Online Appendix

The Curse of Plenty: The Green Revolution and the Rise in Chronic Disease

Sheetal Sekhri and Gauri Kartini Shastry

A Data appendix

A.1 India Agriculture and Climate Dataset

The India Agriculture and Climate (IAC) dataset was compiled by Apurva Sanghi, K.S. Kavi Kumar, and James W. McKinsey, Jr. and draws on data assembled by James W. McKinsey, Jr. and Robert Evenson of the Yale Growth Center.

The database is a district-level panel from 1956 to 1987, covering 271 of the 334 districts in 13 major states of India. The dataset covers more than 85 percent of India by area; the only agriculturally important areas missing in the data are the southern state of Kerala and the northeastern state of Assam. The spatial boundaries correspond to 1961 boundaries; we take into account all splits in districts when merging this data with other datasets.

The agricultural data is comprehensive. It includes area planted and production for five major and fifteen minor crops, and irrigated area and area planted with high-yielding varieties (HYV) for the five major crops. Data on agricultural inputs including fertilizer, bullocks, and tractors (both quantity and price) are included. Other variables include agricultural labor, cultivators, wages and factory earnings, rural population, and literacy rate. In addition, the IAC includes district-specific climate data (average temperature and rainfall), compiled from data from meteorological stations using surface interpolation techniques, and information on soil quality and type.

The data draws on many national and international sources. As mentioned above, for agricultural data, the main source was the dataset created by James W. McKinsey, Jr. and Robert Evenson of the Yale Growth Center. The data was also cross-checked with government publications, such as the Agricultural Situation in India, Area and Production of Principal Crops in India, Agricultural Prices in India, Fertilizer Statistics (published by Fertilizer Association of India), and Statistical Abstracts of India. Climate data from over 160 meteorological stations are from the Food and Agricultural Organization (FAO) of the United Nations.

The IAC contains data on aquifer depth, which originate from the Water Resources Plates produced by the Government of India's National Atlas of India (WRP-NAI) and depicted in Appendix Figure B7. These maps show contours of water table depths. We used the WRP-NAI plates to confirm these variables. As shown in Figure 2, the IAC is missing data on aquifer depth for some districts. These are largely in the mountainous regions of Eastern India, Jammu and Kashmir and Himachal Pradesh in the North, and Kerala in the South. We collected aquifer depth data for these remaining districts using several publications of the Central Groundwater Board, Ministry of Water Resources, India, including the Report of the Groundwater Resource Estimation Committee 1984, 1997, and several groundwater district profile brochures.

A.2 Health outcomes from the IHDS and the NFHS-IV

In the IHDS, the chronic health-related questions are self-reported in response to the 'Education and Health Questionnaire.' The survey asks if a doctor has ever diagnosed an array of diseases, including cataracts, tuberculosis, high blood pressure, heart disease, diabetes, leprosy, cancer, asthma, polio, paralysis, epilepsy, and HIV. To construct the metabolic syndrome index, we make use of responses for three diet-related chronic

diseases: diabetes, heart disease, and high blood pressure for men. Since IHDS measures height and weight for women, we create an indicator for obesity and add it to the index for women. Kling et al. (2007) discuss the benefits of aggregating multiple measures in a given area (e.g., metabolic syndrome). Specifically, it improves statistical power and also addresses concerns about multiple hypothesis testing. We follow Hoynes et al. (2016) and construct the index by taking the simple average across standardized z-score measures of each component. The z-score is calculated by subtracting the mean and dividing by the standard deviation for individuals in cohorts born before 1966. Since all components are indicators for 'bad' health (diabetes, heart disease, high blood pressure, and obesity) an increase in metabolic syndrome index indicates worsening health.

The NFHS 2015-2016 collected biomarkers for respondents. This is the first NFHS with both district identifiers and biomarkers. Specifically, the survey recorded blood glucose levels and diastolic and systolic blood pressure. Blood glucose was measured using a finger-stick blood specimen using the FreeStyle Optium H glucometer with glucose test strips. Respondents were asked if they had eaten anything in the last 6 hours. We define diabetes based on charts developed by Vanessa D. Rozzario for MedIndia and the cutoffs identified in Somannavar et al. (2009). Specifically, these charts characterize blood glucose levels of 140 mg/dL or lower after having eaten recently and 90-100 mg/dL or lower on an empty stomach as normal. The analogous values identified by the American Diabetes Association are a bit higher. We follow the Indian data and define those who have eaten within the past 6 hours as diabetic if they have a blood sugar level of 140 mg/dL or more, and those who have not eaten within the past 6 hours as diabetic if they have a blood sugar level of 90 mg/dL or more. We define an individual as having high blood pressure if they have a systolic reading of 120 or greater or a diastolic reading of 80 or

greater.

A.2.1 Age heaping in the IHDS

As described in Section 4, age heaping creates noise in our measures of exposure to the Green Revolution. Appendix Figure B8 provides a histogram of birth years from the sample of men born between 1951 and 1981, illustrating the problem; respondents disproportionately report ages in multiples of five.[36]

In addition to informing our choice of specifications, as described in Section 4, we address the age heaping in a few ways. First, we confirm that age heaping is uncorrelated with being born in water-abundant districts after 1966 (Appendix Table B3). Second, we take advantage of a second round of the IHDS, where 83 percent of the same respondents reported their age in 2011-12. We exclude observations whose reported age in one of these waves suggests a pre/post classification that conflicts with the age reported in the other wave. We also exclude individuals who claim to have aged 3 or fewer years or 11 or more years in the 6-8 years between the two rounds, since this indicates a lack of attention. Finally, we note that if men born immediately after the Green Revolution (aged 38-39 in the IHDS) report an age of 40, our analysis will consider them not exposed to the Green Revolution, biasing our estimates towards zero.

