I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14487

Early-Life Famine Exposure, Hunger Recall and Later-Life Health

Zichen Deng
Maarten Lindeboom

JUNE 2021

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 14487

Early-Life Famine Exposure, Hunger Recall and Later-Life Health

Zichen Deng
NHH Norwegian School of Economics

Maarten Lindeboom
Vrije Universiteit Amsterdam, Monash University, Tinbergen Institute and IZA

JUNE 2021

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Early-Life Famine Exposure, Hunger Recall and Later-Life Health*

We use newly collected individual-level hunger recall information from the China Family Panel Survey to estimate the causal effect of undernourishment on later-life health. We develop a Two-Sample Instrumental Variable (TSIV) estimator that can deal with heterogeneous samples. We find a non-linear relationship between mortality rates, a commonly used famine indicator, and the individual hunger experience. The nonlinearity in famine exposure may explain the variation in the famine's effect on later life health found in previous studies. We also find that exposure to famine-induced hunger early in life leads to worse health among females fifty years later. This effect is much larger than the reduced-form effect found in previous studies. For males, we find no impact.

JEL Classification: I12, J11, C21, C26

Keywords: famine, hunger, developmental origins, two-sample instrumental variable

Corresponding author:
Maarten Lindeboom
Vrije Universiteit Amsterdam
Centre for Health Economics
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

E-mail: m.lindeboom@vu.nl

* This version: Monday 14th June, 2021. We gratefully acknowledge the valuable comments from the co-editor and three anonymous reviewers.

1 Introduction

A large number of observational studies have demonstrated that health and economic disparities may have roots early in life. This relationship has been shown in studies that examine the association between birth weight and later-life health and socioeconomic outcomes, as well as from studies that use "natural experiments" that cause some individuals (the treated) to be more likely to be affected by adverse conditions than others (the controls). In natural experiments, researchers mostly use contextual factors at an aggregate level to proxy for individual circumstances early in life. Examples of natural experiments include epidemics (Almond, 2006), economic conditions (Van den Berg, Lindeboom, and Portrait, 2006) and famines (Chen and Zhou, 2007; Lumey, Stein, and Susser, 2011). This paper adds to this literature by looking at the long-run consequences of actual exposure to hunger early in life.

In empirical applications, there are often limits to using aggregate indicators as a proxy for individual conditions. First, being born when the event in question took place is not equivalent to actually being exposed to adverse conditions. For example, researchers have used famine in a region as a proxy for being exposed to hunger. However, living in a food-deprived area is not equivalent to actually experiencing hunger, even if the famine's timing and location are precisely known. Wealthier households may still have sufficient food, or some parts of an exposed area may be less affected by the famine. Also, there is often uncertainty about the location and timing of the famine. For instance, the famine may be preceded by a prolonged period of food insecurity, so it is not always clear when the famine started. Most studies rely on historical evidence about the famine's evolution, and in some cases (such as the Chinese Famine), there may be conflicting information from historical sources. In these cases, researchers have limited information to justify the choices underlying their empirical approach, how they define the "treatment" period, which famine indicator they use, or the functional form of the parametric model. Without information on actual exposure among the survivors, the estimates are, at best, attenuated intention-to-treat (ITT) effects, but they may be biased.

One solution is to use more granular data about hunger prevalence, such as excess mortality rates or food prices. But these may also be imperfect proxies for the nutritional environment, since they may not include informal trade systems or other contextual factors that affect mortality rates and exaggerate (or attenuate) the intensity of the famine. The other alternative is to use actual hunger experience. Recently, this information has become available in a few data sets, such as The Survey of Health, Ageing and Retirement in Europe (SHARE) and the Chinese Family Panel Survey (CFPS). These data resolve uncertainty about the precision of famine proxies and may aid in providing causal evidence about the treatment effect of actual hunger exposure on later life outcomes. The data richness is a significant advantage when studying the long-run impacts of early childhood conditions. In this paper, we use hunger recall information to estimate the effect of undernourishment early in life on health, measured by the Metabolic Syndrome Index (Hoynes, Schanzenbach, and Almond, 2016). We develop a Two-Sample Instrumental Variable (TSIV) method that relaxes the homogeneity assumption in standard TSIV methods and can deal with two samples from different populations.

In a recent paper, Van den Berg, Pinger, and Schoch (2016) developed a TSIV approach to examine the causal effect of early-life hunger exposure on later life health outcomes among SHARE respondents from 3 European countries. They estimate the strength of the association between the famine and actual hunger, and the effect of hunger on height. Hunger recall information is also imperfect because those experiencing hunger at very early ages have much less hunger recall than respondents who were older during the famine. To avoid this problem, Van den Berg, Pinger, and Schoch (2016) use another sample of older siblings with more reliable hunger recall information to measure the association between the famine and hunger experience.

The idea of combining two samples can be traced back to Angrist and Krueger (1992) and was further developed in Inoue and Solon (2010). Two-sample methods implicitly assume homogeneity between the two samples. However, this homogeneity assumption may

not always be satisfied in practice. For instance, in a study on intergenerational income mobility Björklund and Jäntti (1997) use independent samples of fathers and sons. Likewise, Currie and Yelowitz (2000) look at the effect of public housing participation on housing quality and educational attainment of children and combine information from CPS on program participation with census data on outcomes. In both applications the sample moments of the common variables differ significantly across the two data sets being combined. In our context, very young children are, in general, frailer than older children, adolescents and prime aged adults. Therefore, very young children surviving a famine are more likely to have better biological traits or to come from more affluent families than their older counterparts. As a consequence, for the primary sample of those who are exposed early in life, the distribution of observed and unobserved confounding factors, is likely to be different from the distribution in the second (auxiliary) sample of older individuals. If there is heterogeneity in confounding factors, our estimates of the causal effect will be biased when the parametric model is misspecified.

Van den Berg, Pinger, and Schoch (2016) solve this by using discrete instruments and covariates and stratifying the samples into a finite set of homogeneous subsamples. When the famine's start and end are not precise, or if there is substantial variation in the famine's intensity across regions, one may want to rely on continuous instruments such as excess mortality rates or prices. Researchers would prefer a method that accommodates continuous instruments and covariates.

To combine information from two different samples in a robust way, we propose a two-step approach. In the first step, we use a non-parametric method to balance the primary and the auxiliary sample. In the second step, we apply a two-sample IV method. The first step decreases the model dependency on functional forms of the parametric causal inference in the second step. We show that the estimator in the matched sample yields unbiased estimates of the analogous regression coefficients in the population of the primary sample. Importantly, our results are valid for continuous treatment variables and continuous instruments. Monte

Carlo simulations show that all estimators perform well in terms of bias when the first stage is correctly specified. In this scenario, traditional two-sample estimators (TSIV and TSTSLS) perform much better on efficiency. The two-step two-sample estimator proposed here performs much better than the other estimators when the first stage is misspecified.

Using the two-step method, we reexamine the long-run health impact of the Great Chinese Famine. The Great Chinese Famine has been studied extensively. Early studies (Chen and Zhou, 2007; Almond, Edlund, Li, and Zhang, 2010) documented substantial effects on height, wealth, and cognitive function in later life. On the other hand, there are also studies that find no effects (see e.g. Kim, Fleisher, and Sun (2017), Meng and Qian (2009) and Xu, Li, Zhang, and Liu (2016)). These differences may be the result of different studies exploiting different historical sources and using different instruments. We focus on individuals born during or shortly before the famine (1957–1962). Unlike previous Chinese famine studies, we use hunger recall information and supplement the primary data set with a second sample of much older individuals born between 1910 and 1947. For these older cohorts, hunger recall error biases are less of a problem. We use this second (auxiliary) sample to estimate the effect of famine exposure on the probability of reporting hunger.

We use nearest-neighbor matching to homogenize the distributions of covariates in the two samples and examine the relationship between famine exposure and the probability of reporting hunger. Virtually all previous papers estimate reduced form relationships between later life health and famine indicators, thereby implicitly assuming a linear relationship between famine indicators and actual hunger exposure among the survivors. We find that the linear approximation for the first stage does not fit the data and requires a logarithmic transformation of excess mortality rates (EDR) to make the relationship linear. The nonlinear relationship between EDR and hunger experiences has consequences for the previous contributions in the Chines Famine literature that estimated reduced-form models that were linear in the instruments. Next, we estimate the impact of hunger on an index of metabolic syndrome and find that early-life hunger for females leads to a 0.4 standard deviation in-

crease in later life metabolic syndrome. This number is much larger than previous estimates in the literature. For males, we find much smaller and insignificant effects.

Our analysis makes four important contributions to the literature on long-run effects. We are the first study on the Chinese famine to provide evidence on the strength of the famine-hunger association. Our first stage results also validate the commonly used instruments in this literature to proxy undernourishment. We find a strong association between the aggregate famine indicators and reported hunger experience. However, this association is not a simple linear relationship. This nonlinearity might explain some of the conflicting findings in the literature. Second, we adapt the standard two-sample instrumental variable model to deal with heterogeneous samples. We propose using a non-parametric method to process the data before the parametric econometric analyses. Monte Carlo simulations show that estimates are less model-dependent when preprocessing balances the primary and the auxiliary samples.

Third, we provide new evidence on the long-run impact of early-life hunger experiences. All but one of the previous papers in this area used reduced-form approaches that include famine indicators rather than actual hunger experience. They thus estimate intention-to-treat (ITT) effects. We show that in the Chines famine, the causal treatment effects are much larger than the intention-to-treat effects found in this literature. Finally, our paper discusses the exclusion restriction required in IV famine studies. In our context, where we aim to assess the effect of undernourishment and use hunger recall information, we have to assume that the famine affects children only via hunger, and that there are no other channels. Although the famine's primary impact is food restriction, we cannot exclude other potential impacts such as stress and/or infectious diseases that often accompany famines. Thus, it is likely that the exclusion restriction is violated. Violations of the exclusion restriction are relevant when interpreting reduced-form ITT estimates. We adopt a recently proposed exercise that bounds the treatment effect under weaker assumptions (i.e., that relaxes the strict exclusion restriction Conley, Hansen, and Rossi, 2012). Our bounding exercises shed light on the

nutrition contribution of the famine's impact relative to all other potential channels. We conclude that our main results hold under much weaker assumptions.

The remainder of the paper is structured as follows. Section 2 briefly describes the historical background, institutional setting, and important features of the Great Chinese Famine. In section 3, we describe the data sets we use in this study and discuss our malnutrition indicator and our outcome variables. Section 4 introduces the framework, discusses identification assumptions and presents the results from model simulations. In section 5 we apply the method to estimate the causal effect of famine induced undernourishment early in life on later life health.

2 Background and Prior Research

2.1 The Great Chinese Famine

The Great Chinese Famine occurred from 1958 to 1961 and is widely considered "the worst famine in human history". During the Famine, at least 16.5 million individuals perished in rural areas (see Sen, 1981; Ravallion, 1997).

