



Is natural experiment a cure? Re-examining the long-term health effects of China's 1959–1961 famine



Hongwei Xu ^{a, *}, Lydia Li ^b, Zhenmei Zhang ^c, Jinyu Liu ^d

^a Institute for Social Research, University of Michigan, 426 Thompson St, ISR 2459, Ann Arbor, MI 48104-2321, USA

^b School of Social Work, University of Michigan, USA

^c Department of Sociology, Michigan State University, USA

^d School of Social Work, Columbia University, USA

ARTICLE INFO

Article history:

Received 13 July 2015

Received in revised form

20 October 2015

Accepted 19 November 2015

Available online 2 December 2015

Keywords:

China

Fetal origins hypothesis

Famine

Biomarker

Cardiovascular and metabolic diseases

Spatial

ABSTRACT

The fetal origins hypothesis posits that adverse prenatal exposures, particularly malnutrition, increase the risk of poor adult health. Studies using famine as a natural experiment to test the fetal origins hypothesis present conflicting findings, partly because of data limitations and modeling flaws. Capitalizing on the biomarker data and prefecture-level geographic information from the 2011 China Health and Retirement Longitudinal Study, this study estimates the effects of prenatal exposure to China's 1959–61 famine on later-life risks of cardiovascular and metabolic diseases. Our analysis addresses the problems of measurement error and intrinsic cohort differences that challenge prior studies. We use provincial and prefecture-level geographic variations in famine severity, a proxy for prenatal malnutrition, for model identification. We construct instrumental variables from geocoded newspaper archive data to adjust for measurement error in famine exposure. We find that estimates of the famine effects are highly sensitive to the choices of health indicators, measures of famine severity, and regression model specifications. Overall, we find little evidence supporting the fetal origins hypothesis. In fact, it appears that prenatal exposure to famine reduces later-life disease risks in certain cases. We interpret this finding as evidence of mortality selection among the famine survivors at work. We conclude that using famine as a natural experiment in itself does not guarantee correct statistical inference about the long-term health impacts of prenatal malnutrition when other analytical challenges remain unresolved.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The controversial fetal origins hypothesis (Barker, 1990, 1995a,b) conjectures that prenatal exposure to an adverse environment, in particular to malnutrition, may “program” the fetus to develop particular metabolic characteristics, likely through environmental effects on the epigenome. Such developmental changes may persist over the life course and increase risks of cardiovascular and metabolic diseases in middle and later ages. The life course perspective of the fetal origins hypothesis, as well as its recent variants such as the developmental origins hypothesis (Bateson et al., 2004; Gluckman et al., 2008, 2005), has made it an attractive theoretical and analytical framework to researchers. However, supportive empirical evidence is uneven, particularly in the area of the long-term health effects of prenatal malnutrition.

Early efforts to test the fetal origins hypothesis using observational data often failed to control for other prenatal confounders (Barker, 1995a,b; Barker and Osmond, 1986). These studies also rely heavily on low birth weight as a crude proxy for fetal malnutrition, even though maternal malnutrition during gestation may induce later-life disease without affecting birth weight (Roseboom et al., 2001). More importantly, the observed association between birth weight and health outcomes in later life could reflect many unobserved joint determinants such as genetic, socioeconomic, and environmental factors. Failing to control for such factors can introduce omitted variable bias and preclude causal inference (Almond and Currie, 2011; Paneth and Susser, 1995; Portrait et al., 2011; Song, 2013a).

To adjust for potential confounders more effectively, researchers have increasingly used exposure to prenatal famine as a natural experiment in studies of long-term health effects of prenatal malnutrition (for a recent review, see Lumey et al., 2011). Well-known famine examples include 19th-century crop failures in Sweden and Finland, the Siege of Leningrad of 1941–44, the Dutch

* Corresponding author.

E-mail address: xuhongw@umich.edu (H. Xu).

Hunger Winter of 1944–45, the Chinese Great Leap Forward famine of 1959–61, and the Bangladesh famine of 1974. Because exposure to famine is beyond the control of most individuals, regardless of their genetic traits, personality, or socioeconomic status, the process governing an individual's prenatal exposure to famine-induced malnutrition is arguably exogenous and resembles random assignment. Therefore, casual effects of prenatal malnutrition on adult health can be inferred by comparing two similar sub-populations that differ in the famine exposure.

However, famine studies have produced conflicting evidence about the effects of prenatal exposure on adult health. For example, in comparisons of cohorts born before or after a famine, prenatal malnutrition has been related to higher older age mortality rates in the 1846–47 Dutch Potato Famine (Lindeboom et al., 2010), but not in the 1866–68 Finnish famine (Kannisto et al., 1997), the Dutch Hunger Winter (Painter et al., 2005), or the 1959–61 Chinese famine (Song, 2009). Even studies of the same famine have produced mixed findings on similar health outcomes. For example, in studies focused on the Siege of Leningrad, Koupil et al. (2007) reported significantly increased risks of cardiovascular diseases (CVD) and mortality in adult life among those exposed to famine in childhood, while Stanner et al. (1997) found no elevated risk for CVD among individuals exposed to famine either in utero or infancy. Similarly, in studies of the Dutch Hunger Winter, Ravelli et al. (1998) found an association between prenatal famine exposure and insulin resistance, but de Rooij et al. (2006) failed to do so. In their review of 30 studies of the relationship between prenatal famine and adult health, Lumey et al. (2011) found a lack of consistent associations for most measures of adult health, except for adult body size, diabetes, and schizophrenia.

These inconclusive findings may reflect incomparable analytic strategies among these studies. For instance, while the most common estimation approach is simple cohort difference (SCD) in health between individuals born during versus after/before a famine, some studies use an approach known as difference-in-differences (DID) to exploit spatial variation in famine exposure in addition to temporal (cohort) variation. The SCD approach hinges on stronger assumptions and presumably yields less robust results than the DID approach (see details below).

In this study, we seek to advance the literature by examining the effects of prenatal and infancy exposure to the 1959–61 Chinese famine, also known as the Great Leap Forward (GLF) famine, on adult cardiovascular and metabolic disease risks. We choose the GLF famine for a case study because of its greater magnitude relative to other famines in terms of duration (three years), geographic scope (pandemic as opposed to endemic), and level of damage (16.5–30 million excess deaths with a mortality rate of over 3.0%) (Song et al., 2009; Susser and St Clair, 2013). Our analysis tackles two empirical challenges of prior studies: (1) measurement errors related to famine exposure and adult health outcomes, and (2) inappropriate estimation strategies. We overcome the first challenge by drawing on biomarker data from the 2011 wave of the China Health and Retirement Longitudinal Study (CHARLS), matching CHARLS respondents to a variety of historical famine data, including fine-grained geographical variation in famine severity. We address the second challenge by assessing results obtained from four discrete estimation strategies – SCD, deviation from cohort trend (DCT), DID, and instrumental variable (IV) – which vary in their ability to address measurement error and to make causal inference. We provide a sketch of the GLF famine in the next section before reviewing the related literature.

2. Background of the GLF famine

From 1958 to 1961, the Communist Party of China (CPC)

launched a massive campaign, known as the Great Leap Forward (GLF), mobilizing the entire country to adopt radical economic and social policies to rapidly transform China from a predominantly agrarian society to an industrialized socialist economy through a Soviet-style high investment in heavy industry, supported by agricultural collectivization. To appeal to their zealous superiors in the CPC, and avoid being labeled “anti-revolution,” local cadres began to make fictitious high-yield agricultural reports of grain output to the *People's Daily*, the CPC's official newspaper (Kung and Chen, 2011). The first such instance was on June 8, 1958, when the front-page headline of the *People's Daily* reported that a People's Commune in Henan Province achieved a significantly higher than average wheat yield of 2105 catties per *mu* (1 catty = 1/2 kg; 1 *mu* = 1/6 acre). This exaggeration was topped the next day when the *People's Daily* reported that another commune in Hubei Province harvested an average of 2357 catties of wheat per *mu*. Quickly, other regions throughout the country began to over-report grain yields. From June to September, more than 800 false reports (based on our calculation) of abnormally high grain yields were published in the *People's Daily*. At the end of 1958, the national grain production was reported to be 375 million metric tons (MMT), roughly double the yield of 1957. Subsequent verification in 1961, however, placed the actual 1958 yield at 200 MMT (Bernstein, 1984).