A.3 Household food diaries

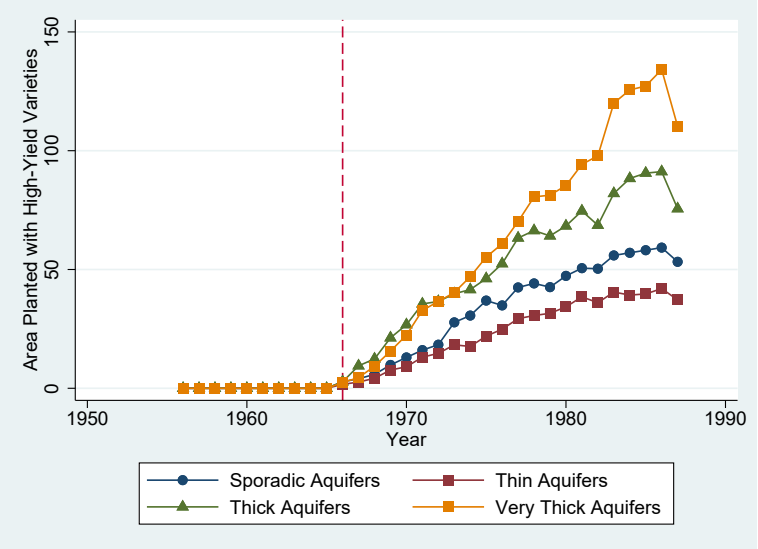
The data on calorie, protein, and fat consumption is computed from the Household Consumption Expenditure Surveys conducted by the National Sample Survey Organization (NSSO). The "thick" round of the survey is conducted every 5 years and is nationally representative. We use the

[36] Age heaping is not as severe a problem in the NFHS as seen in Appendix Figure B9, possibly because age is elicited in multiple ways and corroborated.

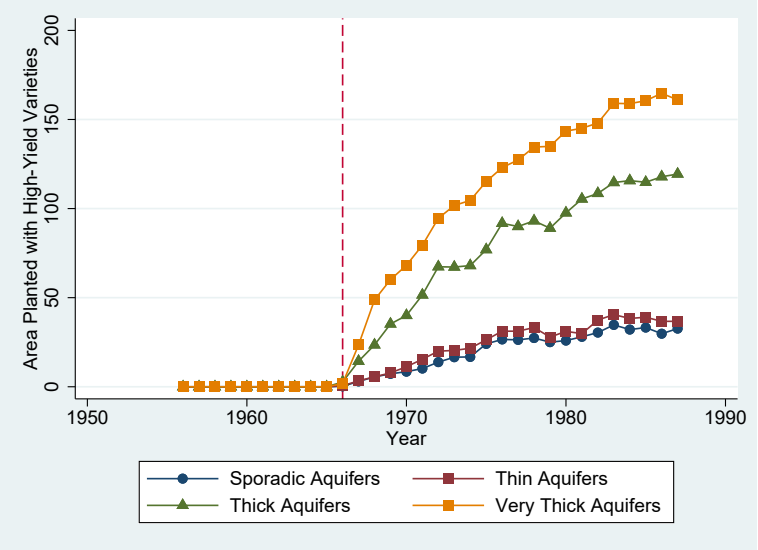
wave conducted in 1999-2000 (the 55th round), 5 years prior to the first wave of the IHDS. The survey measures all consumption, including consumption out of home-grown produce (imputed at producer prices) and out of in-kind wages, gifts, loans, free collections, all imputed at prevailing local retail prices. We use the reported quantities of specific food items consumed, converted into calorie intake based on standard food conversion tables, and aggregated across all food items.[37] Following Deaton and Drèze (2009), we use the "nutritive value of Indian foods" published by the National Institute of Nutrition (Gopalan et al. 1980) for the conversion of food items into caloric equivalents; the conversion factors have remained stable over time. We then follow Subramanian and Deaton (1996) and convert food intake into per-capita calorie, protein, and fat intake in adult equivalents. Information on consumption of food was collected independently for two different reference periods of 7 days and 30 days from the same households. We use the 30 day reference period to estimate daily consumption.

[37]The survey covers over 300 food items. The IHDS data also has information on food expenditures, but does not provide as exhaustive a list of items, making it difficult to calculate the number of calories consumed. We use food expenditures in the IHDS only to determine whether a household spends more on rice or wheat.