Since 1949, the central government adopted the Stalinist development model, which emphasized investment in industrial sectors. The rural sector had to provide resources for investment and raw materials for production. To accommodate high investment in industry, the government initiated a large scale land reform, followed by an aggressive collectivization policy. During the land reform period (1950–52), redistribution of landlord-held land and other property boosted agricultural production. Major indicators of productivity in the rural sector, such as grain and cotton outputs, had double-digit growth rates during this period. Collectivization of the rural sector followed immediately after this rapid growth period. It started with the "Five-Year Plan" (1953–57), in which peasant households were organized into agricultural producers' cooperatives. This reform dramatically slowed agricultural growth rates.

On the eve of the famine, the central government in China controlled food production, distribution, and consumption. Approximately 80% of the population worked in the agriculture sector. Grain was harvested and stored communally, and private stores of grain were prohibited. The central government procured grain produced in rural areas from communal depots after the fall harvest. Procured grain was fed to urban workers, exported to other countries in exchange for industrial equipment and expertise, and stored in reserves as insurance against natural disasters. The grain retained by the rural regions was used to feed the peasants in communal kitchens, which were established so that the collective could control the preparation and consumption of food. Furthermore, the government prevented peasants from migrating and, consequently, peasants could only consume the food distributed to their collective.

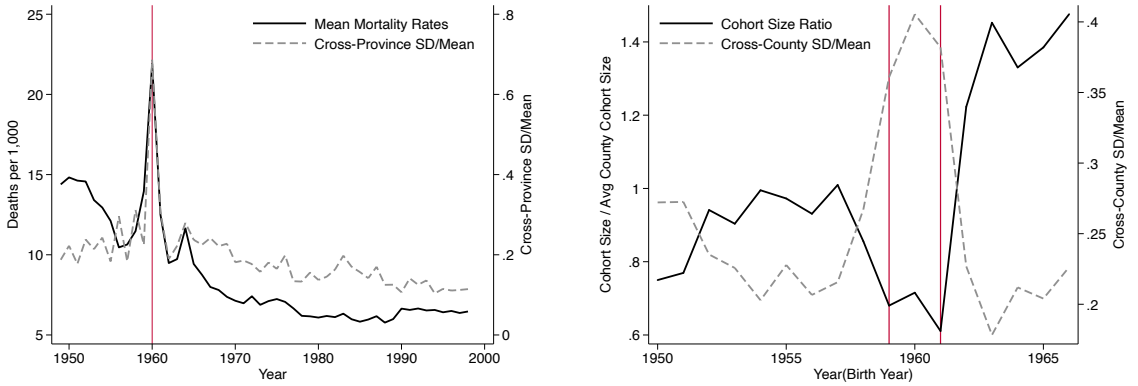
There is a consensus in the literature that the Great Chinese Famine was a direct consequence of Mao's Great Leap Forward, an economic and social campaign led by the Chinese Communist Party from 1958 to 1961 (Kung and Lin, 2003; Meng, Qian, and Yared, 2015). During the campaign, the political climate encouraged provincial leaders to overstate grain production and even export grain to signal the success of Mao's Great Leap Forward (see Meng, Qian, and Yared, 2015). Despite a severe shortage of food, China was a net grain exporter in 1960 (Yao, 1999; Lin and Yang, 2000).

2.2 Relevant Features of the Famine

The famine lasted until 1962, but some researchers have argued (see Tan, Zhibo, and Zhang, 2015, for an example) that birth and death rates in some provinces had already returned to normal levels by 1961. The precise end date for the famine is not clear for all provinces. With hunger recall data from the CFPS, we can address this issue in more detail (see section 3). The famine also featured considerable variation in severity across regions. In 1960 death rates for two adjacent provinces could differ by more than five-fold. For instance, in 1960 the province of Anhui had a death rate of 1.84%, while the neighboring province of Jiangsu

had a death rate of 0.29%.

Figure 1. Average and Spatial Variation in Famine Severity



(a) Province-level mortality rates **(b)** County-level survivor birth cohort sizes

Notes: Figure 1a: The solid line plots mean mortality rates, which are average mortality rates across provinces in each year. The dashed line is the standardized variance in mortality rates across provinces in year t. Figure 1b: The solid line plots the detrended 1% size of the birth cohort born in year t. The dashed line is the normalized cross-county variance in birth cohort sizes. Source: Meng, Qian, and Yared (2015)

To better depict the variation in severity across regions during the Great Chinese Famine, we present some graphical evidence from Meng, Qian, and Yared (2015). Figure 1a plots average mortality rates and the normalized variance in mortality rates over time (the cross-province standard deviation divided by the cross-province mean). The figure shows that during the famine (denoted by the two vertical lines), both mean mortality and the variance in mortality rates spiked. Our empirical strategy exploits the variation in famine induced mortality rates across provinces. Figure 1b provides complementary county-level evidence. The figure plots mean and cross-county standardized variance in cohort size. This figure shows a clear drop in cohort size and increased variance during the famine.

2.3 Selected Famine Studies

For an overview of famine studies, we refer to Van den Berg and Lindeboom (2018). Here we highlight the results from the two most widely studied famines: the "Dutch Hunger Winter" famine and the Chinese Famine. The Dutch Hunger Winter (December 1944–April 1945)

famine is the most studied famine in the epidemiological and demographic literature. The famine has a number of features that are advantageous for researchers: it arrived unexpectedly, lasted for a short period, and took place in a relatively stable society with thorough data collection. Studies using this famine found effects on blood glucose levels, diabetes, severe obesity, high blood pressure (hypertension), and schizophrenia (See Lumey, Stein, and Susser (2011) for an excellent review). Scholte, van den Berg, and Lindeboom (2015) find negative effects of the famine on labor market outcomes and hospitalization outcomes.

The Great Chinese Famine is the second most used famine in the literature on the long-run effect of exposure early in life. Li and Lumey (2017) provide an extensive review and meta-analysis of the medical and epidemiological research. They conclude that the literature has found effects for overweight, type 2 diabetes, hyperglycemia, metabolic syndrome, and schizophrenia. However, most studies vary substantially in exposure definition, control selection, and analytical methods. When controlling for these differences, they conclude that "most effects commonly attributed to the famine can be explained by uncontrolled age differences between exposed and control groups". This is in line with Xu, Li, Zhang, and Liu (2016), who find that estimates of the famine effects are sensitive to the choice of health indicators, measures of famine severity, and regression model specifications. One of the earliest economics papers (Chen and Zhou, 2007) finds substantive effects for height, labor supply, and earnings. Their findings are confirmed by Meng and Qian (2009). Almond, Edlund, Li, and Zhang (2010) look at the effect of famine exposure on literacy, labor market status, wealth, and marriage market outcomes. They find that exposed women marry later and have less educated spouses. They also find evidence for the Trivers-Willard hypothesis that the sex-ratio of the offspring of exposed parents favors daughters. Few economic studies target specific chronic conditions such as diabetes and hypertension. One exception is Kim, Fleisher, and Sun (2017). They do not find effects on chronic diseases such as hypertension.

3 Data

3.1 Hunger Recall

Our main data come from the China Family Panel Study (CFPS), a large-scale, nationally representative panel survey conducted by the Institute of Social Science Survey at Peking University. Currently, four waves are available, 2010, 2012, 2014, and 2016. The baseline wave (hereafter CFPS-2010) is collected through a multistage probability sampling procedure and consists of 14,798 households. All adults living in the household are interviewed, leading to a total sample of 34,425 adult observations.

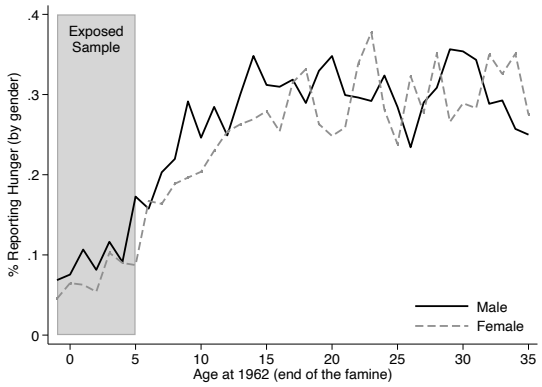
Similar to the SHARE survey, the CFPS-2010 survey included a question on hunger recall. The survey asked: "Have you experienced starvation for more than one week? If so, when did it start, when did it end[1], and where did it happen?" Since the question only requires the experience to last more than one week, we don't know how many weeks in total respondents have experienced food shortage in each year. Note that the Great Chinese famine was not explicitly mentioned in the questionnaire. Therefore, respondents were not primed towards a specific answer. The non-response rate for the hunger experience question is very low (about 0.055). This is similar to the non-response rate in the SHARE survey. Most of the hunger experiences happened during the Great Chinese Famine, which occurred more than 50 years before the survey. Hunger responses related to the great Chinese famine are likely to be subject to recall bias, especially for respondents born close to the famine. Figure 2 supports this suspicion.

The figure displays the fraction of individuals in the raw data who report hunger as a fraction of those who were alive during the famine period. The horizontal axis is the respondent's age in 1962; the vertical axis is the fraction of individuals who experienced hunger during the famine (1958–1962). The fraction of hunger recall increases with age during the famine and stabilizes at about 30 percent after age 12.[2] We see no gender

[1] The public data only provides information at the year level.

[2] Several validation studies match recall data with actual outcomes and find that the recall data is reliable

Figure 2. Probability of Reporting Hunger Conditional on Famine Exposure



Notes: The exposed sample includes individuals who were born between 1958–1962.

differences in reports of hunger during the famine period. Our primary sample contains individuals born before and during the famine, whose own recall of malnutrition around birth and in the first years of life is likely to suffer from significant recall bias. To overcome this problem, we follow the idea introduced by Van den Berg, Pinger, and Schoch (2016) to use recall information from individuals who experienced the famine at an older age to proxy for actual hunger exposure for individuals in the primary sample.

3.2 The Primary and Auxiliary Sample and Summary Statistics

The primary sample includes individuals born between 1958 and 1962 and who lived in rural areas. After dropping individuals with missing information on the outcome variable (see below), we are left with 958 males and 972 females from 27 provinces. Very few people migrated during the famine (Chen and Zhou, 2007). This is also true for our sample of people living in rural China: only 4% live outside their province of birth at the time of the interview. For the reported health outcome, we pool data sets from the first three CFPS waves (CFPS-2010, CFPS-2012, and CFPS-2014).[3] We look at three chronic diseases: hypertension, diabetes, and obesity. The first two conditions are derived from the question

when individuals reach adult ages (see, for example, O'malley, Bachman, and Johnston (1983) on teen drinking behavior).

[3]The fourth wave, CFPS-2016, does not collect information on some of the health conditions we need for our health index (hypertension, diabetes, and height).

"Has a doctor ever told you that you suffer from…". Obesity is defined as a body mass index (BMI) exceeding 29. We used self-reported height and weight to calculate BMI. We construct a Metabolic Syndrome Index by grouping information on all three chronic conditions. The index is the average of the standardized z-scores for each component. High values of the index are associated with worse health.

