 2008, and the 2013 data were the most comprehensive available to us. Officially qualified 2013 NCCR data summarized information from 255 registries covering approximately 226 million people (17). We analyze data from the 193 registry subset that recorded information at the county-level. Sample NCCR data include incidence and mortality for all cancers and for the four most common cancers in the country: trachea, bronchus, and lung; liver; stomach; and esophageal cancer. Data are age-standardized using cumulative rates, i.e. probability of cancer incidence or mortality to age 74.

NCCR summary statistics show cumulative incidence rates to age 74 of 26.2% (s.d. 7.2) for males and 17.2% (s.d. 4.0) for females. We find cumulative mortality rates to age 74 of 18.0% (s.d. 5.0) for males and 9.1% (s.d. 2.4) for females. Site-specific cumulative incidence rates to age 74 for males are as

follows: 6.4% (s.d. 2.0) for trachea, bronchus, and lung; 4.7% (s.d. 3.7) for stomach; 3.4% (s.d. 3.0) for esophageal; and 3.7% (s.d. 1.7) for liver. Site-specific cumulative incidence rates to age 74 for females are as follows: 2.7% (s.d. 1.0) for trachea, bronchus, and lung; 1.7% (s.d. 1.3) for stomach; 1.4% (s.d. 1.7) for esophageal; and 1.3% (s.d. 0.6) for liver. As a general rule, these statistics are similar to published estimates (17).

Cancer registry data are considered reliable. A potential disadvantage is spatial autocorrelation and the modifiable areal unit problem that can arise with spatially aggregated data (39). However, tests suggest spatial autocorrelation in the county-level data is statistically significant but small. Moran's I spatial correlation coefficients range from 0.02 to 0.03 for age-adjusted all cancer incidence and mortality. A greater issue is that, although registry data are useful for identifying trends or program evaluation, they may be less useful for this study's goal of systematically understanding *local*-level correlates of cancer. As such, much of our analyses rely on local-level data from around 600,000 villages across mainland China. These village-level "cancer cluster" or "cancer village" data cover the overwhelming majority of the population.

"Cancer village" data primarily represent media-reported indicators for village-level cancer clusters (*Text S2*). "Cancer villages" are widely documented in sanctioned Chinese media and in sustainability sciences scholarship (28-31). Following the earlier literature, we identified spatial cancer clusters reported by Chinese media, of which 380 were village-level clusters matching villages in our sample dataset. >40% of identified "cancer villages" were identified by official government-sanctioned news sources like China Central Television (CCTV), the Xinhua news agency, People Daily, or a government agency website. >75% of "cancer villages" were classified as spatial cancer clusters by at least one official Chinese government news source or a reputable nationally-circulated Chinese magazine or journal. "Cancer villages" are widespread, with at least one village-level cancer cluster in all but one of our sample provinces (*Fig. S6*).

"Cancer village" data have potential disadvantages. Documented cancer cluster data are surely measured with error. One concern is non-differential misclassification of outcomes (classical measurement error on the dependent variable), but that simply attenuates statistical precision without biasing estimates. A greater concern is differential misclassification of the outcome (non-classical measurement error on the dependent variable). We explore this issue in detail elsewhere in the paper, but we note here some relevant institutional context. False positives that appear prevalent in the developed country "cancer clusters" literature may be less likely in the Chinese setting where official Chinese media outlets may have incentives to minimize attention and public concerns (35-37). Government authorities have publicly acknowledged the existence of "cancer villages" and have identified many of the same specific clusters that we analyze (28-31). False negatives may also be less likely, since our village-level analyses exploit within-county variation only. Misclassification requires one or more true "cancer villages" in a given county to be designated while other true "cancer villages" in that same county are not designated.

Three other limitations of “cancer village” data bear noting. First, drawing inference from data aggregated above the individual-level may raise concerns about “ecological fallacy” or related issues (38-39). We draw the same conclusions from village-, town-, and county-level analyses, minimizing the likelihood that these concerns drive our results. Nevertheless, stimulating follow-ups on individual-level data is one of this study’s goals. Second, village-level cancer cluster data are cross-sectional. Summary statistics indicate that sample industrial enterprises were typically constructed well before the villages were designated as cancer clusters (*Text S2*). Nevertheless, cross-sectional data preclude statistical identification from within-group variation over time. Third, cancer cluster data do not allow us to identify site-specific neoplasms.