B Appendix figures and tables



(a) High-yield rice

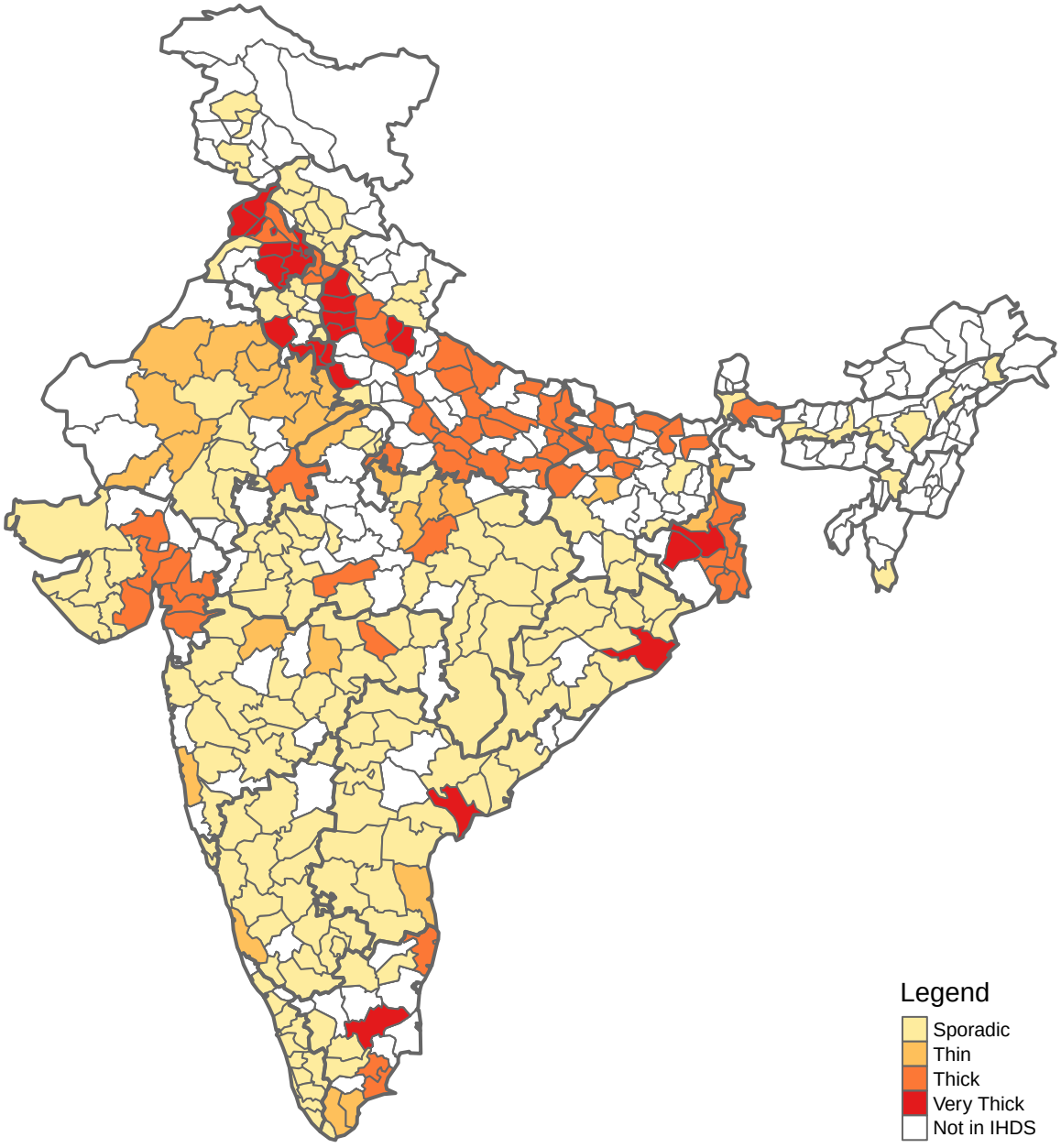


(b) High-yield wheat

FIGURE B1: Adoption of high-yield rice and wheat

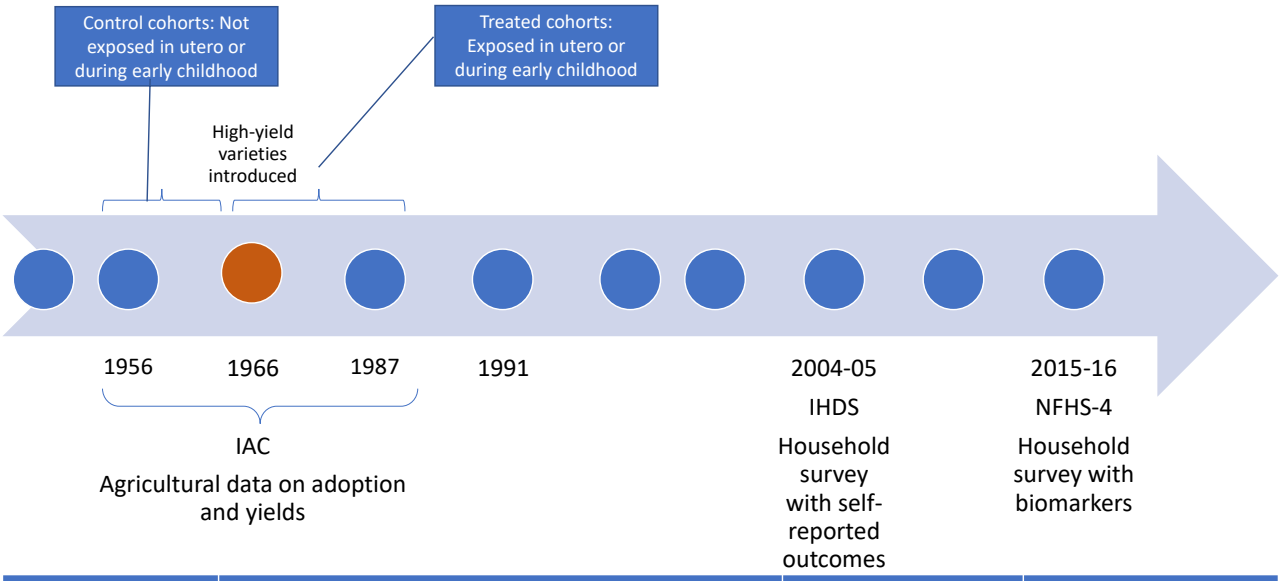
Notes: These figures graph the mean area planted with high-yield rice (Panel a) and wheat (Panel b) in 1000s of hectares separately for districts with sporadic aquifers (blue), thin aquifers (red), thick aquifers (green) and very thick aquifers (yellow).

FIGURE B2: Map of Indian districts in IHDS shaded by aquifer thickness



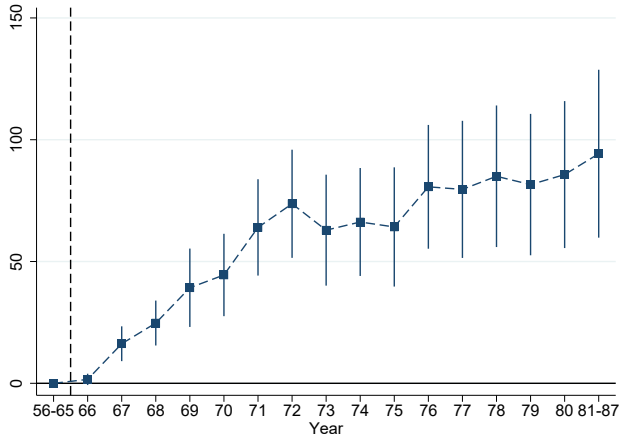
Notes: This map displays Indian districts in the IHDS shaded by aquifer thickness. Unshaded districts were not sampled in the IHDS. Thin borders indicate district boundaries while thick borders indicate state boundaries.

FIGURE B3: Timeline and age of first exposed cohorts

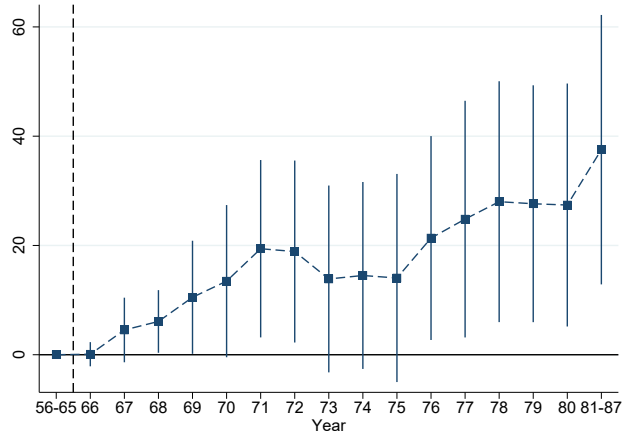


	1966	2004-05	2015-16
Age of first cohort exposed in-utero or during early childhood	Birth	38-39	49-50

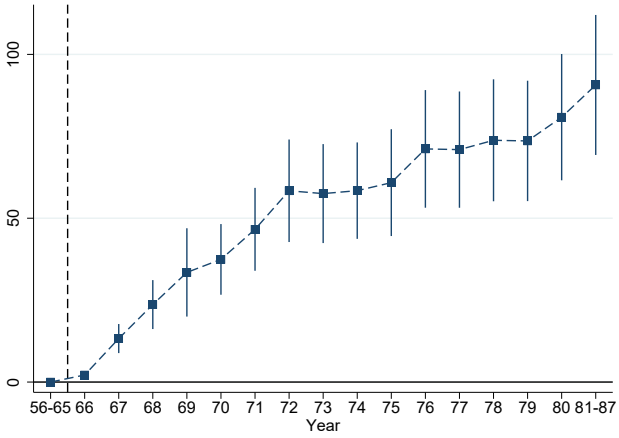
Notes: This figure depicts the timeline of events and survey data collection. The bottom panel indicates the age of the first cohort exposed to the Green Revolution in utero or during early childhood when each dataset was collected.