To construct an auxiliary sample that is less susceptible to recall bias, we select females born between 1910 and 1947, aged 15–52 in 1962. If the female in the auxiliary sample is the mother of the individual in the primary sample, we include this observation as a proxy for the individual's hunger exposure. 10% of the matches are between the individual from the primary sample and his/her mother. For the remaining individuals in the primary sample, we match on the village of birth (or if not available, county of birth, or province of birth) and next on age and literacy.

We report summary statistics for the primary and auxiliary samples in Table 1. Panel A reports summary statistics in the primary sample for the main outcome and age and literacy status for the individual's mother. We report summary statistics for the health outcomes collected in CFPS2010, CFPS-2012, and CPFS-2014 in Panel B. Panel C reports summary statistics for the auxiliary sample (born between 1910 and 1947). We use the same auxiliary sample for the male and female sub-samples. Comparing the primary sample with the auxiliary sample, we see that mothers' literacy rates in the primary sample are much lower than mothers' literacy rates in the auxiliary sample. We also see that mothers in the auxiliary sample are about six years older than the mothers of individuals in the primary sample. This invalidates classical two-sample IV methods. Below we present our two-sample IV method that can be applied when the two samples have different distributions of observed and unobserved characteristics.

Table 1. Summary Statistics

	Female sample			Male sample		
	Obs.	Mean	SD	Obs.	Mean	SD
Panel A: Primary sample - basic information						
Age at 2010	958	50.40	1.88	972	50.59	1.88
Mother literate	958	0.15	0.36	972	0.18	0.39
Mother birth year	958	1931.39	7.26	972	1931.03	7.67
Father literate	958	0.44	0.50	972	0.44	0.50
Father birth year	958	1928.53	7.72	972	1928.18	8.33
Panel B: Primary sample - health outcomes						
Hypertension	2522	0.04	0.19	2618	0.02	0.15
Diabetes	2522	0.01	0.10	2618	0.01	0.09
Obesity	2522	0.04	0.21	2618	0.04	0.19
Metabolic syndrome(index)	2522	0.02	0.64	2618	-0.02	0.54
Panel C: Auxiliary sample						
Mother literate	3682	0.24	0.43	3682	0.24	0.43
Mother birth year	3682	1925.88	15.58	3682	1925.88	15.58
Father literate	3682	0.47	0.50	3682	0.47	0.50
Father birth year	3682	1923.04	15.82	3682	1923.04	15.82

Notes: Author's tabulations of CFPS-2010, CFPS-2012, and CFPS-2014. Panel A summarizes background information for individuals born in rural area between 1957 and 1962. Panel B pools chronic conditions data from three waves of CFPS. The Metabolic Syndrome Index is the z-score from subtracting the mean and dividing by the standard deviation. Both the mean and standard deviation are calculated using the analysis sample (individuals born between 1957 and 1962 in all three waves of CFPS). High values of the index are associated with worse health. Panel C displays the background information in the auxiliary sample, which includes all individuals born prior to 1947.

4 Two-Sample IV Models with Heterogeneous Samples

4.1 Model

Many researchers have used data combination methods to identify the causal effect when no single sample contains all relevant variables (see Ridder and Moffitt, 2007, for a review). Most empirical applications of two-sample methods implicitly assume homogeneity between the primary sample and the auxiliary sample. Table 1 in section 3 showed substantial differences in age and literacy status between the (proxy) mothers in the primary and auxiliary samples. Ignoring that these samples differ in important ways will result in first-stage estimates that

are not relevant for the primary sample and thus irrelevant for the treatment effect in the second stage (in the primary sample). We consider the following general framework:

$$Y_i = \psi(D_i, X_i, U_i), \tag{1}$$

where D_i denotes severe hunger during childhood for individual i, Y_i denotes health in adulthood. X_i denotes a vector of observed covariates. We are interested in the causal effect of hunger experiences in early life (D) on later-life outcomes (Y). There are a number of challenges to identifying the causal effect: D is likely to be endogenous, and D is systematically misreported or not in the same sample as Y. To solve the endogeneity problem, researchers usually instrument D with a contextual factor (Z) and estimate the intention-to-treat (ITT) effect. For instance, researchers have used being born in a famine-stricken area or excess mortality rates in an area as an instrumental variable for undernourishment early in life.

When Z and D are in the same sample, the local average treatment effect can be estimated using two-stage least squares (TSLS) or Instrumental Variable (IV) techniques. Mogstad, Torgovitsky, and Walters (2019) show that when Z is not a binary variable, the TSLS estimator can be understood as a weighted local average treatment effect with some additional monotonicity assumption. In practice, researchers often use a linear instrument-exposure model. From Vansteelandt and Didelez (2018) and Buja, Brown, Berk, George, Pitkin, Traskin, Zhang, and Zhao (2019), we know that the TSLS estimator is consistent even when the relationship between the treatment and the instrument is misspecified.

With variables in two different samples, we can use the primary sample to estimate the reduced-form equation and the auxiliary sample to estimate the first-stage equation. In our study, we use the sample of children born during the famine to estimate the reduced-form equation that relates Y to Z. The auxiliary sample of adults during the famine is used to estimate the relationship between D to Z (the first-stage equation). Although TSLS is robust to model misspecification in the one-sample setting, this robustness property does not

carry over to the two-sample setting Angrist and Krueger (1992). As has been pointed out by Graham, Pinto, and Egel (2016), early applications all (implicitly) assume that both samples are random samples from the study population (i.e., the samples are "compatible"). When the "compatible" assumption is not met, both TSIV and TSTSLS estimates are biased.

Our study uses children born during or shortly before the famine to estimate the reduced form equation and people who were already adults to estimate the first-stage equation. The mortality impact of a famine at very young ages is likely to be different from the mortality impact at adult ages. Therefore, the famine will affect different cohorts differently, resulting in differential mortality selection across different cohorts. Further, young children surviving the famine are more likely to come from families with favorable biological traits and/or (wealthier) families with better access to food. Table 1 showed that there were substantial differences between the distribution of covariates in the primary sample and the auxiliary sample. Two-sample estimates using the original (i.e., raw unbalanced) samples are therefore biased.

Ho, Imai, King, and Stuart (2007) propose preprocessing the data with matching methods to balance the treatment and control group to reduce the problem of model-dependent causal estimates. Similarly, we adopt a two-step approach to address heterogeneous samples in two sample settings. In the first step, we employ non-parametric preprocessing, such as nearest-neighbor matching, to balance the covariate distributions between the primary sample and the auxiliary sample. In the second step, we perform a parametric (or semi-parametric) analysis using the primary sample and the matched individuals from the auxiliary sample. The simplest way to understand our approach is to consider one-to-one-exact matching. This matches each individual in the primary sample to a close match in the auxiliary sample. After the matching procedure, the preprocessed auxiliary sample is balanced with the primary sample, with any unmatched auxiliary units discarded. With all units in the primary sample matched, this procedure eliminates dependence on the functional form of a parametric analysis in the second step. As a result, misspecification in the second step is less likely to

be a source of bias. In Appendix A.1, we show that the proposed two-step estimator (in the remainder referred to as the two-step-TSTSLS or two-step-TSIV) yields unbiased estimates of the analogous regression coefficients in the population of the primary sample.

Graham, de Xavier Pinto, and Egel (2012); Graham, Pinto, and Egel (2016) develop a method of Inverse Probability Tilting (IPT) that uses propensity score estimates, and next apply a weighting estimator.[4] The propensity score is then used to reweigh the auxiliary sample. In this paper, we pursue an approach that uses matching as a preprocessing of the data to balance the auxiliary and primary sample. It is well known that the matching approach has the disadvantage that part of the data will not be used, which causes inefficiency. However, recent literature (see, for instance, Armstrong and Kolesár, 2021) argues that matching is to be preferred when the conditional expectation function is not smooth enough. In that case, putting positive weight on observations other than the closest increases the bias too much, while one-to-one matching minimizes the bias. This observation also echoes early simulation results from Busso, DiNardo, and McCrary (2014), who show that matching methods outperform re-weighting when the overlap is sufficiently poor. The theoretical justification of the trade-off between matching and weighting can also be found in Hirshberg and Wager (2017); Kallus (2020). We don't want to repeat these arguments. Instead, we perform simulations in the section below to show the advantage and disadvantages of our estimator.

4.2 Monte Carlo Simulations

We performed two sets of Monte Carlo simulations to compare the performance of our two-step-TSTSLS estimator with the TSIV, the TSTSLS, and the IPT estimator.[5] Below we provide the general findings of these simulations. More detail can be found in Appendix A.3.

In the first set of simulations, we vary the degree of overlap in the distribution of X in the primary and auxiliary sample and, at the same time, vary the degree of misspecification

[4]Similar concepts have been developed and extended by other papers (Imai and Ratkovic, 2014).
[5]Simulations with a two-step-TSIV gave very similar results.

in the first stage regression. The simulations show that all estimators perform well in terms of bias when the first stage is correctly specified. In this secenario traditional estimators (TSIV and TSTSLS) perform much better on efficiency. The two-step-TSTSLS estimator does not use the information of all data points and therefore has much lower efficiency. The two-step-TSTSLS performs much better than the other estimators when the first stage is misspecified.

In the second set of simulations, we fix the overlap (we take it as poor) and vary the degree of misspecification in the first stage equation. These simulations show that when we gradually increase the degree of misspecification, the performance of the TSIV, the TSTSLS, and the IPT estimators deteriorate quickly. The two-step-TSTSLS estimator performs relatively well. The IV estimate can be understood as the ratio of the intention to treat estimate and the first stage estimate. Therefore, misspecifications in the first-stage regression translate into relatively large biases. Similarly, misspecifications in the first stage will increase the bias of the TSTSLS. The simulation results show that the two-step-TSTSLS is robust against misspecifications. We suggest that this method can be used as a robustness check in empirical applications.

5 Reexamining the Long-run Effect of Undernourishment using the Chinese Famine

In this section, we apply the method developed in the previous section to estimate the causal effect of famine induced undernourishment early in life on later-life health. The Great Chinese famine has been studied extensively. All studies use reduced form regressions that relate the outcome variable Y to instruments Z. The findings of these studies are mixed and largely due to differences in instruments Z, different data sets, different selections made in the construction of the analysis sample, and different specifications for the reduced-form regression (Li and Lumey (2017)).

5.1 Nearest-Neighbor Matching

We use nearest-neighbor matching based on village of birth, mother's age, and literacy to balance the two samples. When village of birth information is not available, we match individuals on county of birth, ensuring that the matched pairs have similar family background characteristics. Cao, Xu, and Zhang (2020) have documented that counties with large family clans experienced lower mortality during the famine. Table 2 presents the balanced primary and auxiliary samples after applying nearest-neighbor matching.[6] As an out-of-sample test, we also show summary statistics for two father's characteristics, which the matching algorithm does not target. The percentage of literate fathers is balanced between the primary and auxiliary samples. Our matching algorithm significantly improves balance for the average age of the father. The mean age difference decreases from 5 to 1.5 after the matching procedure. This improvement signals that the balance between the primary and auxiliary samples has been improved substantially, even for variables we did not explicitly target.