Summary statistics. We calculate sample means and standard deviations for county-level NCCR cancer registry data and for village-level cancer cluster data. At various points, we illustrate county-level and village-level characteristics with sample means and standard deviations.

County-level correlations analysis. We graphically document county-level associations between industrial activity and age-standardized cancer incidence and mortality as reported in official cancer registries. We plot cumulative cancer incidence or mortality to age 74 against the natural log of value added from industrial activity (measured in hundred million CNY) on a scatter diagram. We log industrial activity because the baseline distribution is restricted to the positive domain and right skewed. We highlight possible associations by overlaying the scatter plot with fitted fractional polynomial regression predictions and associated 95% confidence intervals (CIs). Fractional polynomial regressions are in the spirit of standard polynomial regressions but allow more flexible parameterization (43). Best fit is determined by sums of squares, as in standard OLS. We note the standard issue that CIs are imprecise in the tails of the distribution due to limited data.

We also document county-level associations between industrial activity and the presence of a “cancer village” cluster within the county. We overlay the physical location of cancer villages on a map illustrating the quartile of county-level share of employment from industrial activity.

Baseline village-level analysis. The baseline village-level analysis is a standard logistic case-control analysis (44, *Text S4*). We compare the odds that a village contains a spatial cancer cluster across villages with and without polluting industries. We analyze the full 599,822 village sample. Then, in order to make villages more comparable, we analyze the 73,157 villages in counties with at least one identified cancer cluster. This latter sample restriction retains all village-level “cases” but ensures that control villages are located in the same counties and thus more similar. Some baseline analyses include additional village-level control variables.

We extend the simplest model. Innovations relative to the simplest case-control regressions are county-level fixed effects and village-level covariates. The models can be thought of as case-control analyses using statistical techniques to control for *all* factors varying at the county-level and using observed explanatory variables to adjust for other economic, political, and geographic confounders varying at the village-level within counties. Fixed effects control for county-level confounders such as

average economic activity, socio-demographics, genetic differences, and health behaviors. Fixed effects sweep out systematic differences in reporting across counties. Village-level control variables include the presence or size of non-polluting manufacturing facilities in the village, the size of the village's educational sector, the size of the village's health care and social services sector, and/or total retail sales. These measures proxy for economic activity possibly correlated with polluting industrial activity and health at the village-level (5). Controls may also include distances to the nearest big city, the nearest county border, the nearest province border, and the nearest large river. These measures proxy for political and geographic factors possibly correlated with polluting industrial activity and health at the village-level (45).

The related literature suggests that individual-level behavioural factors like smoking and diet influence cancer incidence and mortality (*Text S1*). Due to data limitations we are unable to control for these factors directly. We show, however, that smoking and diet are uncorrelated with the presence of polluting industrial facilities so their omission should not bias our main results (*Table S12*). The use of fixed effects model also alleviate this concern with the plausible assumption that smoking and diet may not dramatically differ across different villages within the same county.

To parallel standard case-control logistic analysis, fixed effects results emphasize conditional fixed-effects logit models (*Text S4*). We also explore robustness to different functional form choices with fixed effect linear probability models (*Text S4*). In principle, coefficients of interest from all models represent the effect of one or more polluting industrial facilities on the probability or log-odds that the village is a spatial cancer cluster, after controlling for county-level fixed effects and observable village-level covariates. To enhance clarity with logistic results, we often interpret the odds ratio (the natural exponentiation of the coefficient), which represents the effect of polluting industrial facilities on the odds that the village is a spatial cancer cluster.

Instrumental Variables and Bivariate Probit Methods. Despite advantages related to familiarity and interpretation, conventional regression approaches may be subject to bias that could hinder causal interpretation. Omitted variable bias (confounding) could arise if the specific locations of industrial facilities are correlated with unobserved factors also correlated with cancer incidence, mortality, and cluster designations. We attempted to minimize such concerns by including fixed effects and controls. Nevertheless, we allow the possibility that estimates remain biased by unobservable differences like smoking and diet across villages in the same county. Bias could also arise if household locations within counties are not randomly assigned and socioeconomic confounders not included in the model are correlated with both the industrial facility locations and cancer via mobility and migration-based sorting.