(a) High-yield varieties



(b) High-yield rice

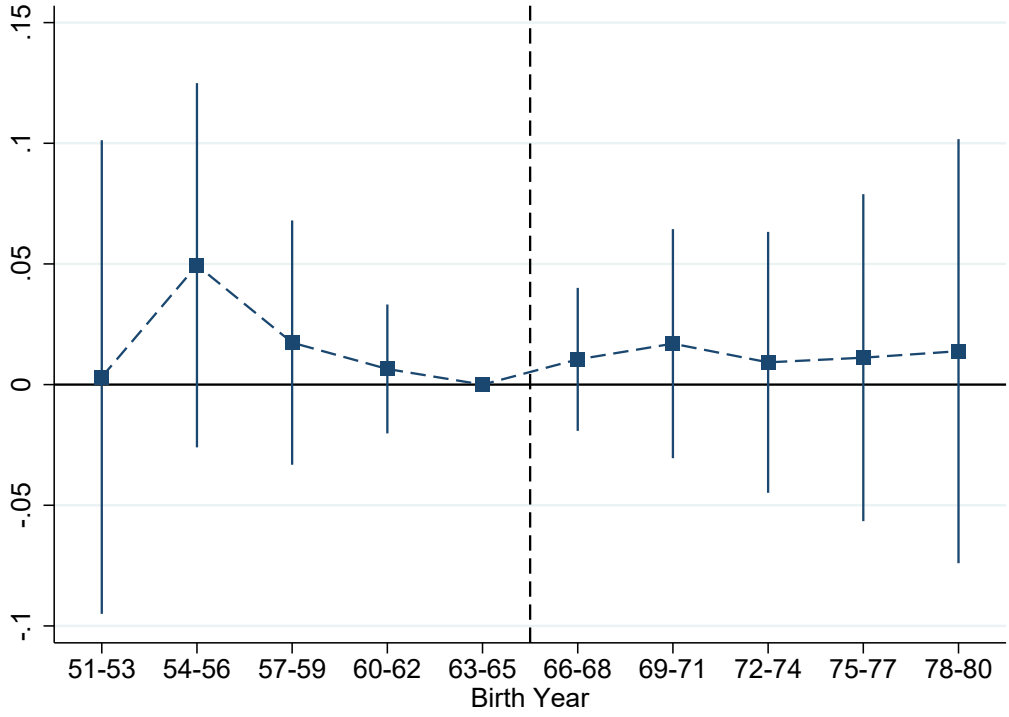


(c) High-yield wheat

FIGURE B4: Adoption of high-yield rice and wheat

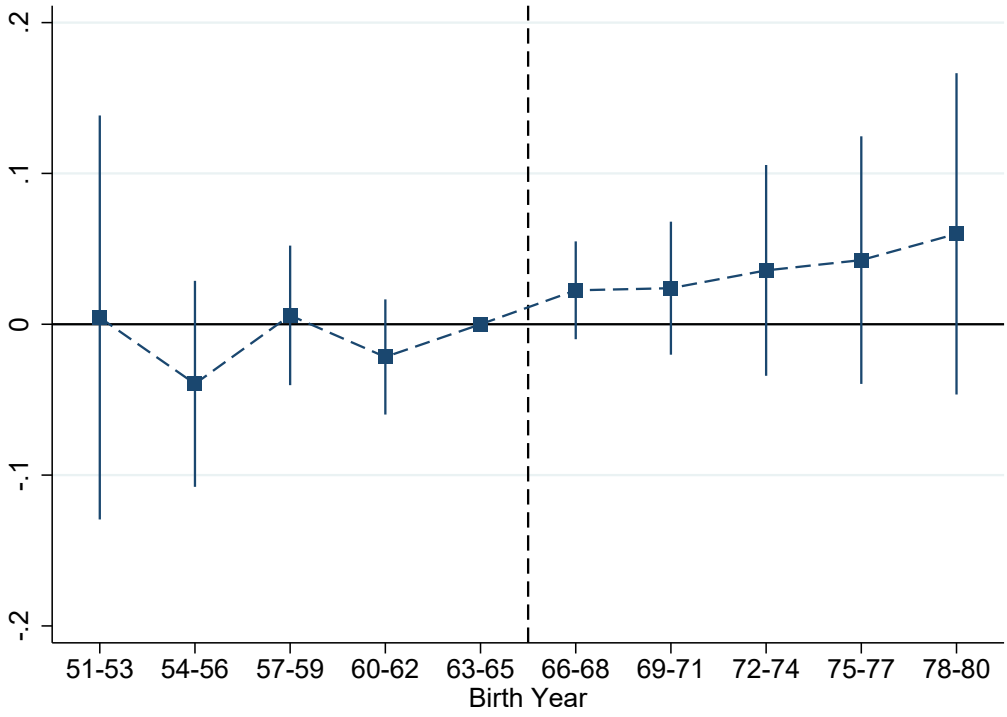
Notes: These figures plot the coefficients from estimating an event-study model using the area planted with all high-yield varieties (Panel a), high-yield rice (Panel b) and high-yield wheat (Panel c) in 1000s of hectares as the dependent variable. Since the outcome variable is 0 for all districts prior to 1966, we are unable to estimate separate coefficients for those years. The regression includes district and birth year fixed effects. Vertical bars indicate 95% confidence intervals.