These grain yield exaggerations led to food shortages in several ways. Top political leaders, believing China was facing a grain surplus rather than a shortfall (Bernstein, 1984; Yang, 1996), raised compulsory procurement levels — amounts that collectives must deliver to the state (Ashton et al., 1984; Bernstein, 1984). The total grain procurement in the 1958 grain year was 22.3 percent higher than that in 1957 (Yang, 1996). In addition, new policies were implemented to divert labor and resources from agriculture to fruitless projects such as the so-called backyard furnace movement in later 1958 and to reduce sown acreage in 1959 (Ashton et al., 1984). Together, these changes resulted in sharp declines in grain production, and rural villages suffered from severe food shortage after compulsory procurement to support urban and industrial growth. Coupled with other manmade and natural devastating factors, the resulting GLF famine of 1959–61 caused an estimated 16.5 to 30 million excess deaths, depending on the data sources, underlying assumptions, and methods of estimation employed (Ashton et al., 1984; Banister, 1987; Coale, 1984; Peng, 1987; Yao, 1999).

3. Studies of the GLF famine and limitations

In prior studies of the long-term effects of the GLF famine on physical health, the most frequently examined health outcome is adulthood nutritional status derived from anthropometric measures. A key data source in this research is the China Health and Nutrition Survey (CHNS), a longitudinal study of Chinese households in nine provinces since 1989. For example, using data from the 1991 wave of the CHNS, Chen and Zhou (2007) found that those born in 1959 and 1960 attained shorter height in adulthood than those born between 1963 and 1967 (i.e., after the famine). The negative association between prenatal famine exposure and adult height was also confirmed in the 1989 wave (Meng and Qian, 2009) and the pooled 1989–93 waves (Fung and Ha, 2010). However, using the 1989–97 CHNS data, Gørgens et al. (2012) reported that the young female famine cohort (born 1957–61) grew about 2 cm taller in adulthood than did the control cohorts (born 1938–47 or 1962–71) – a finding attributed to a positive selection among famine survivors. Findings on BMI-related measures are even more inconclusive. Studies comparing adults from the famine and post-famine birth cohorts report the famine cohort has no significant difference in BMI (Meng and Qian, 2009), borderline significantly

higher (at 0.1 level) BMI (Fung and Ha, 2010), or significantly higher rates of 'overweight' BMI ($\geq 25 \text{ kg/m}^2$) (Luo et al., 2006).

In other research, Huang et al. (2010b) analyzed a cross-sectional sample of women in three provinces and found reduced height in the famine cohort, increased BMI and risk of hypertension in the pre-famine cohort, compared to the post-famine cohort. In a series of studies using the biomarker data from the 2002 China National Nutrition and Health Survey (CNNHS), Li et al. (2010; 2011a; 2011b) reported that only in severely affected famine areas did the 1959–61 famine cohort have significantly higher systolic and diastolic blood pressures (BPs) and higher risks of hyperglycemia and metabolic syndrome than did the 1962–64 post-famine cohort.

Using retrospective data on family deaths from the 1988 National Survey of Fertility and Contraception, Song (2009) compared mortality between the famine and non-famine cohorts, finding no significant difference after adjusting for the cross-cohort temporal trend of mortality rate. In another study using the same data, Song (2010) found a higher mortality rate among the famine cohort up to ages 11 and 12, after which the non-famine cohort exhibited a higher mortality rate up to age 22. But given their vulnerability to recall error, and without a vital registration system for confirmation, such retrospective household survey data on family member death dates can be unreliable.

Other GLF famine studies have relied on subjective health indicators. For example, Fan and Qian (2015) used data from the 2005 Chinese General Social Survey to create a composite measure of health based on multiple measures of self-rated health (e.g., overall health, physical functioning, disability, body pain, vitality, emotion, and mental health). They found significantly worse general health in women born during the famine than in those born after the famine.

Several limitations are notable in famine studies in China. First, only a handful have directly measured risks of cardiovascular and metabolic diseases—health outcomes that have clear physiological mechanisms related to prenatal malnutrition, and that are specified in the fetal origins hypothesis. Anthropometric measures (height, weight, and BMI) are indicative of nutritional status, but are not biomarkers of cardiovascular and metabolic diseases. Some GLF famine studies have analyzed hypertension, but the empirical findings vary, with a significant famine effect reported in one study (Li et al., 2011a) and a null effect in others (Huang et al., 2010b; Meng and Qian, 2009). To the best of our knowledge, the only studies that directly examine metabolic disease effects are by Li and colleagues (2010; 2011a; 2011b), who used biomarker data collected in the 2002 CNNHS. Unfortunately, the CNNHS data are not publicly available, preventing expansion or replication of their studies.

A second limitation of past studies is that many relied on comparisons between the famine cohort (born or conceived during the famine years) and pre- or post-famine cohorts (born before or after the famine years) to infer famine effects. This SCD approach cannot rule out intrinsic cohort differences due to factors other than famine, especially when the famine and non-famine cohorts are broadly defined on age (Almond et al., 2010; Chen and Zhou, 2007; Song, 2013a). A refined alternative is to examine deviations of health outcomes, due to famine exposure, from smooth cohort trends (Almond, 2006; Almond et al., 2010). This approach distinguishes famine-induced cohort difference from cohort difference due to other factors (e.g., economic growth over time), but requires additional model specification about the secular cohort trend. Other studies (Chen and Zhou, 2007; Fan and Qian, 2015; Huang et al., 2010b; Li et al., 2011b; Luo et al., 2006) exploit within-cohort geographic variations in famine severity, in addition to between-cohort differences, to calculate DID estimates of famine

effects. The DID approach may be more robust than the SCD approach because it permits intrinsic cohort differences as long as such differences do not vary by severity of famine exposure (Song, 2013a).

However, the DID strategy suffers from several flaws in itself. First, because researchers generally collect salient data after the famine, famine severity is often approximated by a period measure of total famine-caused excess mortality, regardless of birth cohort, during the famine years (Almond et al., 2010; Chen and Zhou, 2007; Luo et al., 2006), or a cohort measure of famine-induced cohort size shrinkage (Huang et al., 2010a, 2010b; Meng and Qian, 2009). These proxies are susceptible to measurement error, which in turn can bias the estimate of true famine effect. In particular, estimates of famine severity can be seriously biased if other correlated factors that likely contributed to excess mortality or cohort size shrinkage — such as degraded medical or social services — are omitted in regression models (Tan et al., 2014). Second, the selection effect could dominate the health damaging effect of famine where high mortality levels reduced populations to only the fittest individuals. In fact, several studies reviewed above have recognized the fitness selection effect with respect to famine survivors in terms of such outcomes as height (Gørgens et al., 2012), BMI (Huang et al., 2010b), mortality (Song, 2010), and schizophrenia (Song et al., 2009).

In addition, most studies of the GLF famine using the DID approach have relied on variations in famine severity at the provincial level. When the study area covers only a few provinces, researchers must include cohorts born many years after the famine in the control group to make it large enough to analyze between-group variations (Chen and Zhou, 2007; Fan and Qian, 2015; Fung and Ha, 2010), thereby further contaminating the famine effect with intrinsic cohort effect. In addition, because sub-provincial-level famine intensity may vary widely, even within the hardest-hit province, the assumption of within-province homogeneity can result in biased estimates of famine effects. Constrained by data confidentiality restrictions, however, only a handful of studies have attempted to investigate sub-provincial variations by matching survey respondents to county-level cohort size shrinkage (Huang et al., 2010b, 2013; Meng and Qian, 2009). Even these studies, however, have shortcomings. Meng and Qian (2009) treated county of residence at the time of survey as one's birth county among respondents reporting no migration in the prior five years — a problematic assumption after domestic migration began to surge in the early 1990s. And Huang et al.'s (2010b, 2013) analyses were restricted to a limited number of counties (24 or 35) concentrated in three or four provinces, thereby broaching problems of insufficient regional variation.

In short, despite conventional assumptions about the long-term negative health effects of prenatal famine exposure, empirical evidence remains scarce given challenges regarding data limitations, measurement errors, and estimation drawbacks. In the next section, we describe our measurement and modeling strategies to address these challenges.

4. Data and method

4.1. Data

This study draws on individual-level data from the 2011 baseline survey of the China Health and Retirement Longitudinal Study (CHARLS), a nationally representative longitudinal survey of adults aged 45 and older and their spouses, if available. CHARLS sampled 17,708 residents from 150 counties across 28 provinces in China, with a response rate of 80.5% (Zhao et al., 2014b), and collected anthropometric and physical performance measures from 78.9% of the sample (13,978 respondents) and fasting blood samples with

valid test results from 60.0% (10,627). Details on sampling procedures, field operations, and blood collection and tests are described elsewhere (Zhao et al., 2014a, 2014b).

The CHARLS data has several appealing features for our analysis. First, it consists of a nationally representative sample with sufficient provincial-level variations in famine severity, making findings generalizable. Second, both the survey and biomarker data are publicly available, allowing other researchers to replicate our study. Third, respondents' places of residence at the time of survey are known at the prefecture-level, allowing us to match adult health to sub-provincial famine severity of birthplace for respondents who did not move from their birth prefecture. On average, each province consists of about 10 prefectures and each prefecture consists of about nine counties. Unfortunately, CHARLS has not released the birth prefecture for those who moved away.