Table 2. Summary Statistics – Matched Sample

	Female sample			Male sample		
	Obs.	Mean	SD	Obs.	Mean	SD
Panel A: Matched primary sample						
Age at 2010	956	50.40	1.88	970	50.59	1.88
Mother literate	956	0.15	0.36	970	0.18	0.39
Mother birth year	956	1931.37	7.26	970	1931.04	7.68
Father literate	956	0.44	0.50	970	0.44	0.50
Father birth year	956	1928.52	7.72	970	1928.18	8.33
Panel B: Matched auxiliary sample						
Mother literate	956	0.12	0.33	970	0.16	0.37
Mother birth year	956	1930.63	9.60	970	1930.55	10.17
Father literate	956	0.42	0.49	970	0.38	0.49
Father birth year	956	1926.89	11.57	970	1926.74	11.76

Notes: Author's tabulations of CFPS-2010. Panel A summarizes background information for individuals in the matched primary sample. Panel B summarizes background information for individuals in the matched auxiliary sample.

[6]For less than 5% of the individuals in the main sample we can't find a match from the auxiliary sample.

5.2 Results from Two-sample IV Models

In the second step, we set up a linear model, to approximate the true causal model (1):

$$Y_i = \gamma D_i + \pi X_i + U_i, \tag{2}$$

where i indexes the individual. γ is the causal effect of hunger early in life on later-life health, our parameter of interest. As the model has only one endogenous variable and one instrumental variable, the estimates of γ consist of two components: the reduced-form (or ITT) estimates (3)

$$Y_i = \gamma_0 Z_i + \pi_0 X_i + W_i; \tag{3}$$

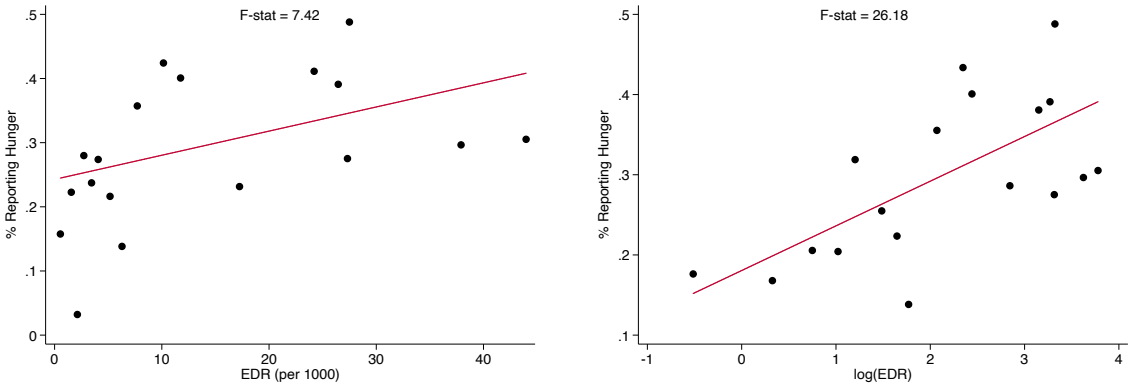
and, when hunger experience information (D_i) is available, the first stage regression (4)

$$D_i = \gamma_1 Z_i + \pi_1 X_i + V_i. \tag{4}$$

The excess mortality (death) rate (EDR) is commonly used as a famine intensity measure in studies of the great Chinese famine (Chen and Zhou, 2007; Almond, Edlund, Li, and Zhang, 2010). Other studies (for instance, Bleakley (2007)) use it as a measure of disease prevalence. We take province-level mortality rates from Meng, Qian, and Yared (2015) and construct our instrumental variable (i.e., the excess death rate in 1960) following Chen and Zhou (2007). We define the excess death rate in 1960 as the gap between the death rate in 1960 and the average death rate in the three years before 1959.

We first check that our instrument is relevant for the excess mortality rate. Figure 3a presents a binned scatter plot of the relationship between hunger experiences and excess mortality rates. The points on the figure plot the percentage of respondents who had hunger experiences during the famine. Interestingly, a linear relation, which is implicitly assumed by most famine studies, does not do a good job capturing the relationship between EDR and hunger. The percentage of individuals who reported any hunger experience remains

Figure 3. Mean Hunger Experiences versus EDR/log(EDR)



(a) Mean Hunger Experiences vs. EDR **(b)** Mean Hunger Experiences vs. log(EDR)

Notes: Figure 3a and 3b present two binned scatter plots of the relationship between the percentage of respondents who had hunger experiences during the famine and excess mortality rates in 1960. The excess death rate (EDR) in 1960 is the gap between the death rate in 1960 and the average death rate in the three years before 1959. Figure 3a uses raw excess mortality rates, while Figure 3b uses the log-transformation of excess mortality rates. The points on the figure plot the mean hunger experiences within each EDR/log(EDR) percentile bin. The best-fit line is estimated using an OLS regression on the underlying micro data. F-statistics for both regressions are reported separately.

more or less stable at 40% for excess mortality rates above about 15. This implies that, for high excess mortality rates, the instrument is no longer informative for actual hunger exposure. Indeed, the associated F-statistics of the linear regression is 7.42, suggesting that the (linear) EDR is a weak instrument. This may explain why studies that restrict their analysis to provinces with high mortality rates generally find small effects.[7]

An arbitrary solution would be to restrict the sample by leaving out provinces with extremely high mortality rates. However, this will considerably decrease the sample size and may generate selection bias, since the marginal survivor at high mortality rates will differ from the marginal survivor at lower mortality rates. Instead, we decided to take a log-transformation of excess mortality rates in the province where the individual lived (log(EDR)). Figure 3b shows that the relation between hunger experiences and the log(EDR) can be captured well by a linear function. The associated F-statistic for the regression is well above 10. The log-transformation is simple and convenient, but has the disadvantage of

[7] A large proportion (7/10) of regional studies surveyed in Li and Lumey (2017), which look at high mortality areas, finds insignificant impacts.

the functional form restriction. Therefore, we also used a more flexible specification (a set of EDR quartile dummies). We report these results in section 5.4.

Table 3. First-stage – the Effect of Famine Intensity on Hunger

	(1)	(2)	(3)	(4)	(5)	(6)
	Hunger					
	All	All	Female	Female	Male	Male
log(EDR)	0.075***	0.074***	0.087***	0.086***	0.062**	0.062**
	(0.015)	(0.015)	(0.015)	(0.014)	(0.016)	(0.016)
Mother literate		-0.020		-0.0065		-0.034
		(0.030)		(0.041)		(0.025)
Age		-0.29		-0.22		-0.32
		(0.39)		(0.63)		(0.50)
Age squared(/100)		0.28		0.21		0.32
		(0.38)		(0.62)		(0.50)
Observations	1926	1926	956	956	970	970
F-Stat	26.18	26.17	32.92	32.91	14.93	14.93

Notes: Each parameter is from a separate regression of hunger between 1957–1962 on log(EDR) (EDR is short for excess death rates). We estimate the model on the matched auxiliary sample. The stratification by gender is based on the gender variable in the primary sample. Standard errors, clustered by province of the birth, are in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table 3 presents estimation results from the first-stage linear regressions based on the matched auxiliary sample, with an indicator for hunger (recall) as the dependent variable. The table includes regressions without controls (columns 1, 3, and 5) and regressions with controls (2, 4, and 6). As controls, we use age and literacy status of the (proxy) mother. The table presents estimates for the full sample and estimates by gender. Across all specifications, we find highly significant effects of the instrument (log(EDR)) and F-statistics that well exceed 10, indicating that there is no problem of weak instruments (Staiger and Stock, 1997).

Table 4 presents two-step-TSIV estimates of the treatment effect on the matched primary sample. We follow Abadie and Spiess (2019) for the calculation of the standard errors, meaning that we form pairs from individuals, one individual from the primary sample and the other from the auxiliary sample. Next, in the bootstrapping procedure we sample provinces

Table 4. Effects of Hunger at Age 0–5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.13)	(0.20)	(0.21)
Mother literate		0.029		0.031
		(0.034)		(0.028)
Age		0.47***		0.14*
		(0.18)		(0.084)
Age squared(/100)		-0.43**		-0.13*
		(0.17)		(0.081)
Observations	2517	2517	2612	2612

Notes: The results are based on TSIV estimates from separate regressions. All regressions use the log(EDR) as the instrumental variable on the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

(that in turn consists of the different pairs) in order to allow for correlation among pairs born in the same province. The table shows that famine-induced hunger increases the standardized metabolic syndrome index by about 0.37. The ITT estimate from the reduced-form regression (3) equals 0.031 (see Table B2 of Appendix B for the full table). For males, the coefficients are small and insignificantly different from zero. These findings are at odds with a number of studies on the long-run health effects of adverse conditions in early childhood. Most of these studies find boys to be more sensitive to adverse conditions early in life. Our results are, however, in line with findings from other studies of the Great Chinese Famine. See e.g. Almond, Edlund, Li, and Zhang (2010) who also find stronger effects for females.[8] Table B1 in Appendix B shows results for the separate components of the metabolic syndrome index. For females, we find significant effects for all individual components (albeit only at the 10% level). Interestingly, for males we only find a significant effect on obesity.

[8]An explanation might be that in the south-east Asian context, gender preferences play a role. In the presence of such effects, parents may have redistributed food towards boys. This may affect the first stage estimates as well as the second stage estimates (parents may redistribute resources towards boys in the later stages of childhood).

We also use OLS regressions, where we use the individual's own hunger recall information. These OLS estimates are subject to two sources of bias: a bias due to the endogeneity of the hunger variable and a bias due to the recall error. The latter is likely to lead to a downward bias. We can also use the proxy hunger measurement directly (i.e., without instrumenting). This should correct the recall errors, but estimates are still subject to endogeneity bias. The results are presented in Table B4 of Appendix B. Using their own recall information, we get very small and insignificant coefficients. We also find small and insignificant coefficients when we use the recall of matched (proxy) mothers.

Height is often used as a proxy for health, and we also estimated the same model with height as a dependent variable. For both genders, we did not find effects on height. We also examined whether the metabolic syndrome maps into health care use. Specifically, we looked at the association between metabolic syndrome and the number of hospital visits. In line with our expectations, we find a positive association for females but not for males. (see Table B5 of Appendix B).

A significant proportion of previous studies that examine the Chinese Famine use the linear EDR in regressions like (3) as an instrument/proxy for the severity of the famine. To highlight that our two-step two-sample method is robust to functional form misspecification, we also present the instrumental variable estimate using the linear EDR in Table B3 of Appendix B. The table shows that all four IV estimates are very close to our main estimates using the log-transformed EDR (Table 4). Due to the weak instruments, the standard errors are much larger. These results support our claim that homogeneity between the two samples decreases biases due to model misspecifications.