FIGURE B5: Event-study estimates for impact on heart disease



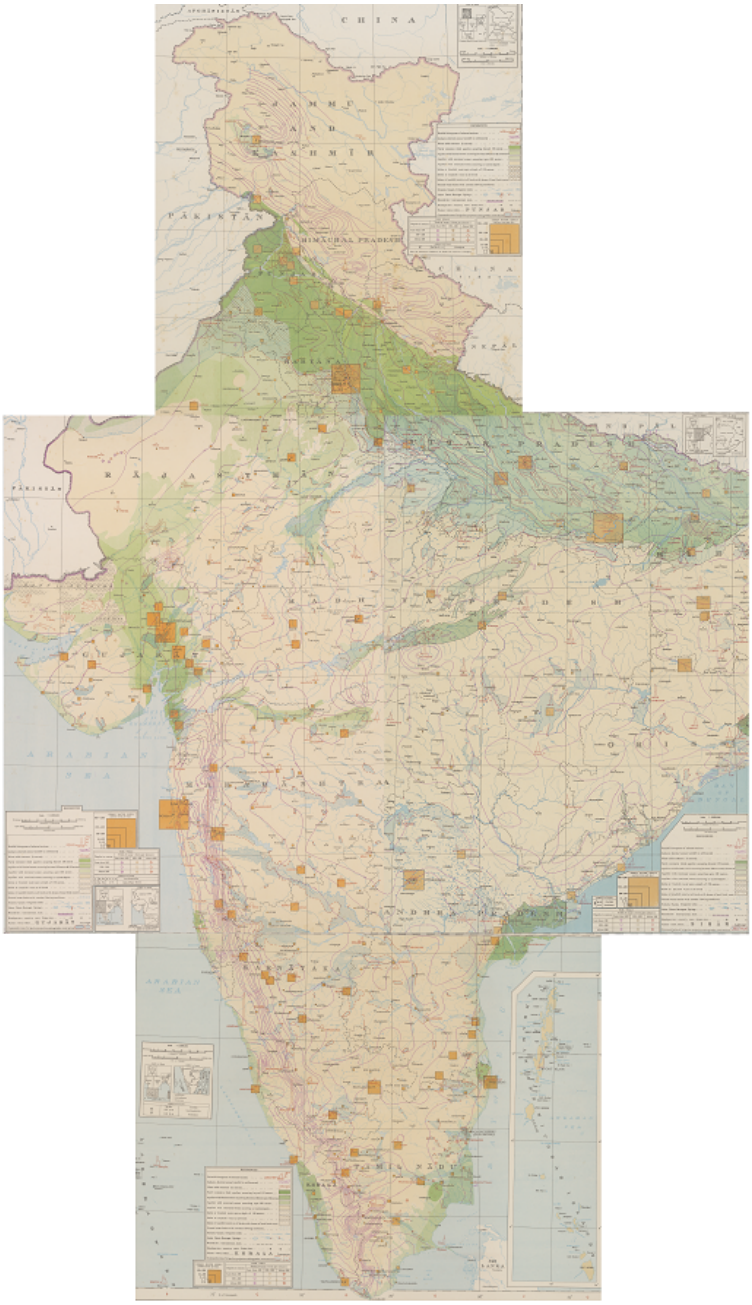
Notes: This figure plots the coefficients from estimating equation (4) using heart disease as the dependent variable. Those born between 1963 and 1965 are the omitted group. The regression includes district, birth year, and state X birth year fixed effects. It also includes district-specific trends and uses sampling weights. Vertical bars indicate 95% confidence intervals.

FIGURE B6: Event-study estimates for impact on high blood pressure



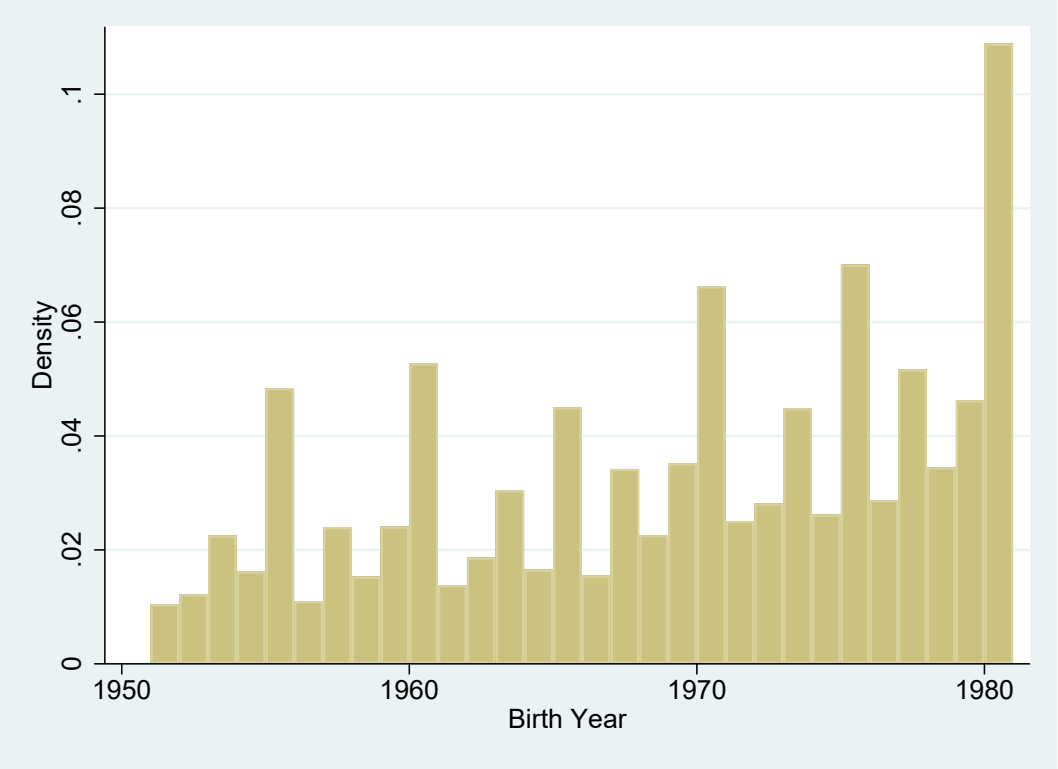
Notes: This figure plots the coefficients from estimating equation (4) using high blood pressure as the dependent variable. Those born between 1963 and 1965 are the omitted group. The regression includes district, birth year, and state X birth year fixed effects. It also includes district-specific trends and uses sampling weights. Vertical bars indicate 95% confidence intervals.