To replicate existing studies as closely as possible, we adopted the most common cohort definitions: the famine cohort consists of those born during the famine period of 1959–61; the pre-famine cohort consists of those born 1956–58; and the post-famine cohort consists of those born 1962–1964. Some famine cohort members (born in the early 1959) were actually conceived in 1958 and thus did not experience prenatal famine exposure in full term, while post-famine cohort members (born in the early 1962) conceived in 1961 experienced some prenatal famine exposure. Additional analyses that narrowed the birth years of the famine cohort to 1960–1961 and of the post-famine cohort to 1963–1964 did not alter our main findings.

Our analytical sample is restricted to 4812 pre-famine, famine, and post-famine cohort members who were born and lived in rural

areas during childhood and hence were hit harder by the famine than their urban peers, who were protected through state-controlled food rationing. Among them, we dropped 141 (2.9%) pre-famine, 109 (2.3%) famine, and 191 (4.0%) post-famine cohort members whose birth prefecture was unknown because they had moved since birth. We further dropped 64 respondents with missing covariates. Chi-square tests show no significant cohort difference in missing data on birth prefecture or the other covariates. Every remaining respondent had at least one valid biomarker and, to maximize statistical power, we allowed the analytical sample size to vary depending on the number of valid responses for each health outcome. As a result, the sample sizes ranged from 1003 to 1308 for the pre-famine cohort, 664 to 884 for the famine cohort, and 1110 to 1550 for the post-famine cohort.

4.2. Dependent variables

To overcome the limitations of previous studies in assessing the cardiovascular and metabolic disease risks of famine exposure, we used the CHARLS biomarker data to construct nine dichotomous indicators of high disease risks in three domains: *cardiovascular* (diastolic BP, systolic BP, and resting pulse); *dyslipidemia* (HDL cholesterol, LDL cholesterol, total cholesterol, and triglyceride); and *diabetes* (glucose and HbA1c), using clinical cut-points (shown in Table 1; also see Zhao et al., 2014a).

In addition, we combined these nine biomarkers with respondents' self-reported disease history to create another four dichotomous disease indicators: *hypertension* (high-risk diastolic or systolic BP, or ever diagnosed with or treated for hypertension);

Table 1
Cohort-stratified distributions of chronic disease risks and control variables in rural Chinese: CHARLS-2011.

High-risk cut-points		Pre-famine (1956–58)	Famine (1959–61)	Post-famine (1962–64; reference)
		% (N)	% (N)	% (N)
Biomarkers				
<i>Cardiovascular</i>				
Diastolic BP	≥90 mmHg	16.0 (1179)	17.5 (793)*	14.3 (1394)
Systolic BP	≥140 mmHg	25.0 (1179)***	21.8 (793)*	17.4 (1394)
Resting pulse	>100 beats/minute	11.7 (1308)	11.3 (884)	11.4 (1550)
<i>Dyslipidemia</i>				
HDL cholesterol	<40 mg/dL in men; <50 mg/dL in women	39.8 (1012)	42.3 (673)	41.8 (1116)
LDL cholesterol	>160 mg/dL	11.1 (1010)**	10.4 (671)**	6.9 (1116)
Total cholesterol	≥240 mg/dL	11.9 (1010)*	9.5 (673)	9.1 (1116)
Triglyceride	≥150 mg/dL	15.5 (1012)	17.2 (673)	15.7 (1116)
<i>Diabetes</i>				
Glucose	≥126 mg/dL	12.7 (1009)*	12.9 (672)*	9.9 (1114)
HbA1c	≥6.5%	4.3 (1021)	4.1 (676)	3.6 (1126)
Biomarkers & self-reports				
Hypertension ^a		34.9 (1179)***	32.3 (793)*	27.7 (1394)
Heart problem ^b		18.7 (1308)	17.9 (884)	17.4 (1550)
Dyslipidemia ^c		52.8 (1009)	55.1 (671)	52.2 (1116)
Diabetes ^d		15.7 (1003)†	16.4 (664)*	12.7 (1110)
Anthropometries				
Height (cm; mean)		159.6 (1167)	159.6 (780)	160.0 (1387)
BMI (mean)		23.5 (1163)***	24.0 (779)	24.2 (1383)
Overweight	BMI ≥25	31.6 (1163)**	35.2 (779)	37.1 (1383)
Abdominal obesity	Waist circumference >90 cm in men; >80 cm in women	44.8 (1173)**	48.7 (785)	50.4 (1391)
Control variables				
Male				
Birth quarter				
Quarter 1		28.9 (1326)***	27.9 (887)**	21.9 (1565)
Quarter 2		23.8 (1326)†	24.5 (887)†	21.2 (1565)
Quarter 3		23.5 (1326)*	22.3 (887)**	27.1 (1565)
Quarter 4		23.8 (1326)***	25.4 (887)*	29.8 (1565)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ for pairwise cohort differences using the post-famine cohort as the reference.

^a High diastolic or systolic blood pressure, doctor-diagnosed or on treatment for hypertension.

^b High resting pulse rate, doctor-diagnosed or on treatment for any heart problem.

^c Low HDL cholesterol, high LDL cholesterol, high total cholesterol, high triglycerides, doctor-diagnosed or on treatment for dyslipidemia.

^d High glucose, high HbA1c, doctor-diagnosed or on treatment for diabetes.

heart problem (high-risk resting pulse, or ever diagnosed with or treated for any heart problem); *dyslipidemia* (high-risk HDL/LDL/total cholesterol or triglyceride levels, or diagnosed with or treated for dyslipidemia); and *diabetes* (high-risk glucose or HbA1c, or ever diagnosed with or treated for dyslipidemia). As a comparison to prior studies, we also considered several anthropometric measures typically used in the literature, including height, BMI, overweight (BMI ≥ 25 kg/m²), and abdominal obesity (waist circumference >90 cm in men or >80 cm in women).

4.3. Famine severity

We employed two measures of famine severity from the literature on the GLF famine. The first is a period measure of total famine-caused excess deaths, regardless of cohort, during the famine years (Almond et al., 2010; Chen and Zhou, 2007; Fan and Qian, 2015; Luo et al., 2006). The excess death rate (EDR) was calculated as the difference between mortality in famine years (1959–61) and the average of death rates in the three years before the famine (1956–58). Following some previous studies (Chen and Zhou, 2007; Fan and Qian, 2015; Huang et al., 2013; Luo et al., 2006), we used the provincial-level EDR constructed by Lin and Yang (2000). No reliability statistics of the EDR was reported by Lin and Yang, but using the widely adopted EDR allows us to compare our results against previous studies that used the same measure, and to explore the origins of potential discrepancies. However, this measure suffers two shortcomings. First, the death rates are published by China's State Statistical Bureau and hence subject to potential data distortion. Second, they are provincial mortality rates, which make sub-provincial analyses impossible.

As an alternative, we derived prefecture-level cohort size shrinkage indices (CSSI) from the publicly available 1% sample of the 1990 China Population Census (<https://international.ipums.org/international/>). Let $N_{nonfamine}^i$ denote the average cohort size of those born during the three years preceding the famine (1956–58) and the three years after the famine (1962–64) in the i th prefecture, and N_{famine}^i denote the average cohort size of those born during the three famine years (1959–61). The CSSI for the i th prefecture is calculated as a ratio:

$$CSSI_i = \frac{N_{nonfamine}^i - N_{famine}^i}{N_{nonfamine}^i} \quad (1)$$

where a larger value indicates a greater reduction in cohort size due to reduced fertility and increased infant mortality, both presumably induced by the GLF famine (Huang et al., 2010a, 2010b; Meng and Qian, 2009). We matched provincial EDRs and prefecture-level CSSIs to the CHARLS respondents based on their self-reported birth prefecture. Both EDR and CSSI can more or less capture famine severity under the assumptions of accurate census data on fertility and mortality, stable secular trends in fertility and mortality in the counterfactual absence of the famine, and strictly restricted migration. When one or more assumptions are violated, measurement error will propagate and the inference of famine effect will be biased.

4.4. Instrumental variables

As mentioned above, it is very hard to accurately measure individual-level famine exposure experienced decades ago. Instead, we have to rely on proxies of famine severity such as EDR and CSSI at certain aggregate levels (e.g., provinces and prefectures) and assume homogeneous famine exposure among individuals from the same area. To address the challenges of measuring famine

exposure, we adopted an IV approach. An IV is an exogenous variable that correlates with the endogenous independent variable (i.e., famine severity) but not with the error term (i.e., the IV affects the dependent variable only indirectly through its effect on the endogenous independent variable being instrumented). A valid IV approach allows consistent estimation even in the presence of measurement error in the treatment variable (i.e., famine exposure) by effectively mimicking random assignments of respondents into the treatment and control groups in cross-sectional data (Angrist and Krueger, 2001).