5.3 Robustness to Violations of Perfect Exogeneity

While we expect hunger to be the most important channel through which a famine affects later life outcomes, it is likely that alternative channels directly affect later-life health as well. Famines may be accompanied by increased stress. Epidemiological studies find that

prenatal stress exposure in humans is associated with worse later-life health outcomes, in particular memory problems, decreased learning, depression, and dementia (Selten, van der Graaf, van Duursen, de Wied, and Kahn, 1999; Heffelfinger and Newcomer, 2001). Further, during a prolonged period of undernourishment the disease environment may change, which may increase the prevalence of infectious diseases. This may affect later-life health and socioeconomic outcomes (Almond, 2006). For such reasons the usual exclusion restriction required for a causal interpretation of IV estimates may not hold in our context.

To examine the robustness of our TSIV estimates, we relax the assumption of perfect exogeneity and derive bounds on the true effect of malnutrition early in life on later-life health following Conley, Hansen, and Rossi (2012). We only perform this analysis for females.[9] Consider a generalization of the standard IV model that allows the instrument Z to enter linearly in the second-stage regression,

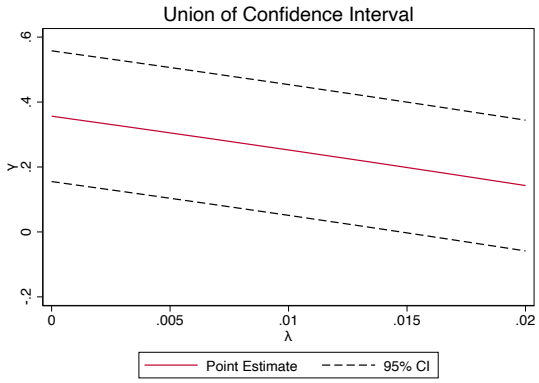
$$Y_i = \gamma D_i + \lambda Z_i + \pi X_i + U_i. \tag{5}$$

Conley, Hansen, and Rossi (2012) shows how to obtain consistent estimates of the effect of interest (here γ) if λ is known. By choosing a set of fixed values for $\lambda = \lambda_0$ and running separate regressions for each value of $\lambda = \lambda_0$, we can evaluate the sensitivity of the IV estimate of γ to violations of the exclusion restriction. Conley, Hansen, and Rossi (2012) choose the ITT estimate of the reduced-form regression (3) to determine the range of relevant values of $\lambda = \lambda_0$. With an ITT estimate 0.031 (see Table B2 of Appendix B) for females, we consider values $\lambda \in [0, 0.02]$.

We plot the TSIV estimates for different values of λ in Figure 4. The (straight) red line is the point estimate of γ under different values of λ. The dotted line is the 95% confidence interval. $\lambda = 0$ corresponds to the assumption that the famine affects later life health only via hunger (D). Indeed, the point estimate at $\lambda = 0$ is 0.37. Higher values of λ are associated with a more important role for alternative channels and hence a smaller role for hunger.

[9]The previous section only showed significant effects for females.

Figure 4. 95% Confidence Intervals to Exclusion Restriction Violations



Notes: The graph shows how the IV estimate of the effect of hunger in early life on hypertension changes when the exclusion restriction $\lambda = 0$ is violated. Estimates and confidence intervals come from estimating Equation (5) for females at different fixed values $\lambda = \lambda_0$. The red line is the point estimate under different values of lambda; the dotted line is the confidence interval. We control for fixed effects for year of birth, province of birth and year of interview. Family controls include mother's literacy and age.

For instance, with λ equal to 0.016, we assume that 50% of the famine affects health via undernourishment.

Figure 4 shows that the straight line of point estimates for γ is relatively flat. Specifically, at $\lambda = 0.016$, γ is about 0.2 and still significantly different from zero. The IV estimate only becomes insignificantly different from zero for higher values of λ. To reject the long-run impact of undernourishment, one must assume that at least 50% of the effect of the famine is due to channels other than undernourishment. Overall, this exercise indicates that our results are robust to moderate violations of the exogeneity assumption. The strength of the first-stage estimates may be important for this finding. Indeed, minor deviations from the exclusion restriction may greatly decrease precision when instruments are weak (see Conley, Hansen, and Rossi, 2012). We have a strong first stage (cf Table 3) and consequently, relatively small biases in the effect of hunger on later life health when the exclusion restrcition is violated.

5.4 Additional robustness checks and selection issues

In this section, we explore some additional specification checks to examine the robustness of our findings. The results are presented in Table 5. Columns 2 and 4 present estimates of with controls (age and literacy status of the mother); columns 1 and 3 report estimates without additional controls.

Table 4 uses bootstrapped standard errors based on matched pairs, but ignored uncertainty that arises in balancing the two samples in the first step. To shed light on this, we adjust the bootstrap by resampling the original data and next use the bootstrap two-step estimates to form alternative estimates of standard errors. To allow for arbitrary correlation of the errors for individuals born in the same province, we do the resampling at the province level. Standard errors based on this approach (in square brackets), are only slightly larger than those reported in Table 4. The estimates for females remain significant at the 5% level.[10]

We also restricted our sample to individuals born between 1958 and 1961. Although recent studies (Tan, Zhibo, and Zhang, 2015) show that the food shortage problems in some provinces were still present in 1962, the earlier literature often assumed that the famine ended in 1961. As a result the instrument may be less noisy, but the smaller sample size may reduce statistical power. Panel B of the table shows that the effects for females becomes larger and the standard errors are similar.

Third, we check the impact of migration. CFPS collects detailed information on the province of birth. While across province migration is limited (only 4%), our matching algorithm may perform worse for individuals who migrated. To check this, we drop 4% (82 cases) of individuals who do not live in their province of birth at the time of the survey. We report these results in Panel C of the table; our estimates are hardly affected by this restriction.

We used the natural logarithm of EDR as the instrumental variable. The ad hoc choice

[10]Standard errors clustered at the county of birth gave similar results.

Table 5. Robustness

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Panel A: alternative clustering				
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.13)	(0.20)	(0.21)
	[0.17]	[0.17]	[0.28]	[0.29]
Observations	2517	2517	2612	2612
Panel B: small sample window				
Hunger before age 5	0.44***	0.43***	0.059	0.079
	(0.12)	(0.14)	(0.22)	(0.23)
Observations	1859	1859	2040	2040
Panel C: control for migration				
Hunger before age 5	0.36***	0.36***	-0.014	0.0054
	(0.13)	(0.13)	(0.20)	(0.23)
Observations	2387	2387	2534	2534
Panel D: non-linear instruments				
Hunger before age 5	0.34***	0.35***	0.13	0.17
	(0.13)	(0.13)	(0.26)	(0.28)
Observations	2517	2517	2612	2612
Panel E: placebo test using cohort 1964-1967				
Hunger before age 5	-0.22	-0.22	-0.26	-0.19
	(0.20)	(0.20)	(0.35)	(0.36)
Observations	2707	2707	2519	2519

Notes: Each coefficient is from a separate regression. All regressions use the log(EDR) as the instrumental variable. In columns (2) and (4), we control for mother's literacy and age. In Panel A, alternative bootstrap that takes into account the matching step appear in square brackets. In Panel B, we drop individuals born in 1962. In Panel C, we drop individuals for whom the residing province is different from the province of birth. Panel D uses quartile dummies as the instrumental variable. In Panel E, we estimate the same model using individuals born between 1964 and 1967. Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

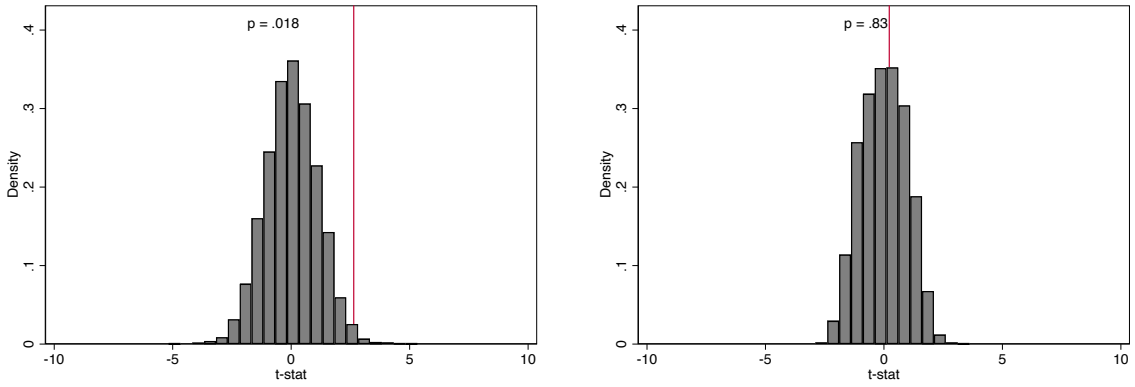
for the functional form may be questionable. We therefore also considered a more flexible approach where we use quartile dummies. The results, reported in Panel D of Table 5, shows that all four IV estimates are very close to the estimates in Table 4.

Additionally, we conduct a placebo exercise. Recall that we use variation in the peak excess mortality rate across provinces and link these mortality rates to individuals born in these provinces between 1958 and 1962. To examine whether other (province-level) confounding effects may drive our results, we apply our two-step estimator to the cohorts born between 1964 and 1967. Studies that use a DiD design (Chen and Zhou, 2007; Almond, Edlund, Li, and Zhang, 2010) use this cohort as a control group. The estimates, reported in Panel D, have the wrong sign and are all insignificant. These results suggests that our findings are driven by famine-induced hunger and not other mortality-related confounding factors that vary by province.

We conduct falsification tests to demonstrate the statistical power of our inferences by assigning a pseudo-treatment. We randomly assign province of birth and thus log(EDR) to each respondent in the primary sample. If our identification strategy is valid, estimates using these pseudo-samples should be centered around zero. In Figure 5, we plot the distribution of the t-statistics from 5,000 estimated pseudo-treatment effects for males and females. The two distributions are both centered around zero. To address our model's statistical power, we mark the location of the t-statistic of the baseline treatment effects in the distribution of pseudo-treatment effects in Table 4. We also report the share of the pseudo-treatment t-statistics that exceed the actual t-statistic of the baseline model (in absolute values). These p-values support our design and the statistical power of our exercises.