FIGURE B7: Water Resources Plates, National Atlas of India



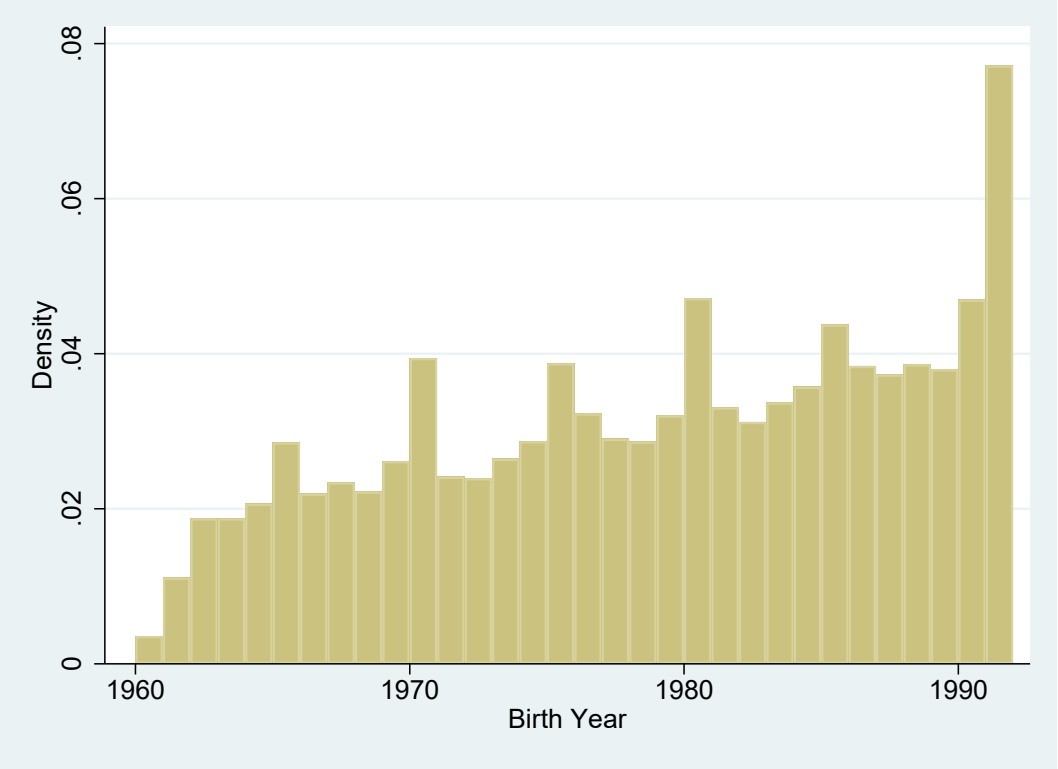
Source: National Atlas India.

FIGURE B8: Histogram of birth year in IHDS sample



Notes: This figure presents a histogram of birth year derived from reported age in the IHDS for those born between 1951 and 1981.

FIGURE B9: Histogram of birth year in NFHS sample



Notes: This figure presents a histogram of birth year derived from reported age in the NFHS for those born between 1951 and 1992. While there are still small spikes at ages divisible by 5, only the first spike, at 1966, would affect our results; our results are robust to excluding those born in 1966.

TABLE B1: Robustness of DID impact on long-term health

VARIABLES	(1) Metabolic syndrome index	(2) Metabolic syndrome index	(3) Diabetes	(4) Diabetes
Panel A: Cluster by state				
Born $\geq$ 1966 x groundwater-rich	0.089*	0.13	0.020**	0.039*
	(0.047)	(0.085)	(0.0090)	(0.021)
Observations	35,160	35,085	35,160	35,085
Panel B: Unweighted				
Born $\geq$ 1966 x groundwater-rich	0.037**	0.062	0.010**	0.022**
	(0.018)	(0.042)	(0.0045)	(0.010)
Observations	35,160	35,085	35,160	35,085
Panel C: Dropping 40-year-olds				
Born $\geq$ 1966 x groundwater-rich	0.076**	0.094*	0.013**	0.027*
	(0.038)	(0.057)	(0.0061)	(0.015)
Observations	33,210	33,133	33,210	33,133
Panel D: Household fixed effects				
Born $\geq$ 1966 x groundwater-rich	0.14**	0.11	0.058***	0.070**
	(0.063)	(0.15)	(0.017)	(0.034)
Observations	10,744	10,534	10,744	10,534
Panel E: District controls				
Born $\geq$ 1966 x groundwater-rich	0.093***	0.14**	0.023***	0.044***
	(0.034)	(0.064)	(0.0074)	(0.016)
Observations	35,160	35,085	35,160	35,085
Panel F: Individual controls				
Born $\geq$ 1966 x groundwater-rich	0.079**	0.12*	0.018***	0.037**
	(0.033)	(0.064)	(0.0063)	(0.016)
Observations	35,010	34,935	35,010	34,935

Notes: Each column presents the results from estimating equation (3) using a health outcome from the IHDS as the dependent variable. The specification is as in Table 5 Panel A (odd columns) or Panel B (even columns) with the following exceptions: Panel A clusters standard errors by state. Panel B does not use sample weights. Panel C drops those who report being exactly 40 years old. Panel D includes household fixed effects. Panel E includes district-level control variables and Panel F includes individual-level control variables. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B2: DID estimates for long-term health using different indexes

VARIABLES	(1) PCA	(2) PCA	(3) Factor	(4) Factor
Born $\geq$ 1966 x groundwater-rich	0.26**	0.41**	0.12**	0.19**
	(0.11)	(0.19)	(0.046)	(0.086)
Observations	35,160	35,085	35,160	35,085
R-squared	0.094	0.124	0.095	0.124
Kmo	0.536	0.536		

Notes: Each column presents the results from estimating equation (3) using alternative methods of aggregating the different health outcomes from the IHDS as the dependent variable. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district and birth year fixed effects. They also include district-specific trends and use sampling weights. Even columns also include state X birth year fixed effects. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table B3: DID estimates for impact on migration or age heaping

VARIABLES	(1) Migrated in	(2) Migrated in	(3) Migrated out	(4) Migrated out	(5) Reported age divisible by 5	(6) Reported age divisible by 5
Born $\geq$ 1966 x groundwater-rich	0.024	0.022	-0.018	-0.018	0.0070	0.0061
	(0.049)	(0.049)	(0.019)	(0.019)	(0.022)	(0.022)
x Diabetes		0.090		0.17		-0.010
		(0.17)		(0.10)		(0.044)
x Heart disease		0.088		-0.13		0.037
		(0.11)		(0.099)		(0.048)
x High blood pressure		-0.012		-0.016		0.0074
		(0.13)		(0.084)		(0.029)
Observations	35,040	35,040	33,640	33,639	35,086	35,085
R-squared	0.331	0.332	0.130	0.131	0.893	0.893

Notes: Each column presents the results from estimating equation (3) using indicators of migration or age heaping as the dependent variable. Having migrated in (Col 1-2) is defined as having a household head who has not lived in the village/town/city for their entire life. Having migrated out (Col 3-4) is defined as having migrated out of the current district between the 2005 and 2011 IHDS waves. Age heaping (Col 5-6) is defined as reporting an age divisible by 5. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district, birth year, and state X birth year fixed effects. They also include district-specific trends and use sampling weights. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B4: DID estimates for long-term health using the second wave of the IHDS

VARIABLES	(1) Metabolic syndrome index	(2) Diabetes	(3) Heart disease	(4) High blood pressure
Born $\geq$ 1966 x groundwater-rich	0.018	0.014	0.0024	-0.0044
	(0.035)	(0.020)	(0.0079)	(0.012)
Observations	38,443	38,443	38,443	38,443
R-squared	0.150	0.135	0.080	0.126