The third and fourth research designs enhance the plausibility of causal attribution by exploring robustness to methods designed specifically to reduce confounding above and beyond fixed-effect regressions. We use techniques that exploit variation in an exogenous proxy correlated with the explanatory variable of interest but uncorrelated with other cancer determinants. Our chosen proxy is the natural log of the village's distance to the nearest major rail line. For village i in county c :

$$[1] \quad FACILITY_{ic} = \delta + \phi Z_{ic} + X_{ic} \Theta + \gamma_{ic}$$

$$y_{ic} = \tau + \rho FACILITY_{ic} + X_{ic} \Delta + \mu_{ic},$$

where $FACILITY_{ic}$ is an indicator equal to 1 if the village contains at least one polluting industrial facility and 0 otherwise; Z_{ic} is the railway proxy; y_{ic} is equal to 1 if the village is a spatial cancer cluster and 0 otherwise; X_{ic} is a vector of economic, political, and geographic control variables measured at the village-level, and γ_{ic} and μ_{ic} are error terms. The first proxy variable approach to [1] is a classical instrumental variables (IV) approach which analyzes y_{ic} in a linear probability model and estimates Eq. 1 using linear two-stage least-squares (IV 2SLS) (22-25). The second proxy variable approach to [1] analyzes y_{ic} in a binary-dependent variable model and estimates Eq. 2 using bivariate probit (BP) methods (26,27).

Plausible causal interpretation of the coefficient of interest, ρ in [1], relies on assumptions (26, 27). Assumption [A1] requires variation in industrial activity to be related to variation in proximity to major rail lines. Assumption [A2] requires that major rail lines induce variation in industrial activity rather than vice versa, conditional on the covariates. Assumption [A3] requires that major rail lines induce variation in cancer clusters through their influence on industrial activity alone and not through other channels, conditional on the covariates. To be precise, IV requires [A1], [A2], and [A3]. If one assumes that γ_{ic} and μ_{ic} are normally distributed in Eq. [2] above, the BP model only requires assumption [A1]. [A2] and [A3] are not strictly necessary.

We believe the assumptions necessary for casual identification are reasonable in this setting. [A1] is directly testable. A growing literature on China, Europe, the United States, and elsewhere supports the broader assertion that industrial activity follows major historic rail construction (46-49). [A2] is plausible, since the specific location of most major Chinese railways was determined by terrain, various war and foreign policy considerations, and straight-line (linear) connections between historic large cities of China (46-48). Results are robust to omitting large urban areas.

The exclusion restriction [A3] strikes us as reasonable conditional on the other covariates. Estimates from Eq. [1] are robust to the BP model, which does not necessarily require [A3]. Moreover, major railways may be unlikely to directly influence detected village-level cancer incidence or mortality via their own pollution emissions. Although an existing literature links railways to pollution or health risks, this literature focuses on highly local pollution within, adjacent to, or within a few kilometers of the railyard (50-52). Effects decline with distance on even this small scale (50). Similarly, meta-analyses of pollution near roads or other transportation corridors suggest that contaminants and health risks tend to decline with distance and become small within 10 km from the source (53-54). The average distance between a village and a major rail line in this study is 44.8 km.

Variation: We explore heterogeneity across industries by running regressions for each industry separately. We then analyze a simultaneous regression analysis to allow for possible correlations in locations of multiple types of polluting industries, such that we regress the probability of a village-level

spatial cancer cluster on the presence of cement enterprises, chemical enterprises, coking enterprises, and so on, simultaneously. We explore heterogeneity across size, geographic regions, and land use type using standard regression interactions.

Standard errors. We cluster standard errors at the county-level. This allows for spatial correlations between villages within a county.

DATA AVAILABILITY

The raw data can be obtained from the China Bureau of Statistics's 2010 Population Census and 2008 Second Economic Census, the PRC National Bureau of Surveying and Mapping, the China Ministry of Health 2013 National Central Cancer registry, and the China Data Center at the University of Michigan. The datasets that support the findings of this study are available from the corresponding authors upon request, along with code. Processed datasets will be de-identified.

Declarations

Additional information. Correspondence and requests for materials should be addressed to J.P.S. and H.Y.

Acknowledgments: We are grateful to Alan Barreca, Randall Lutter, and Chris Ruhm for helpful comments. H.Y. acknowledges funding from the Natural Science Foundation of China (No. 71974124, 71322305) and Shanghai Institute of International Finance and Economics.

Author Contributions: J.P.S., H.Y., and X.Z. designed research; H.Y. and X.Z. collected data; J.P.S. and X.Z. analyzed data; J.P.S., H.Y., and X.Z. wrote the paper. Author names presented in reverse alphabetical order.