A famine, especially when it lasts for a prolonged period, leads to selective mortality and selective fertility. Starting with selective fertility, Roseboom (2010) show that during the Dutch Hunger Winter of 1944 about half of the women in exposed areas did not menstruate. Besides, in utero mortality is more likely to occur among frail fetuses. Frailty may depend on biological traits or poor living conditions for women during the gestational period. In addition to in utero mortality, mortality may also occur between birth and the survey in 2010. The extent to which these selection effects take place will vary with the intensity of the famine. For the Chinese famine, with substantial regional variation in famine intensity,

Figure 5. Pseudo-treatment Effects



(a) Females **(b)** Males

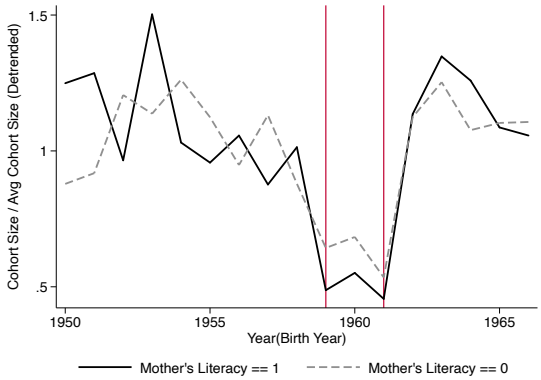
Notes: Pseudo-treatment vs. actual hunger exposure: the distribution of t-statistics resulting from 5,000 random assignments of treatment to individuals, as well as the t-statistics from the actual treatment through hunger exposure (red line). "p-values" report the share of the pseudo-treatment t-statistics that are larger than the actual t-statistics.

this will lead to systematic differences in province populations, leading to biased inferences.

We examine the possible influence of selection effects by looking at cohort sizes. We expand the CFPS data by including all individuals born between 1950 and 1966. We count the number of observations for each birth year and regress these cohort sizes on a linear time trend. Next, we use the ratio between observed cohort sizes and predicted sizes to plot the detrended cohort loss series. The resulting detrended series based on the CFPS closely resembles the detrended series based on the Census in 1990 (see Figure B1 of the Appendix B).

Figure 6 plots the detrended time series for literate and illiterate mothers. The figure shows a clear drop in cohort size of about 40–50 percent, indicating the importance of selective fertility and mortality. These cohort size reductions are in line with the findings from the Dutch Hunger Winter (Roseboom (2010) and Scholte, van den Berg, and Lindeboom (2015)). It is important whether the cohort size-reduction differs by socioeconomic status. Figure 6 shows that there are only small differences in cohort size loss by literacy status.

Figure 6. Cohort Loss by Mother's Literacy



Notes: Figure 6 plots the detrended relative cohort sizes by mother's literacy using CFPS-2010.

6 Conclusion

Undernourishment early in life may have lasting effects on health and socioeconomic outcomes later in life. Most of the epidemiological and economic literature has produced intention-to-treat (ITT) effects from reduced-form regressions that relate later life outcomes to indicators of famine exposure early in life. This paper uses an indicator of *actual* hunger experience early-in-life from a hunger recall question in the China Family Panel Survey (CFPS). The outcome variable is observed in the primary sample, and the hunger information is obtained from an auxiliary sample. We develop a two-step Two-Sample Instrumental Variable (TSIV) approach that can deal with differences in the distribution of observable and unobservable factors in the two samples. In the first step, we balance the primary and the auxiliary sample, next we apply the classical TSIV estimator on the balanced samples. Monte Carlo simulations show that this two Sample estimator performs much better than other two sample estimators when the first stage is misspecified. Using the CFPS data, we find evidence for long-term impacts on late-life health of early-life malnutrition for females, but not for males.

Using hunger recall information has two clear advantages over previous studies. First, it allows us to examine the relationship between hunger experiences and the commonly used

famine indicator (excess death rates). This helps to justify instruments, as well as providing insight into the proper specification of the reduced-form regressions used in the extensive famine literature. We find a non-linear relationship between hunger recall and excess death rates. A linear specification, generally used in the famine literature, leads to weak instrument problems. This may, in part, explain the differences in findings across studies. Further, with hunger recall information, we can estimate the causal effect of undernourishment on later-life health. We show that these effects are much larger. Importantly, they are robust to potentially mild and moderate violations of the exogeneity assumption.

References

Abadie, Alberto and Guido W Imbens. 2012. "A martingale representation for matching estimators." *Journal of the American Statistical Association* 107 (498):833–843.

Abadie, Alberto and Jann Spiess. 2019. "Robust Post-Matching Inference." *Working paper* .

Almond, Douglas. 2006. "Is the 1918 Influenza Pandemic Over? Long-term Effects of in Utero Influenza Exposure in the Post-1940 U.S. Population." *Journal of Political Economy* 114 (4):672–712.

Almond, Douglas, Lena Edlund, Hongbin Li, and Junsen Zhang. 2010. *Long-Term Effects of Early-Life Development: Evidence From the 1959 to 1961 China Famine*. University of Chicago Press.

Angrist, Joshua and Alan Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association* 87 (418):328–336.

Armstrong, Timothy and Michal Kolesár. 2021. "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness." *Working paper* .

Björklund, Anders and Markus Jäntti. 1997. "Intergenerational Income Mobility in Sweden Compared to the United States." *The American Economic Review* 87 (5):1009–1018.

Bleakley, Hoyt. 2007. "Disease and Development: Evidence from Hookworm Eradication in the American South." *The Quarterly Journal of Economics* 122 (1):73–117.

Buja, Andreas, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. 2019. "Models as Approximations I: Consequences Illustrated with Linear Regression." *Statist. Sci.* 34 (4):523–544.

Busso, Matias, John DiNardo, and Justin McCrary. 2014. "New Evidence on the Finite Sample Properties of Propensity Score Reweighting and Matching Estimators." *Review of Economics and Statistics* 96 (5):885–897.

Cao, Jiarui, Yiqing Xu, and Chuanchuan Zhang. 2020. "Clans and Calamity: How Social Capital Saves Lives during China's Great Famine." *Working paper* .

Chen, Xiaohong, Han Hong, Alessandro Tarozzi et al. 2008. "Semiparametric Efficiency in GMM Models with Auxiliary Data." *Annals of statistics* 36 (2):808–843.

Chen, Yuyu and Li-An Zhou. 2007. "The Long-Term Health and Economic Consequences of the 1959–1961 Famine in China." *Journal of Health Economics* 26 (4):659–681.

Conley, Timothy G, Christian B Hansen, and Peter E Rossi. 2012. "Plausibly exogenous." *Review of Economics and Statistics* 94 (1):260–272.

Currie, Janet and Aaron Yelowitz. 2000. "Are Public Housing Projects Good for Kids?" *Journal of Public Economics* 75 (1):99–124.

Graham, Bryan S, Cristine Campos de Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *The Review of Economic Studies* 79 (3):1053–1079.

Graham, Bryan S, Cristine Campos de Xavier Pinto, and Daniel Egel. 2016. "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)." *Journal of Business & Economic Statistics* 34 (2):288–301.

Heffelfinger, Amy K. and John W. Newcomer. 2001. "Glucocorticoid effects on memory function over the human life span." *Development and Psychopathology* 13 (3):491–513.

Hirshberg, David A and Stefan Wager. 2017. "Augmented Minimax Linear Estimation." *Working paper* .

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3):199–236.

Hoynes, Hilary, Diane Whitmore Schanzenbach, and Douglas Almond. 2016. "Long-Run Impacts of Childhood Access to the Safety Net." *American Economic Review* 106 (4):903–934.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B: Statistical Methodology* :243–263.

Inoue, Atsushi and Gary Solon. 2010. "Two-sample Instrumental Variables Estimators." *The Review of Economics and Statistics* 92 (3):557–561.

Kallus, Nathan. 2020. "Generalized Optimal Matching Methods for Causal Inference." *Journal of Machine Learning Research* 21 (62):1–54.

Kim, Seonghoon, Belton Fleisher, and Jessica Ya Sun. 2017. "The Long-term health effects of fetal malnutrition: Evidence from the 1959–1961 China great leap forward famine." *Health economics* 26 (10):1264–1277.

Kung, James Kai-sing and Justin Yifu Lin. 2003. "The Causes of China's Great Leap Famine, 1959–1961." *Economic Development and Cultural Change* 52 (1):51–73.

Li, Chihua and LH Lumey. 2017. "Exposure to the Chinese famine of 1959-61 in early life and long-term health conditions: a systematic review and meta-analysis." *International journal of epidemiology* .

Lin, Justin Yifu and Dennis Tao Yang. 2000. "Food Availability, Entitlements and the Chinese Famine of 1959–61." *The Economic Journal* 110 (460):136–158.

Lumey, Lambert H, Aryeh D Stein, and Ezra Susser. 2011. "Prenatal Famine and Adult Health." *Annual Review of Public Health* 32:237–262.

Meng, Xin and Nancy Qian. 2009. "The Long Term Consequences of Famine on Survivors: Evidence from a Unique Natural Experiment using China's Great Famine." *Working paper* .

Meng, Xin, Nancy Qian, and Pierre Yared. 2015. "The Institutional Causes of China's Great Famine, 1959–1961." *The Review of Economic Studies* 82 (4):1568–1611.

Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters. 2019. "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables." *Working paper* .

O'malley, Patrick M., Jerald G. Bachman, and Lloyd D. Johnston. 1983. "Reliability and Consistency in Self-Reports of Drug Use." *International Journal of the Addictions* 18 (6):805–824. PMID: 6605313.

Ravallion, Martin. 1997. "Famines and Economics." *Journal of Economic Literature* 35 (3):1205–1242.

Ridder, Geert and Robert Moffitt. 2007. "The Econometrics of Data Combination." *Handbook of Econometrics* 6:5469–5547.

Roseboom. 2010. *Baby's van de Hongerwinter. De onvermoede erfenis van ondervoeding.* Augustus.

Scholte, Robert S., Gerard J. van den Berg, and Maarten Lindeboom. 2015. "Long-run Effects of Gestation During the Dutch Hunger Winter Famine on Labor Market and Hospitalization Outcomes." *Journal of Health Economics* 39:17–30.

Selten, Jean-Paul, Yolanda van der Graaf, Rozemarijn van Duursen, Christien C Gispen de Wied, and René S Kahn. 1999. "Psychotic illness after prenatal exposure to the 1953 Dutch Flood Disaster." *Schizophrenia Research* 35 (3):243–245.

Sen, Amartya. 1981. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford University Press.

Shu, Heng and Zhiqiang Tan. 2020. "Improved methods for moment restriction models with data combination and an application to two-sample instrumental variable estimation." *Canadian Journal of Statistics* .

Staiger, Douglas and James Stock. 1997. "Instrumental Variables with Weak Instruments." *Econometrica* 65 (3):557–586.

Tan, Chih Ming, Tan Zhibo, and Xiaobo Zhang. 2015. "Sins of the Fathers: The Intergenerational Legacy of the 1959-61 Great Chinese Famine on Children's Cognitive Development." *Working paper* .

Van den Berg, Gerard J and Maarten Lindeboom. 2018. "Famines, Hunger, and Later-Life Health." *The Oxford Research Encyclopedia of Economics and Finance* .

Van den Berg, Gerard J, Maarten Lindeboom, and France Portrait. 2006. "Economic Conditions Early in Life and Individual Mortality." *The American Economic Review* 96 (1):290–302.

Van den Berg, Gerard J, Pia R Pinger, and Johannes Schoch. 2016. "Instrumental Variable Estimation of the Causal Effect of Hunger Early in Life on Health Later in Life." *The Economic Journal* 126 (591):465–506.

Vansteelandt, Stijn and Vanessa Didelez. 2018. "Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators." *Scandinavian Journal of Statistics* 45 (4):941–961.