Notes: Each cell presents the results from estimating equation (3) using a health outcome from the second wave of the IHDS as the dependent variable. The dependent variable in Columns (2)-(4) indicate whether the individual was diagnosed with these conditions. The dependent variable in Column (1) is an index of metabolic health, averaging standardized z-scores of the three conditions. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district, birth year, and state X birth year fixed effects. They also include district-specific trends and use sampling weights. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B5: DID estimates for long-term health for women

VARIABLES	(1) Metabolic syndrome index	(2) Diabetes	(3) Heart disease	(4) High blood pressure	(5) Obese
Born $\geq$ 1966 x groundwater-rich	0.011	0.00035	0.0072	-0.013	0.016
	(0.049)	(0.0096)	(0.0100)	(0.020)	(0.024)
Observations	33,924	33,924	33,924	33,924	23,102
R-squared	0.118	0.086	0.077	0.105	0.096

Notes: Each column presents the results from estimating equation (4) using a health outcome from the IHDS as the dependent variable. The sample is restricted to women born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district, birth year, and state X birth year fixed effects. They also include district-specific trends and use sampling weights. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B6: DID estimates for impact on dietary habits

VARIABLES	(1) Rice-eating household
Born $\geq$ 1966 x groundwater-rich	-0.016
	(0.021)
Observations	35,086
R-squared	0.666

Notes: This table presents the results from estimating equation (3) using an indicator for rice-eating as the dependent variable. Rice-eating is defined from household expenditure data. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). The regression includes district, birth year, and state X birth year fixed effects. It also includes district-specific trends and uses sampling weights. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B7: Heterogeneity estimates for diabetes, with district X birth year fixed effects

VARIABLES	(1) All	(2) All	(3) All	(4) Rural Origins
Born $\geq$ 1966 x groundwater-rich				
x Rice-eating household	0.014			0.026
	(0.030)			(0.027)
x Rural		-0.055		
		(0.050)		
x Rural origin			0.026	
			(0.032)	
Observations	32,587	32,689	32,565	23,321
R-squared	0.318	0.347	0.355	0.317
Prob > F (joint)	0.63	0.28	0.42	0.35
Prob > F (main + interaction)	0.63	0.28	0.42	0.35

Note: Each column presents the results from estimating equation (3) using diabetes from the IHDS as the dependent variable. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district X birth year fixed effects, district-specific trends, and use sampling weights. All fixed effects and trends are also interacted with the relevant household characteristic. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B8: DID impact on calorie consumption from various food groups

VARIABLES	(1) Total	(2) Fat	(3) Protein	(4) Carbs (Total -Protein-Fat)
Difference-in-differences				
Born $\geq$ 1966 x groundwater-rich	76.6**	19.8	10.8**	46.0**
	(32.1)	(18.9)	(4.41)	(18.1)
Pre-trends (rel. to 1951-55 cohorts)				
Born $\geq$ 1956 x groundwater-rich	-58.9	-8.96	-9.21	-40.7
	(48.6)	(20.1)	(9.09)	(32.6)
Born $\geq$ 1961 x groundwater-rich	-19.5	-3.23	-3.27	-13.0
	(43.1)	(22.3)	(5.16)	(26.7)
Observations	138,741	138,741	138,741	138,741
R-squared	0.093	0.081	0.067	0.079

Notes: This table estimates regressions of daily calorie consumption in adult equivalents from various food groups on an indicator for being born after 1966 interacted with an indicator for water-abundant districts. Pre-trends are considered by including indicators for being born after 1956 and 1961, each interacted with water-abundant districts. The data is from the Household Consumption Expenditure Survey, collected in 1999-2000 by the NSSO. The sample is restricted to men born after 1951. All regressions include district, birth year, and state X birth decade fixed effects. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B9: Heterogeneity estimates for diabetes, by type of location

VARIABLES	(1) All	(2) All	(3) Rural Origins	(4) Rural Origins
Born $\geq$ 1966 x groundwater-rich	0.032	0.011	0.0045	0.0034
	(0.020)	(0.019)	(0.0046)	(0.0042)
x Rural	0.0092			
	(0.026)			
x Rural origin		0.037		
		(0.023)		
x Rice-eating household			0.045**	
			(0.021)	
x Rice-eating state				0.055
				(0.033)
Observations	34,976	34,864	25,658	25,834
R-squared	0.169	0.181	0.204	0.168
Prob > F (joint)	0.051	0.023	0.035	0.15
Prob > F (main + interaction)	0.039	0.0061	0.017	0.079

Notes: Each column presents the results from estimating equation (3) using diabetes from the IHDS as the dependent variable. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district, birth year, and state X birth year fixed effects. They also include district-specific trends and use sampling weights. All fixed effects and trends are also interacted with the relevant household characteristic. Standard errors clustered at the district level (Columns 1-3) and state level (Column 4) are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B10: Heterogeneity estimates for diabetes, controlling for fertilizer

VARIABLES	(1) All	(2) All	(3) All	(4) All	(5) All	(6) Rural Origins	(7) Rural Origins
Born $\geq$ 1966 x groundwater-rich	0.035**	0.0011	0.0012	0.014	-0.0027	0.0022	-0.00033
	(0.017)	(0.0047)	(0.0048)	(0.019)	(0.021)	(0.0053)	(0.0069)
x Rice-eating household		0.039**				0.044**	
		(0.020)				(0.021)	
x Rice-eating state			0.043				0.050
			(0.028)				(0.036)
x Rural				0.020			
				(0.025)			
x Rural origin					0.044*		
					(0.026)		
Amount of fertilizer used	-0.15	-0.068	-0.084	-0.16	-0.18	0.030	0.028
	(0.18)	(0.16)	(0.21)	(0.17)	(0.18)	(0.18)	(0.27)
Observations	24,245	24,141	24,245	24,241	24,166	18,052	18,169
R-squared	0.075	0.107	0.075	0.115	0.134	0.144	0.117
Prob > F (joint)		0.11	0.30	0.21	0.084	0.10	0.26
Prob > F (main + interaction)		0.037	0.17	0.087	0.029	0.033	0.23

Note: Each column presents the results from estimating equation (3) using diabetes from the IHDS as the dependent variable. The sample is restricted to men born between 1951 and 1981. Individuals with substantially discrepant reported ages across the two waves of the IHDS are excluded (see text for more details). All regressions include district, birth year, and state X birth year fixed effects. They also include district-specific trends and use sampling weights. All fixed effects and trends are also interacted with the relevant household characteristic. Standard errors clustered at the district level (Columns 1-2,4-6) and state level (Columns 2,7) are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B11: Impact estimates for cohort size in 2001