Competing interests: The authors declare no competing financial interests.

References

1. Zheng, S., & Kahn, M. E. (2013). Understanding China's urban pollution dynamics. *J Econ Lit*, **51**(3), 731-772.
2. Li, X.*et al.* (2016). Urbanization and health in China, thinking at the national, local and individual levels. *Env Health*, **15**(1), 113.
3. Dockery, D. W. *et al.* (1993). An association between air pollution and mortality in six US cities. *NEJM*, **329**(24), 1753-1759.

4. Pope III, C. A. *et al.* (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA*, **287**(9), 1132-1141.
5. Chay, K. Y., & Greenstone, M. (2003). The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession. *Q J Econ*, **118**(3), 1121-1167.
6. Dominici, F., Sheppard, L., & Clyde, M. (2003). Health effects of air pollution: a statistical review. *Intl Stat Rev*, **71**(2), 243-276.
7. Zhang, P., et al. (2011). Long-term exposure to ambient air pollution and mortality due to cardiovascular disease and cerebrovascular disease in Shenyang, China. *PLoS One*, **6**(6), e20827.
8. Huang, W., et al. (2012). Seasonal variation of chemical species associated with short-term mortality effects of PM_{2.5} in Xi'an, a central city in China. *Am J epidemiol*, **175**(6), 556-566.
9. Chen, R. *et al.* (2012). Association of particulate air pollution with daily mortality the China air pollution and health effects study. *Am J epidemiol*, **175**(11), 1173-1181.
10. Chen, Y., Ebenstein, A., Greenstone, M., & Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy. *PNAS*, **110**(32), 12936-12941.
11. Pope III, C. A., & Dockery, D. W. (2013). Air pollution and life expectancy in China and beyond. *PNAS*, **110**(32), 12861-12862.
12. Tanaka, S. (2015). Environmental regulations on air pollution in China and their impact on infant mortality. *J Health Econ*, **42**, 90-103.
13. Yin, P. *et al.* (2017). Particulate air pollution and mortality in 38 of China's largest cities: time series analysis. *The BMJ*, 356, 667.
14. Tie, X., Wu, D., & Brasseur, G. (2009). Lung cancer mortality and exposure to atmospheric aerosol particles in Guangzhou, China. *Atmos Env*, **43**(14), 2375-2377.
15. Cao, J. *et al.* (2011). Association between long-term exposure to outdoor air pollution and mortality in China: a cohort study. *J Haz Mat*, **186**(2), 1594-1600.
16. Ebenstein, A. (2012). The consequences of industrialization: evidence from water pollution and digestive cancers in China. *Rev Econ Stat*, **94**(1), 186-201.
17. Chen, W., et al. (2017). Cancer incidence and mortality in China in 2013: an analysis based on urbanization level. *Chin J Cancer Res*, **29**(1), 1-10.
18. Ferlay, J., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Intl J of cancer*, **136**(5), E359-E386.

19. Zhou, M., et al. (2016). Cause-specific mortality for 240 causes in China during 1990–2013: a systematic subnational analysis for the Global Burden of Disease Study 2013. *Lancet*, **387** (10015), 251-272.
20. He, J., & Chen, W. Q. (2012). Chinese cancer registry annual report 2012. Beijing: Press of Military Medical Sciences, 68-71.
21. Holdaway, J. (2010). Environment and health in China: an introduction to an emerging research field. *J Contemporary China*, **19**(63), 1-22.
22. Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *J Econ Perspectives*, **15**(4), 69-85.
23. Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *Intl J epidemiology*, **29**(4), 722-729.
24. Newhouse, J. P., & McClellan, M. (1998). Econometrics in outcomes research: the use of instrumental variables. *Annual review of public health*, **19**(1), 17-34.
25. Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.
26. Heckman, J. J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**(4), 931-959.
27. Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
28. Liu, L. (2010). Made in China: cancer villages. *Environment: Science and Policy for Sustainable Development*, **52**(2), 8-21.
29. Gong, G., & Zhang, T. (2013). Temporal–spatial distribution changes of cancer village in China. *China Popul Resour Environ*, **23**, 156-164.
30. Shagun, S. (2016). Development and Disease: A Study of Cancer Villages in China. *Politikon*. 29, 189.
31. Lora-Wainwright, A. (2010). An Anthropology of 'Cancer Villages': villagers' perspectives and the politics of responsibility. *J Contemporary China*, **19**(63), 79-99.
32. Hu, Y. et al. (2018). A novel evolution tree for analyzing the global energy consumption structure. *Energy*, **147**, 1177-1187.
33. Wang, J. F. et al. (2012). Cities evolution tree and applications to predicting urban growth. *Popul. Environ.*, **33**(2-3), 186-201.