Xu, Hongwei, Lydia Li, Zhenmei Zhang, and Jinyu Liu. 2016. "Is natural experiment a cure? Re-examining the long-term health effects of China's 1959–1961 famine." *Social Science & Medicine* 148:110–122.

Yao, Shujie. 1999. "A Note on the Causal Factors of China's Famine in 1959–1961." *Journal of Political Economy* 107 (6):1365–1369.

Appendix A Theoretical Results

A.1 Model

Suppose we are interested in estimating the treatment effect of a binary treatment D on outcome Y in a primary population of interest, which is confounded by measured covariates X and unmeasured ones U, with the aid of an instrumental variable Z. However, we only observe $(Y_i, Z_i, X_i), i = 1, ..., N_1$ from this population F_p. As a remedy, suppose an additional sample $(D_j, Z_j, X_j), j = 1, ..., N_0$ is available from an auxiliary population F_a, possibly different from the primary population. Let R be an indicator variable equal to 1 if drawn from the primary population and 0 otherwise. We use the notation $D(0)$ to represent the latent D in the primary sample. The following assumptions give a formal definition of the data combination model.

Assumption A1 (Random Sampling)**.** *With probability $Q \in (\xi, 1-\xi)$ for $0 < \xi < 0.5$, we draw a unit at random from F_p and record its realizations of Y, Z, and X, otherwise we draw a unit at random from F_a and record its realizations of D, Z, and X.*

Assumption A2 (Weak Overlap)**.** *Let $\mathcal{X}_p = \{x : f_p(x) > 0\}$ and $\mathcal{X}_a = \{x : f_a(x) > 0\}$, then $\mathcal{X}_p \subseteq \mathcal{X}_a$.*

Assumption A3 (Conditional Distributional Equality)**.** $F_p(D(0)|Z, X) = F_a(D|Z, X)$, $F_p(Y|Z, X) = F_a(Y|Z, X)$

Similar to Graham, Pinto, and Egel (2016); Shu and Tan (2020), Assumption A1 defines how the data are generated. Assumption A2 states that the support of the common variables (Z, X) in the primary sample is contained within the support of the auxiliary sample. This ensures that for each unit in the study (primary) sample, there will be matching units with similar values of X in the auxiliary sample.[11] Assumption A3 requires predictive invariance

[11] In the empirical application this is also verified: for less than 5% of the individuals in the main sample, we can't find a match from the auxiliary sample.

for the treatment between the two heterogeneous populations. The distributions of (Y, Z, X) and (D, X, Z) in the two populations differ only in terms of their marginal distributions for the always measured variable, (Z, X). This assumption is similar to the idea of selection-on-observables.

Let P^* be the matched sample generated by matching each unit in the primary sample, i, to the auxiliary sample, $J(i)$ with replacement. We only consider one-to-one matching, since the auxiliary sample in our empirical application (section 5) is only slightly larger than the primary sample. We choose the sets of matches $J(i)$ to minimize the sum of the matching discrepancies, $\sum_{i=1}^{N_1} d(X_i, X_{j(i)})$, where $d(\cdot)$ is the distance metric to measure the matching discrepancies. The commonly used distance metrics include, for example, the Mahalanobis distance. Similar to the matching literature, we assume that the sum of matching discrepancies vanishes (i.e., $\frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} d(X_i, X_{j(i)}) \xrightarrow{p} 0$) quickly enough to allow asymptotic unbiasedness as $N_0, N_1 \to \infty$ with $N_0 > N_1$.

We now describe the population distribution targeted by the matched sample, P^*. Since $F_p(\cdot)$ and $F_a(\cdot)$ are the cumulative distribution functions from the primary and auxiliary samples, we define $E_p[\cdot]$ and $E_a[\cdot]$ as the corresponding expectation operators. We define a matching target distribution, F_p^*, as

$$
\begin{aligned}
&E_p^*[(D, Z, X) \in A | R = 1] = E_p[(D, Z, X) \in A | R = 1] \text{ and} \\
&E_p^*[(D, Z, X) \in A | R = 0] = E_p[E_a[(D, Z, X) \in A | Z, X, R = 0] | R = 1],
\end{aligned}
$$

where E_p^* represents the corresponding expectation operators on matched targeting distribution and R an indicator that equals 1 for the primary sample and 0 otherwise. The first expression holds because the primary sample is our matched targeting distribution. The second expression, the distribution of (D, Z, X) in the auxiliary sample, is generated by integrating the conditional distribution of (D, Z, X) given Z, X over the distribution of Z, X in the primary sample.

Assumptions A1, A2, and A3 allow researchers to balance the primary and the auxiliary sample. To proceed, let $K = g(Z, X)$ be a $(k \times 1)$ vector of functions of (Z, X), and let $\tilde{\beta}$ be the vector of regression coefficients obtained from regressing D on K in the matched sample. The choice of K can be but is not limited to (Z, X). To ensure that matching is working, we also need to assume that conditional expectations are well-behaved and $H = E(KK')$ is invertible. Other assumptions can be found in Abadie and Imbens (2012).

The following Proposition A1 formalizes the idea that the first-stage estimates of the matched sample recover the parameters of the matching target distribution (i.e., the distribution of the primary sample).

Proposition A1. *Under regularity conditions, the regression coefficients ($\tilde{\beta}$) of D on K in the matched sample, P^*, are unbiased estimates of the analogous regression coefficients (β) in the population of the primary sample.*

Proof. We use the notation $D(0)$ to represent the latent D in the primary sample. Therefore, the regression coefficient in the primary (target) sample is defined by $E_p[(D(0) - K'b)^2]$.

$$
\begin{aligned}
E_p[(D(0) - K'b)^2] &= E_p[E_p[(D(0) - K'b)^2|Z, X]] && (6)\\
&= E_p[E_a[(D - K'b)^2|Z, X]] && (7)\\
&= E_p[E_a[(D - K'b)^2|Z, X, R = 0]|R = 1] && (8)\\
&= E^*[(D - K'b)^2|R = 0] && (9)
\end{aligned}
$$

The equality in (6) follows from the law of iterated expectations; the equality in (7) follows from propensity score equality (Assumption 3). Equations (8) and (9) follow from the definition of the matching target distribution. Until here, we have shown that matching under propensity score equality allows us to reproduce the first stage setting for the primary sample. Therefore, the regression coefficient in the primary sample is recovered using the matched sample.

To further establish the large sample property of the estimator, let $\tilde{\beta}$ be the vector of the sample regression coefficients obtained from regressing D on K in the matched sample,

$$\tilde{\beta} = \underset{b \in R^k}{\operatorname{argmin}} \frac{1}{N_1} \sum_{i \in P^*} (D - K'b)^2) = (\frac{1}{N_1} \sum_{i \in P^*} KK')^{-1} \frac{1}{N_1} \sum_{i \in P^*} KD. \tag{10}$$

From 6-9, the matching procedure makes sure that $\frac{1}{N_1} \sum_{i \in P^*} KK' \xrightarrow{p} H$. $H = E(KK')$ is the Hessian, which is invertible by assumption.

$$\tilde{\beta} - \beta = (\frac{1}{N_1} \sum_{i \in P^*} KK')^{-1} \frac{1}{N_1} \sum_{i \in P^*} (KD - KK'\beta) \xrightarrow{p} 0 \tag{11}$$

□

A.2 Nonlinear Models

Above, we used a linear model for the second step after balancing the primary and the auxiliary sample. The results also hold for more complex (nonlinear) models. For example, we can consider the following moment condition proposed in Graham, Pinto, and Egel (2016)

$$\mathbb{E}_p[\psi_p(Y; \beta) - \psi_a(D, Z_1; \beta) e(Z)] = 0 \tag{12}$$

with $Z = (Z_0', Z_1')'$. $E_p[\cdot]$ denotes expectations taken with respect to the primary population. β is the parameter of interest. There exist identification results for moment condition 12, when D and Y are observed in two different samples (Chen, Hong, Tarozzi et al., 2008). Note that both TSIV and TSTSLS methods can be seen as a special case of the moment condition 12. For example, we have the linear model of Angrist and Krueger (1992) if we take in 12 $e(Z) = Z$, $\psi_p(Y; \beta) = Y$ and $\psi_a(D, Z_1; \beta) = D'\gamma_1 + Z_1'\gamma_2$ with $\beta = (\gamma_1, \gamma_2')'$.

A.3 Simulation Results

We perform simulation results for the two classical methods (TSIV and TSTSLS), the Inverse Probabilty Tilting method (IPT), and the two-step-TSTSLS estimator proposed in this paper. In each of our experiments, we assume that X in both the primary sample and the auxiliary sample is distributed according to a truncated normal distribution, with support $[0, 2]$. The location and scale parameters of both distributions, (μ_p, ω_p^2) and (μ_a, ω_a^2), may differ. We assume a multinomial sampling scheme: with probability $Q_0 = 1/2$ a draw of (Y, Z, X) is taken at random from the population to constitute the primary sample; otherwise, a draw of (D, Z, X) is taken from the population to constitute the auxiliary sample. We set $\mu_p = 1.5$ and $\mu_a = 0.5$. We vary ω_p and ω_a to reflect the overlap between the primary sample and the auxiliary sample. In case 1, we take $\omega_p = \omega_a = 1$. Alternatively, in case 2, we take $\omega_p = \omega_a = 0.3$. In case 1, there is much overlap, which means in practice that the distribution of X does not differ too much in both samples. In case 2, there is little overlap, implying that the distribution of X in both samples differs a lot. Finally, we assume that Y and D are generated according to the following data generating process:

For the primary sample we generate data according to

$$Y = 0.5D + U, \tag{13}$$

and the endogenous variable D in the auxiliary sample is generated by

$$D = 0.5Z + \theta XZ + V, \tag{14}$$

where Z is distributed as $N(0, 1)$ and (U, V) are distributed independently of (Z, X) as

$$\begin{pmatrix} U \\ V \end{pmatrix} \sim N\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}. \tag{15}$$

For each simulation, we generated an i.i.d. sample of size $N_0 = 1000$ of (Y, Z, X) from the

population (the primary sample) and an i.i.d. sample of size $N_1 = 1000$ of (D, Z, X) from the population (the auxiliary sample). We then merge the two samples. With $\theta = 0$, the setting is simplified to the classical two sample models. With $\theta \neq 0$, we simulate misspecification.

Table A1 presents the results for four different scenarios. In scenario 1 the model is correctly specified, the overlap is good ($\omega_p = \omega_a = 1$) and there is no misspecification ($\theta = 0$). All four methods (see the first four rows) perform well with a very small bias and a small Root Mean Squared Error (RMSE). However, as expected, the two-step estimator performs worse than the other three methods on efficiency. The two-step-TSTSLS does not use the information of all data points, which results in a larger RMSE. In scenario 2, the overlap is good, but the model is misspecified ($\theta = 0.3$). The IPT estimator performs best with the smallest bias and RMSE. We next turn to situations where the overlap in the distribution of X in both samples is poor. In scenario 3, we take ($\theta = 0$) (i.e., the model is correctly specified), while in scenario 4, we assume that both the model is misspecified ($\theta = 0.3$) and the overlap is poor. In scenario 3, the two-step-TSTSLS estimator performs best on the bias, but the RMSE (like the IPT estimator) is less efficient. In scenario 4, all estimators are biased, but the two-step-TSTSLS estimator performs best.