Our IVs were constructed from newspaper reports of exaggerations of grain yields by county officials in each prefecture as published in the 1958 archives of the *People's Daily*. As described above, local cadres demonstrated their loyalty to the central state through enthusiastic endorsement of the GLF movement in various forms, one of which was to falsely claim unprecedentedly high grain yields resulting from the agricultural measures advocated in the movement such as deep plowing, intensive seeding, and heavy fertilizing. Political loyalty was rewarded with career advancement and associated increases in salary, occupation prestige, authority, and privileged access to bureaucratically controlled goods (Goldstein, 1991). Unfortunately, the falsified harvest led to excessive compulsory grain procurement, reduced sown acreage in 1959, and diverted labor and resources from agriculture to fruitless projects (Ashton et al., 1984), all of which contributed to the subsequent famine. In other words, we should expect the frequency of exaggerating grain yields in a given prefecture to be positively related to famine severity in that prefecture, satisfying the IV relevance requirement. On the other hand, local cadres gradually stopped falsifying grain yields towards the end of 1958 as the harvest season ended. This devastating practice had been completely abandoned, along with other radical measures, by 1961 when the central state suspended the GLF movement as a whole. Therefore, it is unlikely that exaggerations of grain yields in 1958 could directly affect later-life health of the famine cohort other than through its impact on famine severity, thereby satisfying the exclusion restriction for an IV.

We geocoded to county level a total of 558 exaggerations of grain yields, defined as a reported grain yield of 1000 catties per *mu* (or 3000 kg per acre) or more (Kung and Chen, 2011), as reported in the *People's Daily* from June to September, 1958 – the most intensive period of agricultural falsification. Due to the lack of county-level geographic information in the CHARLS data, the two IVs employed in this study are aggregated measures at the prefecture level. The first IV is the total number of exaggerations summed over all the counties in each prefecture, reflecting the famine severity. The second IV is the total number of counties that exaggerated grain yields in each prefecture, capturing the geographic coverage of agricultural falsification. These two IVs together form a measure of famine intensity and coverage across prefectures.

4.5. Control variables

We controlled for gender and birth quarter (to adjust for season effect of birth on mortality, see for example Almond, 2006), as well as provincial fixed effects whenever suitable (i.e., model convergence is not a problem).

4.6. Statistical models

We employed four estimators to infer the health effects of the GLF famine. The SCD estimator ignores geographic variations in famine severity and infers the famine effects from between-cohort differences in health outcomes. To simplify the discussion, let's ignore the pre-famine cohort for now and define $C = 1$ for the

famine cohort and $C = 0$ for the post-famine cohort. Let y_i denote a health outcome for i th individual, a SCD estimator can be obtained from:

$$y_i = \beta_0 + \beta_1 C_i + \beta_2 X_i + \varepsilon_i \tag{2}$$

where β_1 represents the famine effect, X_i denotes the control variables, and ε_i is a random error. The DCT estimator adds a polynomial cohort trend:

$$y_i = \beta_0 + \beta_1 C_i + \beta_2 YOB_i + \beta_3 YOB_i^2 + \beta_4 YOB_i^3 + \beta_5 X_i + \varepsilon_i \tag{3}$$

where YOB denotes birth year and we follow Almond et al.'s (2010) work to allow a flexible, cubic cohort trend. Under the assumption that cohort effects on health tend to be smooth, deviations of health outcomes from the cohort trend for the famine cohort, as reflected by β_1 , indicate famine effects (Almond, 2006).

The DID estimator exploits geographic variations in famine severity in addition to cohort variations. Let S_j denote famine severity in j th prefecture (or province), a regression-based DID estimator can be obtained from:

$$y_{ij} = \beta_0 + \beta_1 C_{ij} + \beta_2 S_j + \beta_3 C_{ij} \times S_j + \beta_4 X_{ij} + \varepsilon_i \tag{4}$$

where β_3 , the coefficient of the interaction between cohort and regional famine severity, is the DID estimate of the famine effect.

Unlike the SCD, DCT, and DID, the IV estimator does not hinge on between-cohort variations. It also alleviates the problems of measurement error by instrumenting the treatment variable – famine severity. Let Z_{1j} and Z_{2j} denote the two IVs in j th prefecture, a two-stage least squares (2SLS) IV estimator can be obtained from two equations:

$$y_{ij} = \beta_0 + \beta_1 S_j + \beta_2 X_{ij} + \varepsilon_i \tag{5}$$

$$S_j = \alpha_0 + \alpha_1 Z_{1j} + \alpha_2 Z_{2j} + u_j \tag{6}$$

These estimators require different assumptions to achieve consistent estimates of the famine effects. The SCD approach assumes no other cohort differences in health exist except the cohort difference in exposure to the famine. This assumption may not hold when there are improvements in nutrition and health across cohorts due to economic growth over time. The DCT approach adjusts for intrinsic cohort differences due to non-famine factors by fitting a nonlinear cohort trend, but the chosen functional form can be misspecified. The DID approach does not require an explicit specification of the cohort trend, but it assumes the intrinsic cohort differences to be the same across areas of different levels of famine severity. It also implicitly assumes that variations in famine exposure are accurately approximated by the measured famine severity (i.e., S_j). The IV approach addresses the measurement error in famine exposure, but it requires the assumptions of instrument relevance and exclusion restriction to be satisfied. Throughout the regression analysis, we calculated p -values based on robust standard errors that adjust for the potential correlation of observations clustered within the same prefectures.

5. Results

5.1. Descriptive statistics

Table 1 reports cohort-specific prevalence of disease risks and nutritional status, as well as Chi-square and t-tests of pairwise cohort differences using the post-famine cohort as the reference group. Comparisons between the famine and post-famine cohorts

on biomarkers alone and on biomarkers combined with self-reports are consistent with the fetal origins hypothesis. The famine cohort had significantly higher proportions of at-risk members with respect to diastolic and systolic BP, LDL cholesterol, glucose, and combined indicators of hypertension and diabetes compared to the post-famine cohort. Similar patterns hold for the pre-famine cohort in comparison with the post-famine cohort. Comparisons on the anthropometric measures, however, yielded no significant differences between the famine and post-famine cohorts, and less risk for the pre-famine than the post-famine cohort in terms of the prevalence of high-BMI, overweight, and abdominally obese individuals.

Table 1 also shows frequency distributions of the control variables. The pre-famine cohort consisted of more male survivors (52.3%) than the famine (46.4%) and post-famine (46.2%) cohorts. In terms of season of birth, both the pre-famine and famine cohorts were more likely to be born in the first quarter and less likely to be born in the third and last quarters than the post-famine cohorts.

Table 2 presents descriptive statistics for measures of famine severity and IVs. At the provincial level, both excessive death rate (EDR) and cohort size shrinkage index (CSSI) exhibit sufficient spatial variations with large standard deviations and ranges. Figs. 1–3 depict prefecture-level spatial variations in CSSI and the two IVs – number of counties with grain yield exaggerations and number of grain yield exaggerations – respectively. Fig. 1 shows notable regional clusterings of high CSSI, corresponding to severe famine both within and across provincial boundaries. Mixed colors in any given province also suggest that within-province variation in CSSI was quite common even in the least (green (in the web version)) or hardest (red (in the web version)) hit provinces. The spatial distributions of the two IVs (Figs. 2 and 3) overlap partially with that of the CSSI (Fig. 1), implying that variation in the exaggeration of grain yields account for part of the variation in famine severity, and hence CSSI.

5.2. Regression results

To save space, Table 3 shows only the main coefficient estimates of interest – that is, the estimated famine effects on health as indexed by biomarkers (see Appendix Table A1 for examples of full model estimates). The SCD estimates provide evidence supporting the fetal origins hypothesis. Compared to the post-famine cohort, both pre-famine and famine cohorts were at significantly higher risk of poor health with respect to systolic BP, LDL cholesterol, and glucose. The pre-famine cohort also had higher total cholesterol. The DCT estimates reveal marginally significant risks of high resting pulse for both pre-famine and famine cohorts, but no difference in other outcomes. By contrast, the DID estimates are either insignificant or negatively significant (most notably for LDL and total cholesterol), regardless of how famine severity was measured (EDR or CSSI) or at what geographic scale (province or prefecture). In other words, the famine cohort living in areas of greater famine

Table 2
Measures of famine severity and instrumental variables during the 1959–1961 Chinese famine.