34. Wang, J. F., Zhang, T. L., & Fu, B. J. (2016). A measure of spatial stratified heterogeneity. *Ecol. Indic.*, **67**, 250-256.
35. Wartenberg, D. (2001). Investigating disease clusters: why, when and how? *J R Stat Soc: Series A*, **164(1)**, 13-22.
36. Goodman, M., Naiman, J. S., Goodman, D., & LaKind, J. S. (2012). Cancer clusters in the USA: What do the last twenty years of state and federal investigations tell us? *Crit rev toxicol*, **42(6)**, 474-490.
37. Kingsley, B. S., Schmeichel, K. L., & Rubin, C. H. (2007). An update on cancer cluster activities at the Centers for Disease Control and Prevention. *Env health perspect*, **115(1)**, 165-171.
38. Greenland, S., & Robins, J. (1994). Invited commentary: ecologic studies—biases, misconceptions, and counterexamples. *Am J epidemiol*, **139(8)**, 747-760.
39. Openshaw, S. (1983). The modifiable areal unit problem. Norwick: Geo Books.
40. Greenland, S. (1996). Basic methods for sensitivity analysis of biases. *Intl J epidemiology*, **25(6)**, 1107-1116.
41. Ezzati, Majid, et al. (2002) Selected major risk factors and global and regional burden of disease. *Lancet* **360.9343**: 1347-1360.
42. China Ministry of Environmental Protection. (2009). Category namelist of environmental regulation industry for listed companies. MEP.
43. Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied statistics*, **43(3)**, 429-467.
44. Rothman, K. J., Greenland, S., & Lash, T. L. (2008). Case-control studies. *Encyclopedia of Quantitative Risk Analysis and Assessment*, 1.
45. Sigman, H. (2002). International Spillovers and Water Quality in Rivers: Do Countries Free Ride? *Am Econ Rev*, **92(4)**, 1152-1159.
46. Atack, J., Bateman, F., Haines, M., & Margo, R. A. (2010). Did railroads induce or follow economic growth?. *Soc Sci Hist*, **34(02)**, 171-197.
47. Banerjee, A., Duflo, E., & Qian, N. (2012). On the road: Access to transportation infrastructure and economic growth in China (No. w17897). NBER.
48. Xue, X., Schmid, F., & Smith, R. A. (2002). An introduction to China's rail transport Part 1: History, present and future of China's railways. *Proceedings of the Institution of Mechanical Engineers, Part F: J of Rail and Rapid Transit*, **216(3)**, 153-163.

49. Baum-Snow, N., Brandt, L., Henderson, J. V., Turner, M. A., & Zhang, Q. (2017) Roads, Railroads and Decentralization of Chinese Cities. *Rev Econ Stat*, **99(3)**, 435-448.

50. Hand, R., Di, P., Servin, A., Hunsaker, L., & Suer, C. (2004). Roseville Rail Yard Study. State of California, Air Resources Board, Stationary Source Division.

51. Jaffe, D. A. *et al.* (2014). Diesel particulate matter emission factors and air quality implications from in-service rail in Washington State, USA. *Atmospheric Pollution Research*, **5(2)**, 344-351.

52. Cahill, T. A., Cahill, T. M., Barnes, D. E., Spada, N. J., & Miller, R. (2011). Inorganic and organic aerosols downwind of California's Roseville Railyard. *Aerosol Sci Tech*, **45(9)**, 1049-1059.

53. Jerrett, M. *et al.* (2005). A review and evaluation of intraurban air pollution exposure models. *Journal of Exposure Science and Environmental Epidemiology*, **15(2)**, 185.

54. Zhou, Y., & Levy, J. I. (2007). Factors influencing the spatial extent of mobile source air pollution impacts: a meta-analysis. *BMC public health*, **7(1)**, 89.

Tables

Table 1. Marginal effects from village-level regressions.