In a second set of simulations, we vary the trade-off between efficiency (RMSE) and bias when the model is misspecified with a varying degree of misspecification. For these simulations we fix the overlap parameter $\omega_p = 0.3$ (bad overlap). We subsequently take 1,000 repeated simulations under four scenarios where we vary the degree of misspecification, i.e. we vary θ. In scenario 1, we take $\theta = 0$, i.e. the model is specified correctly. This scenario is equal to the scenario 3 of Table A1. In scenario 2 to 4, we gradually increase the degree of misspecification, with steps of 0.1, i.e. we take $\theta = 0.1, 0.2, 0.3$ for scenarios 2, 3, 4, respectively. We report the results of these simulations in Table A2.

In scenario 1 The bias of the TSIV, TSTSLS and the two-step-TSTSLS are similar and outperform the IPT estimator. However, the two-step-TSTSLS (and the IPT) is less efficient than the TSIV and TSTSLS estimators. When we gradually increase the degree

Table A1. Monte Carlo Results

	(1)	(2)	(3)	(4)
	N	Asym. Bias	Std dev.	RMSE
Scenario 1: Good overlap and correct specification				
TSIV	1000	0.049	0.515	0.515
TSTSLS	1000	0.007	0.522	0.519
IPT	1000	0.050	0.504	0.504
TWO-STEP-TSTSLS	1000	-0.057	2.954	2.940
Scenario 2: Good overlap and incorrect specification				
TSIV	1000	0.122	0.124	0.174
TSTSLS	1000	0.118	0.126	0.172
IPT	1000	-0.000	0.096	0.095
TWO-STEP-TSTSLS	1000	0.051	0.118	0.128
Scenario 3: Bad overlap and correct specification				
TSIV	1000	-0.054	0.380	0.382
TSTSLS	1000	-0.098	0.347	0.359
IPT	1000	-0.181	1.343	1.349
TWO-STEP-TSTSLS	1000	0.075	1.084	1.081
Scenario 4: Bad overlap and incorrect specification				
TSIV	1000	0.415	0.161	0.445
TSTSLS	1000	0.405	0.158	0.435
IPT	1000	0.403	3.164	3.173
TWO-STEP-TSTSLS	1000	0.147	0.140	0.203

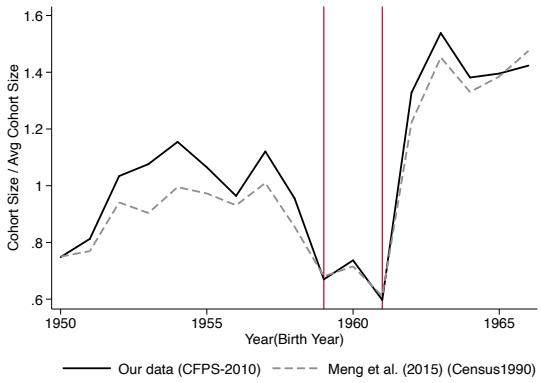
of misspecification (scenarios 2 to 4), the performance of the TSIV, the TSTSLS and the IPT estimators deteriorate quickly. The two-step-TSTSLS estimator performs very well in comparison with the other estimators. The IV estimate can be understood as the ratio of the intention to treat estimate and the first stage estimate. Therefore, misspecifications in the first stage regression translate in relatively large biases. Similarly, misspecifications in the first stage will increase the bias of the TSTSLS. The two-step-TSTSLS is robust against misspecifications and therefore we suggest that this method can be used as a robustness check in empirical applications.

Table A2. Additional Monte Carlo Results

	(1)	(2)	(3)	(4)
	N	Asym. Bias	Std dev.	RMSE
Scenario 1: Bad overlap and correct specification ($\theta = 0.0$)				
TSIV	1000	-0.054	0.380	0.382
TSTSLS	1000	-0.098	0.347	0.359
IPT	1000	-0.181	1.343	1.349
TWO-STEP-TSTSLS	1000	0.075	1.084	1.081
Scenario 2: Bad overlap and incorrect specification ($\theta = 0.1$)				
TSIV	1000	0.211	0.256	0.331
TSTSLS	1000	0.187	0.244	0.307
IPT	1000	-0.121	1.414	1.412
TWO-STEP-TSTSLS	1000	0.129	0.321	0.345
Scenario 3: Bad overlap and incorrect specification ($\theta = 0.2$)				
TSIV	1000	0.348	0.192	0.397
TSTSLS	1000	0.334	0.187	0.383
IPT	1000	-0.375	5.446	5.432
TWO-STEP-TSTSLS	1000	0.137	0.184	0.228
Scenario 4: Bad overlap and incorrect specification ($\theta = 0.3$)				
TSIV	1000	0.415	0.161	0.445
TSTSLS	1000	0.405	0.158	0.435
IPT	1000	0.403	3.164	3.173
TWO-STEP-TSTSLS	1000	0.147	0.140	0.203

Appendix B Additional Figures and Tables

Figure B1. Cohort Loss in CFPS



Notes: The figure compares the relative survivor birth cohort sizes in our data set (CFPS-2010, the solid line) with the relative cohort sizes in Meng, Qian, and Yared (2015) (Census1990, the dashed line).

Table B1. Effects on Separate Components

	(1)	(2)	(3)	(4)
		Components		
	Metabolic syndrome (index)	Diabetes	High blood pressure	Obesity
Panel A: Female				
Hunger before age 5	0.37***	0.032*	0.075*	0.063*
	(0.13)	(0.017)	(0.040)	(0.033)
Mother literate	0.029	0.0019	0.013	-0.0022
	(0.034)	(0.0057)	(0.0090)	(0.0080)
Age	0.47***	0.024	0.11**	0.096**
	(0.18)	(0.027)	(0.058)	(0.045)
Age squared(/100)	-0.43**	-0.022	-0.10*	-0.091**
	(0.17)	(0.026)	(0.056)	(0.043)
Observations	2517	2517	2517	2517
Panel B: Male				
Hunger before age 5	0.072	-0.041	-0.040	0.18***
	(0.23)	(0.031)	(0.069)	(0.058)
Mother literate	0.031	-0.0063*	0.0084	0.022*
	(0.028)	(0.0033)	(0.0070)	(0.011)
Age	0.14*	-0.0098	0.0089	0.097***
	(0.084)	(0.017)	(0.037)	(0.034)
Age squared(/100)	-0.13*	0.0084	-0.0054	-0.091***
	(0.081)	(0.016)	(0.035)	(0.032)
Observations	2612	2612	2612	2612

Notes: Each coefficient is from a separate regression. All regressions use the log(EDR) as the instrumental variable. The sample contains all individuals born between 1957 and 1962 in three waves of CFPS. Three components, diabetes, hypertension, and obesity, are dummy indicators constructed from CFPS. Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level respectively.

Table B2. Reduced-form Estimates at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
log(EDR)	0.031***	0.033***	0.0026	0.0036
	(0.011)	(0.011)	(0.012)	(0.012)
Mother literate		0.028		0.028
		(0.030)		(0.026)
Age		0.45**		0.14
		(0.17)		(0.085)
Age squared(/100)		-0.42**		-0.13
		(0.16)		(0.083)
Observations	2517	2517	2612	2612

Notes: The results are based on reduced-form estimates (3) from separate regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B3. Effects of Hunger at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Panel A: log(EDR) as the instrumental variable				
Hunger before age 5	0.37***	0.38***	0.045	0.062
	(0.14)	(0.13)	(0.20)	(0.21)
Mother literate		0.029		0.031
		(0.034)		(0.028)
Age		0.47***		0.14*
		(0.18)		(0.084)
Age squared(/100)		-0.43**		-0.13*
		(0.17)		(0.081)
Observations	2517	2517	2612	2612
Panel B: EDR as the instrumental variable				
Hunger before age 5	0.42**	0.41**	0.063	0.10
	(0.22)	(0.19)	(0.30)	(0.30)
Mother literate		0.029		0.032
		(0.035)		(0.029)
Age		0.47***		0.15*
		(0.18)		(0.083)
Age squared(/100)		-0.44***		-0.13*
		(0.17)		(0.080)
Observations	2517	2517	2612	2612

Notes: The results are based on TSIV estimates from separate regressions. All regressions are based on the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Panel A uses the log(EDR) as the instrumental variable. Panel B uses the EDR as the instrumental variable. Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B4. OLS Estimates at Age 0-5

	(1)	(2)	(3)	(4)
	Metabolic syndrome (index)			
	Female	Female	Male	Male
Panel A: Matched recall				
Hunger before age 5	-0.049	-0.048	0.039	0.041
	(0.035)	(0.034)	(0.032)	(0.032)
Mother literate		0.027		0.030
		(0.029)		(0.027)
Age		0.43***		0.14*
		(0.16)		(0.083)
Age squared(/100)		-0.40***		-0.13
		(0.15)		(0.080)
Observations	2517	2517	2612	2612
Panel B: Own recall				
Hunger before age 5	0.014	0.0096	0.0019	0.0020
	(0.045)	(0.044)	(0.033)	(0.034)
Mother literate		0.027		0.028
		(0.029)		(0.026)
Age		0.44***		0.14*
		(0.17)		(0.084)
Age squared(/100)		-0.40**		-0.13
		(0.16)		(0.081)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B5. Effects of Hunger at Age 0-5

	(1)	(2)	(3)	(4)
	Hospital Visits			
	Female	Female	Male	Male
Hunger before age 5	0.20*	0.20*	0.100	0.10
	(0.11)	(0.11)	(0.16)	(0.16)
Mother literate		0.0085		-0.024
		(0.020)		(0.016)
Age		0.083		0.13
		(0.13)		(0.093)
Age squared(/100)		-0.076		-0.12
		(0.12)		(0.088)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province based on matched bootstrap (Abadie and Spiess, 2019) with 999 replications appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.

Table B6. Metabolic Syndrome and Hospital Visits

	(1)	(2)	(3)	(4)
	Hospital Visits			
	Female	Female	Male	Male
Metabolic syndrome(index)	0.069***	0.071***	0.027	0.028
	(0.020)	(0.020)	(0.027)	(0.028)
Mother literate		0.012		-0.027*
		(0.018)		(0.015)
Age		0.047		0.13*
		(0.12)		(0.073)
Age squared(/100)		-0.042		-0.12*
		(0.12)		(0.069)
Observations	2517	2517	2612	2612

Notes: The results are based on simple OLS regressions. All regressions use the (matched) primary sample of individuals born between 1957 and 1962 from three waves of the China Family Panel Survey (CFPS). Standard errors clustered by province appear in parentheses. *, **, *** indicates significance at the 10%, 5% and 1% level, respectively.