	Mean	SD	Min	Max	N
Provincial level					
EDR (unit: 0.1%)	6.4	6.9	0.1	28.6	27
CSSI (unit: 1%)	37.1	12.6	19.0	63.0	27
Prefecture level					
CSSI (unit: 1%)	40.8	13.8	9.9	79.1	119
N of exaggerations	3.7	6.1	0.0	41.0	119
N of counties exaggerating grain yields	1.8	2.1	0.0	14.0	119

Note: EDR = excess death rate and is from Lin and Yang (2000); CSSI = cohort size shrinkage index and is from the 1% sample of China's 1990 population census.

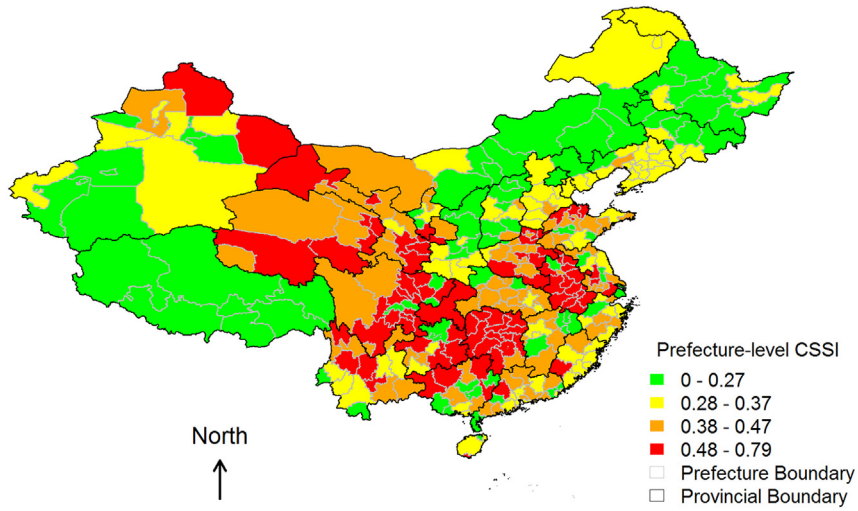


Fig. 1. Prefecture-level cohort size shrinkage index (CSSI).

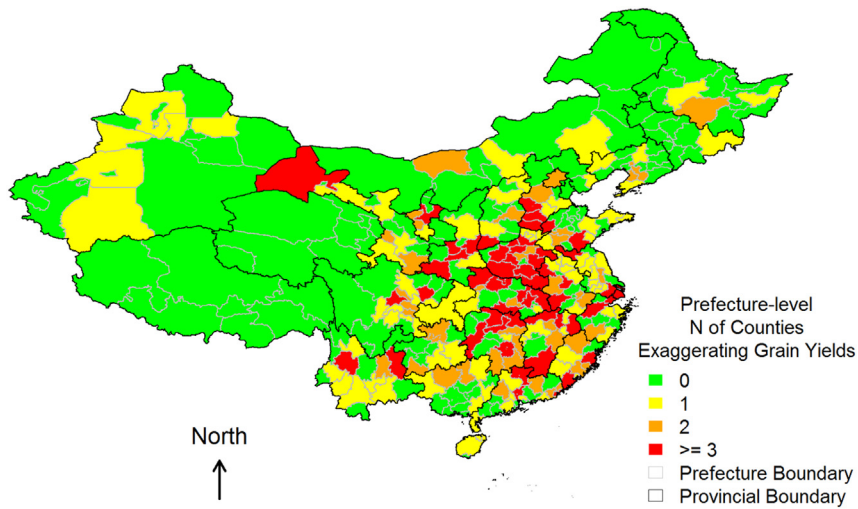


Fig. 2. Number of counties that ever exaggerated grain yields in each prefecture (i.e., the first instrumental variable).

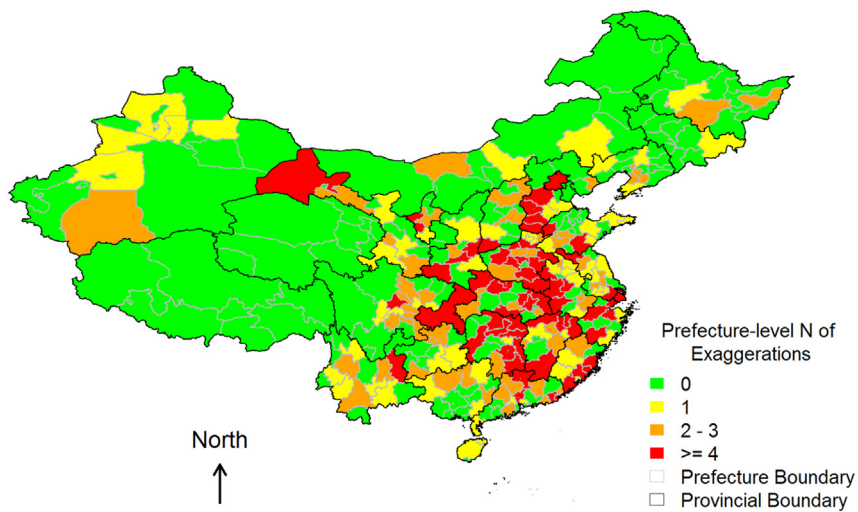


Fig. 3. Total number of exaggerations of grain yields in each prefecture (i.e., the second instrumental variable).

Table 3
Estimated coefficients of famine effects on high-risk biomarkers.

	Diastolic BP	Systolic BP	Resting pulse	Cholesterol			Triglyceride	Glucose	HbA1c
				HDL	LDL	Total			
SCD (ref: post-famine)									
Pre-famine	0.074	0.279***	0.043	-0.036	0.266**	0.144*	-0.020	0.157*	0.077
Famine	0.126	0.163*	0.003	-0.031	0.252**	0.024	0.055	0.175*	0.049
DCT (ref: post-famine)									
Pre-famine	-0.039	-0.047	0.580†	-0.059	0.094	0.035	0.089	-0.107	0.388
Famine	-0.082	-0.038	0.323†	-0.108	0.079	-0.111	0.159	0.040	0.278
DID (ref: post-famine)									
Provincial EDR									
×Pre-famine	-0.026*	-0.012	-0.001	-0.013†	-0.016	-0.019†	-0.015	0.001	0.009
×Famine	-0.011	-0.021*	-0.018†	-0.006	-0.020†	-0.023†	-0.008	-0.014	0.001
Provincial CSSI									
×Pre-famine	-0.009	-0.004	-0.003	-0.001	-0.009	-0.012†	0.001	0.007	0.009
×Famine	-0.005	-0.008	-0.007	0.003	-0.024**	-0.018*	0.000	-0.003	-0.002
Prefecture CSSI									
×Pre-famine	-0.005	-0.003	-0.006	-0.004	-0.007	-0.010*	-0.002	0.005	0.004
×Famine	-0.004	-0.004	-0.001	0.001	-0.022***	-0.019**	-0.001	0.000	0.000
Famine cohort probit									
Provincial EDR	-0.017†	-0.022**	-0.027*	-0.017*	-0.008	-0.012	-0.008	-0.006	-0.004
Provincial CSSI	-0.014**	-0.013**	-0.016*	-0.008	-0.010	-0.010	-0.006	-0.004	-0.005
Prefecture CSSI	-0.012*	-0.009†	-0.008	-0.006	-0.010†	-0.010†	-0.005	-0.002	-0.001
Famine cohort 2SLS									
Prefecture CSSI	0.001	-0.004	0.003	0.006	-0.005†	-0.004	0.000	0.000	0.000
Diagnostic statistics									
First-stage F	16.74	16.74	19.61	15.56	14.98	15.56	15.56	15.49	12.67
Stock-Yogo's test	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed

Note: SCD = simple cohort difference; DCT = deviation from cohort trend; DID = difference-in-differences; 2SLS = two-stage least squares; EDR = excess death rate; CSSI = cohort size shrinkage index. All the models control for gender and birth quarters. The SCD, DCT, and DID models additionally control for provincial fixed effects. † $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ based on prefecture-level cluster standard errors.

severity had similar or even lower levels of health risks in later life (as indicated by the biomarkers) than did the post-famine cohort living in areas less affected by the famine.

A similar pattern that contradicts the fetal origins hypothesis persists when we relaxed the assumption about no intrinsic cohort difference by exploiting only the spatial variation in famine severity within the famine cohort. Famine severity, regardless of how it is measured, was negatively associated with the risks of having abnormal diastolic and systolic BP. Provincial-level famine severity, measured by EDR or CSSI, was also negatively associated with high resting pulse. Famine severity was also negatively associated with some measures of cholesterol: provincial-level EDR was significantly related to HDL cholesterol, and prefecture-level CSSI was marginally related to both LDL and total cholesterol.