	(1)	(2)	(3)	(4)
(A) Fixed-Effects Logistic Regression	0.008	0.007	0.008	0.008
	p < .01	p < .01	p < .01	p < .01
(B) Fixed-Effects Linear Probability	0.008	0.007	0.008	0.008
	p < .01	p < .01	p < .01	p < .01
(C) IV 2SLS	0.042	0.052	0.057	0.059
	p < .08	p < .09	p < .08	p < .15
(D) IV Bivariate Probit	0.059	0.014	0.014	0.008
	p < .04	p < .04	p < .03	p < .26
Other industrial facility controls	-	YES	YES	YES
Other economic controls	-	-	YES	YES
Geo-political controls	-	-	-	YES

Each cell presents marginal effects for the relationship between a dependent variable defined as “village contains a documented cancer cluster” and an explanatory variable defined as “village contains one or more polluting industrial establishments.” Each row (A-D) represents a different empirical modeling approach. Row A presents marginal effects for logistic regressions that use fixed effects and only exploit within-county variation for statistical identification. Row B presents marginal effects for linear probability regressions that use fixed effects and only exploit within-county variation for statistical identification. Marginal effects from two-stage regressions that exploit variation in a railway proxy to help isolate causal effects are in Row C (Instrumental variables 2SLS) and Row D (Bivariate Probit). Each column (1-4) represents a regression with a different number of controls. Full results are in *Table S5*, *Table S6*, *Table S7a*, and *Table S7b*.

Figures

Figure 1

County-level relationships between industrial activity and age-standardized all cancer incidence and mortality. Panels depict scatter plots overlaid with fitted polynomial regression predictions and 95% confidence intervals. Each plotted data point represents a county-level cancer registry. Left panel response variables represent cumulative incidence rate to age 74 for all cancer. Right panel responses variables represent cumulative mortality rate to age 74 for all cancer. Explanatory variables represent the

natural log of value added from industrial production in hundred million CNY. The panels depict positive county-level associations between industrial activity and age-standardized incidence and mortality. Relationships tend to increase at a decreasing rate. Limited data drive imprecise estimates in the distributional tails.

Figure 2

County-level relationships between industrial activity and age-standardized cancer mortality: four most common cancers. Panels depict scatter plots overlaid with fitted polynomial regression predictions and 95% confidence intervals. Each plotted data point represents a county-level cancer registry. All response variables represent cumulative male mortality rate to age 74 for each listed cancer. Explanatory variables represent the natural log of value added from industrial production in hundred million CNY. The panels depict positive county-level associations between industrial activity and age-standardized mortality from lung, trachea, bronchus cancer; stomach cancer; and esophageal cancer. These relationships tend to increase at a decreasing rate. We detect no clear association between industrial activity and age-standardized mortality from liver cancer. Limited data drive imprecise estimates in the distributional tails. Results are similar for male incidence, female mortality, and female incidence (Figures S1, S2, S3).

Figure 3

“Cancer village” locations overlaid on choropleth of industrial activity quartiles. Darker colors represent greater county-level share of employment in industrial activity. Village-level cancer clusters are represented by red dots. White indicates no data; we do not analyze the sparsely populated and less developed far west and northwestern provinces. Cancer villages are, on average, located in counties with greater shares of employment in industrial activities. Counties in the 2nd, 3rd, and 4th quartiles of share of employment in industrial activities have 1.7, 1.9, and 2.8 times more reported spatial cancer clusters than counties in Q1.

Figure 4

Change in log odds (with 95% CIs) associated with the presence of industrial facilities in the village. Upper (darker) bars represent coefficient estimates (circles) and 95% confidence intervals from one regression per industry, i.e. regressing cancer cluster presence on the presence of cement facilities. Lower (lighter) bars represent coefficient estimates (diamonds) and 95% confidence intervals from one regression for all industries, i.e. regressing cancer cluster presence on the presence of cement facilities, chemical facilities, coking facilities, etc. Underlying coefficients (reported in Table S8) suggest that the odds that a village contains a spatial cancer cluster increase 2.62 times (95% CI: 1.64 – 4.17) if the

village contains one or more chemical facilities, 1.87 times (95% CI: 1.02 – 3.27) if the village contains one or more dye facilities, and 7.44 times (95% CI: 3.48 – 15.87) if the village contains one or more leather facilities. The key take-away is that the log odds that a village contains a spatial cancer cluster increase significantly if the village contains a chemical, dye, or tanning facility.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterialfor.docx](#)