Turning to the IV estimates, the first-stage F -statistic unambiguously exceeds the rule of thumb value of 10, and the Stock-Yogo's test of weak IVs is rejected for every biomarker (Staiger and Stock, 1997). The IV estimates showed a marginally significant, negative effect of prefecture-level famine severity on abnormally high LDL cholesterol, contrary to the fetal origins hypothesis.

As a sensitivity check, Table 4 reports the estimates of famine effects on cardiovascular and metabolic diseases measured by combinations of biomarkers and self-reports. A similar pattern emerges as that in Table 3. The SCD estimates provide strong evidence for the fetal origins hypothesis, since both the pre-famine and famine cohorts were at greater risks of hypertension and diabetes compared to the post-famine cohort. The DCT estimates also indicate elevated risks of heart problems in the pre-famine and famine cohorts. However, these significant relationships either disappeared or were reversed in direction in the DID estimates. One exception is that the pre-famine cohort living in areas more heavily affected by the famine remained at a marginally significantly higher risk of diabetes when famine severity was approximated by provincial-level CSSI. When we exploited spatial variations in

famine exposure among the famine cohort alone, we again found evidence against the fetal origins hypothesis. Famine severity, measured by EDR or CSSI at provincial or prefecture-level, consistently reduced the risks of hypertension, heart problem, and dyslipidemia. The IV estimates show no significant famine effects.

As another sensitivity check, we repeated the same analyses in Table 3 with additional controls for self-reported diagnoses and treatments for cardiovascular and metabolic diseases. Including these endogenous variables does not alter the general patterns described above, despite reduced significance levels for several coefficients (see Appendix Table A2).

To further replicate and compare with previous research, Table 4 also reports estimates for nutritional outcomes based on anthropometric measures. Overall, we observed the same results as before, with famine exposure having either no effect or seemingly health-promoting effects through reduced BMI and risks of overweight and abdominal obesity. One exceptional finding pertains to height. Famine exposure significantly undermined height attainment. This result persisted when we examined famine cohort only. Another exceptional finding is that the IV estimate indicates a marginally significant, positive effect of prefecture-level CSSI on the risk of abdominal obesity.

Lastly, the spatial clusterings of high CSSI (see Fig. 1) suggest potential correlation of the error term across nearby prefectures. This may not be a serious problem in our study because ignoring such between-prefecture correlation may result in underestimated standard errors and hence inflated significance levels. Any statistical adjustment would then further attenuate rather than improve the existing evidence of the fetal origin hypothesis – a key argument in this study. Furthermore, the error term may not be spatially correlated across prefectures after controlling for CSSI and other covariates and calculating robust errors as we have done. Because a full spatial model is beyond the scope of this study, we investigated this possibility through a simple residual analysis. We calculated residuals from the models estimated above and regressed them on

Table 4
Estimated coefficients of famine effects on combined disease risks from biomarkers, self-reports, and anthropometries.

	Hypertension	Heart problem	Dyslipidemia	Diabetes	Height	BMI	Overweight	Abdominal obesity
SCD (ref: post-famine)								
Pre-famine	0.218***	0.075	0.020	0.141*	−1.332***	−0.542***	−0.125*	−0.069
Famine	0.128*	0.020	0.039	0.166*	−0.727*	−0.165	−0.060	−0.047
DCT (ref: post-famine)								
Pre-famine	0.013	0.525*	−0.066	−0.166	−0.631	0.451	0.262	0.227
Famine	−0.036	0.261†	−0.074	0.017	−0.300	0.191	0.096	0.047
DID (ref: post-famine)								
Provincial EDR								
×Pre-famine	−0.012†	−0.003	−0.019**	0.007	−0.082**	−0.040*	−0.008	−0.015**
×Famine	−0.012	−0.023*	−0.014†	−0.015	0.007	−0.041*	−0.017*	0.002
Provincial CSSI								
×Pre-famine	−0.004	−0.002	−0.007	0.009†	−0.038	−0.017	−0.002	−0.006
×Famine	−0.004	−0.011*	−0.005	−0.005	0.014	−0.018	−0.009†	0.007
Prefecture CSSI								
×Pre-famine	−0.003	−0.005	−0.007†	0.006	−0.028	−0.026*	−0.004	−0.006
×Famine	−0.003	−0.004	−0.005	−0.002	0.013	−0.021†	−0.008*	0.004
Famine cohort probit/OLS								
Provincial EDR	−0.017*	−0.029**	−0.020**	−0.012	−0.127**	−0.019	−0.009	0.010
Provincial CSSI	−0.012*	−0.022**	−0.012**	−0.007	−0.079*	−0.019	−0.009†	0.007
Prefecture CSSI	−0.010*	−0.012*	−0.010*	−0.005	−0.057*	−0.022†	−0.009*	0.003
Famine cohort 2SLS								
Prefecture CSSI	−0.004	−0.003	0.003	−0.002	−0.107	0.049	−0.002	0.015†
Diagnostic statistics								
First-stage F	16.98	19.91	15.12	15.36	16.67	16.89	16.89	16.77
Stock-Yogo's test	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed

Note: SCD = simple cohort difference; DCT = deviation from cohort trend; DID = difference-in-differences; 2SLS = two-stage least squares; EDR = excess death rate; CSSI = cohort size shrinkage index. All the models control for gender and birth quarters. The SCD, DCT, and DID models additionally control for provincial fixed effects. † $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ based on prefecture-level cluster standard errors.

the longitudes and latitudes of prefecture centroids. We found only a limited number of significant coefficients (see Appendix Table A3), suggesting that between-prefecture correlation was not a serious problem.

6. Discussion

During the past few decades, health researchers have been enthusiastic in testing Barker's fetal origins hypothesis (Barker, 1990, 1995a,b) because, if sustained empirically, it can contribute significantly to our knowledge about the long-term health consequences of early-life events (for reviews, see Bateson et al., 2004; Ben-Shlomo and Kuh, 2002; Gluckman et al., 2008). Evidence supporting the fetal origins hypothesis also has important public health implications in terms of identifying the right target (pregnant women), timing (prenatal), and strategy (improving maternal nutrition) for interventions to reduce health disparities in future generations. Therefore, recent empirical studies have used famines as natural experiments to address the endogeneity problems in inferring the causal health effects of prenatal malnutrition from observational data.

Our analysis addresses the problems of measurement error and intrinsic cohort difference that plague conventional studies. Through a systematic comparison, we show how estimates of the famine effects can be driven by the choice of analytical strategy. We conclude that using famine as a natural experiment does not eliminate the potential for making erroneous inferences when other analytical flaws are present.

The current study is innovative and contributes to the literature in at least two important ways. First, using China's 1959–1961 famine as a natural experiment, we have conducted a comprehensive analysis of the impacts of prenatal malnutrition on later-life chronic disease risks by comparing estimation results from a variety of measurement and estimation strategies. Because we drew on publicly available, nationally representative data sources (the 2011 CHARLS, the 1% sample of the 1990 China Population

Census, and the *People's Daily*), our findings have greater generalizability than existing studies using regional samples (Chen and Zhou, 2007; Fung and Ha, 2010; Gørgens et al., 2012; Huang et al., 2010b, 2013; Luo et al., 2006; Meng and Qian, 2009), and can be replicated by other researchers. We also expanded health indicators from such anthropometric proxies as body weight and height to more direct and accurate measures using biomarkers that capture cardiovascular and metabolic functions. Contrary to the provincial homogeneity assumption often held in previous studies (Chen and Zhou, 2007; Fan and Qian, 2015; Fung and Ha, 2010), we illustrated considerable regional variations in famine severity at sub-provincial level (Table 2 and Fig. 1). We exploited these finer geographic variations to obtain more accurate DID estimates than past studies that focused solely on provincial variations. Last, we went beyond conventional between-cohort comparison and DID approaches by constructing IVs to adjust for measurement errors in famine exposure and focusing on spatial variations within the famine cohort only. These features of our analytical approach together ensure the robustness of our empirical findings.

Second, after comparing results obtained from a variety of analytic strategies, we concluded that evidence supporting the fetal origins hypothesis was weak. Only when using the standard SCD estimates (and in a few cases the DCT estimates) did we find significantly increased chronic disease risks among the pre-famine and famine cohorts. After purging constant cohort difference across regions that were affected by the famine to varying degrees, the DID estimates indicated either null or seemingly positive health effects (reduced chronic disease risks) in later life associated with prenatal famine exposure. Similar results persisted when we entirely discarded cohort comparison and instead leveraged fine-grained geographic variations in famine severity within the famine cohort. The IV estimates were generally not significant; and when they were borderline significant in a few cases, they suggested positive health effects of prenatal famine exposure. These results are robust against different measures of

health.

We acknowledge that our estimates of long-term health effects of the GLF famine are confounded by differential mortality and interpret the null and negative findings as the evidence of selective mortality at work rather than the counterevidence against the fetal origins hypothesis per se. In other words, we do not believe that prenatal exposure to famine could “causally” decrease one’s health risks in later life. Instead, we concur with Song and colleagues that because fetuses, infants, children, and adults of poorer health endowment are more likely to be lost in the famine, the surviving famine cohort are likely to consist of the fittest individuals who are resilient to adverse environments (Song, 2009, 2010, 2013b; Song et al., 2009). Therefore, it is only possible to observe the negative long-term health effects of famine when famine survivors are not dominated by the selective frailty process. The positive estimates in this study may simply reflect that mortality selection outweighs the famine effects. On the other hand, our insignificant estimates suggests that either mortality selection offsets the famine effects, or that early-life exposure to the GLF famine indeed has no long-term health consequence. One appealing strategy to disentangle the long-term health effect of famine and the selection effect is to compare the health outcomes between the children of the pre-famine and famine cohorts and those of the post-famine cohort, the idea being that children inherit their parents’ genotype (the selection effect) but not their phenotype (the famine effect) (Gørgens et al., 2012). Unfortunately, CHARLS does not collect data on children’s cardiovascular and metabolic conditions.

We draw two important methodological implications for future research. First, estimates of famine effects can be highly sensitive to choices of analytic strategies and measures of health outcomes. For example, the typical SCD estimates show opposite effects to the other estimates in this study. The DID estimates turned out more significant for cholesterol-related biomarkers but less so after being combined with self-reported disease diagnosis; while the within-famine cohort estimates exhibited more significant results for measures related to hypertension but not diabetes. Thus, more comprehensive sensitivity analyses using multiple health indicators and modes are warranted in future research to avoid making biased inference.

Second, empirical tests of long-term health effects of prenatal exposure to famine are challenging, because mortality selection may dominate the health-detrimental famine effects in observational data. Although it remains methodologically appealing in terms of accounting for endogeneity, regarding famine as a natural experiment has drawbacks as demonstrated in this study.

Several study limitations are noteworthy. First, similar to many other famine studies, we do not have data to accurately measure individual-level prenatal exposure. Using ecological measures of famine severity at provincial or prefecture-level is a reasonable alternative, but may conceal important individual heterogeneity within the same region. Second, ecological measures such as EDR and CSSI are all estimated from government statistics that may be subject to falsification, although our IV approach helps to reduce measurement error. Third, like other studies of the GLF famine, we have (1) difficulty demarcating the timing of in utero famine exposure due to the lack of reliable and accurate vital statistics for this period (Susser and St Clair, 2013), and (2) a relatively small sample size after excluding the missing data.

Lastly, our IVs may be invalid if regional variation in exaggerations of grain yields was not related to regional variation in famine severity (violating the relevance assumption), or exaggerations of grain yields correlated with, for example, the CCP support which in turn affected subsequent recovery from the famine and population health (violating the exclusion restriction assumption). We have performed an additional analysis by constructing numbers of old revolutionary bases at the prefecture-level from government statistics (Department of Agriculture, 1989) as an indicator of regional political loyalty to the CPC (Kung and Lin, 2003), and found it only modestly correlated (in the range of 0.16–0.19) with the two IVs (results not shown). Furthermore, massive political turmoil and economic disruption continued nationwide for nearly two decades after the GLF, severely restricting government efforts to restore public health resources. Therefore, it is possible that the IV assumptions still hold. Despite these limitations, overall, our study adds new empirical evidence and methodological insights to the controversial literature on the fetal origins hypothesis.

Appendix

Table A1. Estimated coefficients of famine effects on high-risk systolic blood pressure.

	SCD	DCT	DID		
			Provincial EDR	Provincial CSSI	Prefecture CSSI
Male (ref: female)	−0.021	−0.021	−0.020	−0.021	−0.021
Birth quarter (ref: quarter 1)					
Quarter 2	0.065	0.064	0.067	0.066	0.064
Quarter 3	0.120	0.120	0.125	0.122	0.121
Quarter 4	0.117 [†]	0.117 [†]	0.121*	0.120*	0.117 [†]
Birth year		0.084			
Birth year ²		−0.027			
Birth year ³		0.002			
Cohort (ref: post-famine)					
Pre-famine	0.279***	−0.047	0.374***	0.437 [†]	0.391 [†]
Famine	0.163*	−0.038	0.316**	0.477	0.340
Provincial EDR			0.033*		
Provincial EDR × Pre-famine			−0.012		
Provincial EDR × Famine			−0.021*		
Provincial CSSI				−0.081*	
Provincial CSSI × Pre-famine				−0.004	
Provincial CSSI × Famine				−0.008	
Prefecture CSSI					0.003
Prefecture CSSI × Pre-famine					−0.003
Prefecture CSSI × Famine					−0.004

(continued)

	Diastolic BP	Systolic BP	Resting pulse	Cholesterol			Triglyceride	Glucose	HbA1c
				HDL	LDL	Total			
Longitude	0.009	0.008	0.011†	-0.003	0.009	0.006	-0.006	-0.007	-0.005
Latitude	0.002	0.006	0.005	0.002	-0.011†	-0.001	0.021**	0.029***	0.008
Provincial CSSI									
Longitude	0.008	0.007	0.012†	-0.003	0.007	0.004	-0.005	-0.007	-0.006
Latitude	0.003	0.006	0.002	-0.001	-0.006	0.001	0.020**	0.028***	0.009†
Prefecture CSSI									
Longitude	0.004	-0.001	0.010†	-0.005	0.008	0.007	-0.005	-0.002	0.002
Latitude	0.007	0.008	-0.001	0.000	-0.006	-0.001	0.017**	0.010	0.004
Famine cohort 2SLS									
Longitude	0.002*	-0.004**	0.005***	-0.001	-0.002	-0.001	0.002	0.001	0.001
Latitude	0.000	0.004**	-0.003**	0.000	0.000	0.000	-0.001	0.000	0.001

Note: SCD = simple cohort difference; DCT = deviation from cohort trend; DID = difference-in-differences; 2SLS = two-stage least squares; EDR = excess death rate; CSSI = cohort size shrinkage index. All the models control for gender and birth quarters. The SCD, DCT, and DID models additionally control for provincial fixed effects. † $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ based on prefecture-level cluster standard errors.

References

- Almond, D., 2006. Is the 1918 influenza pandemic over? Long-term effects of in utero influenza exposure in the post-1940 U.S. population. *J. Political Econ.* 114, 672–712.
- Almond, D., Currie, J., 2011. Killing me softly: the fetal origins hypothesis. *J. Econ. Perspect.* 25, 153–172.
- Almond, D., Edlund, L., Li, H., Zhang, J., 2010. Long-term effects of early-life development: evidence from the 1959 to 1961 China famine. In: Ito, T., Rose, A. (Eds.), *The Economic Consequences of Demographic Change in East Asia*. University of Chicago Press, Chicago, pp. 321–345.
- Angrist, J.D., Krueger, A.B., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. *J. Econ. Perspect.* 15, 69–85.
- Ashton, B., Hill, K., Piazza, A., Zeitz, R., 1984. Famine in China, 1958–61. *Popul. Dev. Rev.* 10, 613–645.
- Banister, J., 1987. *China's Changing Population*. Stanford University Press, Stanford, CA.
- Barker, D.J.P., 1990. The fetal and infant origins of adult disease: the womb may be more important than the home. *BMJ Br. Med. J.* 301, 1111.
- Barker, D.J.P., 1995a. The fetal and infant origins of disease. *Eur. J. Clin. Invest.* 25, 457–463.
- Barker, D.J.P., 1995b. Fetal origins of coronary heart disease. *BMJ Br. Med. J.* 311, 171–174.
- Barker, D.J.P., Osmond, C., 1986. Infant mortality, childhood nutrition, and ischaemic heart disease in England and Wales. *Lancet* 327, 1077–1081.
- Bateson, P., Barker, D., Clutton-Brock, T., Deb, D., D'Udine, B., Foley, R.A., et al., 2004. Developmental plasticity and human health. *Nature* 430, 419–421.
- Ben-Shlomo, Y., Kuh, D., 2002. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *Int. J. Epidemiol.* 31, 285–293.
- Bernstein, T.P., 1984. Stalinism, famine, and Chinese peasants. *Theory Soc.* 13, 339–377.
- Chen, Y., Zhou, L.-A., 2007. The long-term health and economic consequences of the 1959–1961 famine in China. *J. Health Econ.* 26, 659–681.
- Coale, A.J., 1984. *Rapid Population Change in China, 1952–1982*. National Academy Press, Washington, D.C.
- de Rooij, S.R., Painter, R.C., Phillips, D.I.W., Osmond, C., Michels, R.P.J., Godsland, I.F., et al., 2006. Impaired insulin secretion after prenatal exposure to the Dutch famine. *Diabetes Care* 29, 1897–1901.
- Department of Agriculture, P.s.R.o.C., 1989. *Rural Economic Statistics of China: 1949–1986*. Agriculture Press, Beijing.
- Fan, W., Qian, Y., 2015. Long-term health and socioeconomic consequences of early-life exposure to the 1959–1961 Chinese famine. *Soc. Sci. Res.* 49, 53–69.
- Fung, W., Ha, W., 2010. Intergenerational effects of the 1959–61 China famine. In: Fuentes-Nieva, R., Seck, P.A. (Eds.), *Risk, Shocks and Human Development: On the Brink*. Palgrave-Macmillan, London, UK, pp. 222–254.
- Gluckman, P.D., Hanson, M.A., Cooper, C., Thornburg, K.L., 2008. Effect of in utero and early-life conditions on adult health and disease. *N. Engl. J. Med.* 359, 61–73.
- Gluckman, P.D., Hanson, M.A., Spencer, H.G., Bateson, P., 2005. Environmental influences during development and their later consequences for health and disease: implications for the interpretation of empirical studies. *Proc. R. Soc. B Biol. Sci.* 272, 671–677.
- Goldstein, A., 1991. *From Bandwagon to Balance-of-power Politics: Structural Constraints and Politics in China, 1949–1978*. Stanford University Press, Stanford, CA.
- Gørgens, T., Meng, X., Vaithianathan, R., 2012. Stunting and selection effects of famine: a case study of the great Chinese famine. *J. Dev. Econ.* 97, 99–111.
- Huang, C., Li, Z., Venkat Narayan, K.M., Williamson, D.F., Martorell, R., 2010a. Bigger babies born to women survivors of the 1959–1961 Chinese famine: a puzzle due to survival selection? *J. Dev. Orig. Health Dis.* 1, 412–418.
- Huang, C., Li, Z., Wang, M., Martorell, R., 2010b. Early life exposure to the 1959–1961 Chinese famine has long-term health consequences. *J. Nutr.* 140, 1874–1878.
- Huang, C., Phillips, M.R., Zhang, Y., Zhang, J., Shi, Q., Song, Z., et al., 2013. Malnutrition in early life and adult mental health: evidence from a natural experiment. *Soc. Sci. Med.* 97, 259–266.
- Kannisto, V., Christensen, K., Vaupel, J.W., 1997. No increased mortality in later life for Cohorts Born during famine. *Am. J. Epidemiol.* 145, 987–994.
- Koupil, I., Shestov, D., Sparén, P., Plavinskaja, S., Parfenova, N., Vägerö, D., 2007. Blood pressure, hypertension and mortality from circulatory disease in men and women who survived the siege of Leningrad. *Eur. J. Epidemiol.* 22, 223–234.
- Kung, J.K.-s., Chen, S., 2011. The tragedy of the nomenklatura: career incentives and political radicalism during China's great leap famine. *Am. Political Sci. Rev.* 105, 27–45.
- Kung, J.K.-s., Lin, Justin Y., 2003. The causes of China's great leap famine, 1959–1961. *Econ. Dev. Cult. Change* 52, 51–73.
- Li, Y., He, Y., Qi, L., Jaddoe, V.W., Feskens, E.J.M., Yang, X., et al., 2010. Exposure to the Chinese famine in early life and the risk of hyperglycemia and type 2 diabetes in adulthood. *Diabetes* 59, 2400–2406.
- Li, Y., Jaddoe, V.W., Qi, L., He, Y., Lai, J., Wang, J., et al., 2011a. Exposure to the Chinese famine in early life and the risk of hypertension in adulthood. *J. Hypertens.* 29, 1085–1092.
- Li, Y., Jaddoe, V.W., Qi, L., He, Y., Wang, D., Lai, J., et al., 2011b. Exposure to the Chinese famine in early life and the risk of metabolic syndrome in adulthood. *Diabetes Care* 34, 1014–1018.
- Lin, J.Y., Yang, D.T., 2000. Food availability, entitlements and the Chinese famine of 1959–61. *Econ. J.* 110, 136–158.
- Lindeboom, M., Portrait, F., van den Berg, G.J., 2010. Long-run effects on longevity of a nutritional shock early in life: the Dutch potato famine of 1846–1847. *J. Health Econ.* 29, 617–629.
- Lumey, L.H., Stein, A.D., Susser, E., 2011. Prenatal famine and adult health. *Annu. Rev. Public Health* 32, 237–262.
- Luo, Z., Mu, R., Zhang, X., 2006. Famine and overweight in China. *Appl. Econ. Perspect. Policy* 28, 296–304.
- Meng, X., Qian, N., 2009. *The Long Term Consequences of Famine on Survivors: Evidence from a Unique Natural Experiment Using China's Great Famine*. National Bureau of Economic Research Working Paper Series, No. 14917.
- Painter, R.C., Roseboom, T.J., Bleker, O.P., 2005. Prenatal exposure to the Dutch famine and disease in later life: an overview. *Reprod. Toxicol.* 20, 345–352.
- Paneth, N., Susser, M., 1995. Early origins of coronary heart disease: the Barker hypothesis. *BMJ* 310, 411–412.
- Peng, X., 1987. Demographic consequences of the great leap forward in China's provinces. *Popul. Dev. Rev.* 13, 639–670.
- Portrait, F., Teeuwisz, E., Deeg, D., 2011. Early life undernutrition and chronic diseases at older ages: the effects of the Dutch famine on cardiovascular diseases and diabetes. *Soc. Sci. Med.* 73, 711–718.
- Ravelli, A.C.J., van der Meulen, J.H.P., Michels, R.P.J., Osmond, C., Barker, D.J.P., Hales, C.N., et al., 1998. Glucose tolerance in adults after prenatal exposure to famine. *Lancet* 351, 173–177.
- Roseboom, T.J., van der Meulen, J.H.P., Ravelli, A.C.J., Osmond, C., Barker, D.J.P., Bleker, O.P., 2001. Effects of prenatal exposure to the Dutch famine on adult disease in later life: an overview. *Mol. Cell. Endocrinol.* 185, 93–98.
- Song, S., 2009. Does famine have a long-term effect on cohort mortality? Evidence from the 1959–1961 great leap forward famine in China. *J. Biosoc. Sci.* 41, 469–491.
- Song, S., 2010. Mortality consequences of the 1959–1961 great leap forward famine in China: debilitation, selection, and mortality crossovers. *Soc. Sci. Med.* 71, 551–558.
- Song, S., 2013a. Identifying the intergenerational effects of the 1959–1961 Chinese great leap forward famine on infant mortality. *Econ. Hum. Biol.* 11, 474–487.

- Song, S., 2013b. Prenatal malnutrition and subsequent foetal loss risk: evidence from the 1959-1961 Chinese famine. *Demogr. Res.* 29, 707–727.
- Song, S., Wang, W., Hu, P., 2009. Famine, death, and madness: schizophrenia in early adulthood after prenatal exposure to the Chinese great leap forward famine. *Soc. Sci. Med.* 68, 1315–1321.
- Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. *Econometrica* 65, 557–586.
- Stanner, S.A., Bulmer, K., Andrès, C., Lantseva, O.E., Borodina, V., Poteen, V.V., et al., 1997. Does malnutrition in utero determine diabetes and coronary heart disease in adulthood? Results from the Leningrad Siege study, a cross sectional study. *BMJ Br. Med. J.* 315, 1342–1348.
- Susser, E., St Clair, D., 2013. Prenatal famine and adult mental illness: interpreting concordant and discordant results from the Dutch and Chinese famines. *Soc. Sci. Med.* 97, 325–330.
- Tan, C.M., Tan, Z., Zhang, X., 2014. Sins of the Father: The Intergenerational Legacy of the 1959-61 Great Chinese Famine on Children's Cognitive Development. The Rimini Center for Economic Analysis Working Papers. Rimini, Italy.
- Yang, D.L., 1996. Calamity and Reform in China: State, Rural Society, and Institutional Change since the Great Leap Famine. Stanford University Press, Stanford, CA.
- Yao, S., 1999. A note on the causal factors of China's famine in 1959–1961. *J. Political Econ.* 107, 1365–1369.
- Zhao, Y., Crimmins, E., Hu, P., Hu, Y., Ge, T., Kim, J.K., et al., 2014a. China Health and Retirement Longitudinal Study: 2011-2012 National Baseline Blood Data Users' Guide. China Center for Economic Research, Peking University, Beijing.
- Zhao, Y., Hu, Y., Smith, J.P., Strauss, J., Yang, G., 2014b. Cohort profile: the china health and retirement longitudinal study (CHARLS). *Int. J. Epidemiol.* 43, 61–68.