VARIABLES	(1) Total	(2) Total	(3) Rural	(4) Rural	(5) Urban	(6) Urban
Post x groundwater-rich	-3.23***	-1.10	-1.73***	-0.61	-1.51***	-0.50
	(0.58)	(0.79)	(0.36)	(0.47)	(0.37)	(0.49)
Observations	3,486	3,474	3,486	3,474	3,486	3,474
R-squared	0.997	0.998	0.995	0.997	0.998	0.999

Note: This table presents results from estimating equation (3) using population size (in 1000s) from the 2001 Census, by gender and 5-year age group, as the dependent variable. The sample is restricted to men born between 1952 and 1981. All columns include district and age group fixed effects and district-specific trends (by 5-year birth cohort). Even columns include state X age group fixed effects. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B12: Changes in the share planted with high-yield varieties

VARIABLES	(1) Share HYV
Post-1966 x groundwater-rich	0.079***
	(0.015)
Observations	8,615
R-squared	0.926

Note: This table presents the results from a difference-in-difference regression using the share of farmland planted with high-yield crop varieties as the dependent variable. All regressions include district and year fixed effects, district-specific trends, and controls for average annual rainfall and average annual temperature. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B13: Worst-case scenario DID estimates using trimmed population, random

VARIABLES	(1) Metabolic syndrome index	(2) Metabolic syndrome index	(3) Diabetes	(4) Diabetes
Born $\geq$ 1966 x groundwater-rich	0.077**	0.11*	0.015**	0.028**
	(0.039)	(0.062)	(0.0065)	(0.014)
Bootstrapped confidence interval	[0.050-0.084]	[0.078-0.127]	[0.011-0.015]	[0.015-0.030]
Observations	35,149	35,074	35,149	35,074
R-squared	0.091	0.123	0.088	0.122

Note: Each column presents the results from estimating equation (3) using a health outcome from the IHDS as the dependent variable. The specification is as in Table 5 Panel A (odd columns) or Panel B (even columns). The sample is trimmed to account for differential infant mortality, where a randomly chosen subset of individuals born after 1966 in water-abundant districts with non-zero values of the index or with diabetes are dropped. The number of individuals to be dropped is calculated from the estimates in Bharadwaj et al. (2020) and Table B12. Bootstrapped 95% confidence intervals, calculated over 100 iterations, are presented. *** p<0.01, ** p<0.05, * p<0.1.

TABLE B14: Worst-case scenario DID estimates using trimmed population, by income

VARIABLES	(1) Metabolic syndrome index	(2) Metabolic syndrome index	(3) Diabetes	(4) Diabetes
Born $\geq$ 1966 x groundwater-rich	0.065*	0.087	0.011**	0.015
	(0.038)	(0.055)	(0.0054)	(0.0094)
Observations	35,149	35,074	35,149	35,074
R-squared	0.092	0.124	0.088	0.123

Note: Each column presents the results from estimating equation (3) using a health outcome from the IHDS as the dependent variable. The specification is as in Table 5 Panel A (odd columns) or Panel B (even columns). The sample is trimmed to account for differential infant mortality, where the poorest individuals born after 1966 in water-abundant districts with non-zero values of the index or with diabetes are dropped. The number of individuals to be dropped is calculated from the estimates in Bharadwaj et al. (2020) and Table B12. *** p<0.01, ** p<0.05, * p<0.1.

C Agricultural production

In order to illustrate how the Green Revolution changed the production of various crops, we estimate an event study:

$$Y_{dt} = \Sigma_{l=1961}^{1964} \theta_l^{pre} (\tau_l * W_d) + \Sigma_{l=1966}^{1972} \theta_l^{post} (\tau_l * W_d) + \tau_d + \tau_t + \varepsilon_{dt} \quad \text{(C1)}$$

where Y_{dt} are crop yields in 1000s of tons. τ_d and τ_t are district and year fixed effects, respectively. W_d is an indicator for water-abundant districts and τ_l are a set of year indicators. The year 1965 (the year prior to the Green Revolution in India) is the reference year and we group together the years prior to 1961 and the years after 1972.

We plot the event-study coefficients (θ_l^{pre} and θ_l^{post}) for the production of rice and wheat over time along with 95 percent confidence intervals in Figure C1. Rice and wheat production exhibits a large significant increase after 1966 in water-abundant districts relative to water-scarce districts. However, this figure clearly illustrates the existence of statistically significant pre-trends.

Therefore, we utilize the synthetic difference-in-differences (SDID) method proposed by Arkhangelsky et al. (2021) to eliminate the bias arising from non-parallel pre-trends. The SDID method, in addition to aligning the pre-treatment trends in the outcomes with unit weights, also looks for time weights that balance pre-exposure time periods with post-exposure ones. The time weights can remove bias by eliminating the role of time periods that are very different from the post periods. Unit weights remove bias by focusing on similar units. This can also improve precision if there is heterogeneity in outcomes by units or time periods. We utilize the R package *synthdid* to estimate the district (unit) weights and the

time weights.[38]

Figure C2 demonstrates the SDID results: the blue line plots the production of rice and wheat for the water-abundant districts and the orange line plots the production of rice and wheat for a weighted average of the control districts. The two time series are parallel prior to 1966. The dashed line shows the time trend that treated units would have exhibited in the absence of treatment (which is parallel to the time series graph of the synthetic control units). The linear pivoted ray marks the treatment effect. Production of rice and wheat clearly increased post-Green Revolution.

We also investigate the impact of the Green Revolution on the production of two other important crops: pulses, a significant source of protein in India, and sugar, due to global trends in sugar consumption and the connection between sugar and diabetes. As described in the paper, since we assume rural markets were not fully integrated, crop production may provide insights into dietary changes. As seen in Panel (a) of Figure C3, the production of pulses fell in groundwater-rich districts in absolute terms, as well as relative to other districts. At the same time, sugar production was steadily increasing in groundwater-rich districts even before the Green Revolution and continued along the same trend afterward (Panel b).

We estimate event studies and use the SDID method for pulses and sugar production as well. Figure C4 plots the conventional event-study estimates (equation C1). The figure confirms the decline in the production of pulses in groundwater-rich districts relative to other districts after the Green Revolution. We also see no differential pre-trends. Nevertheless, we still apply the SDID method, estimate a new set of district and time weights, and illustrate the results in Figure C5. The analogous figures for sugar are C6 and C7. We see no evidence of an increase in sugar

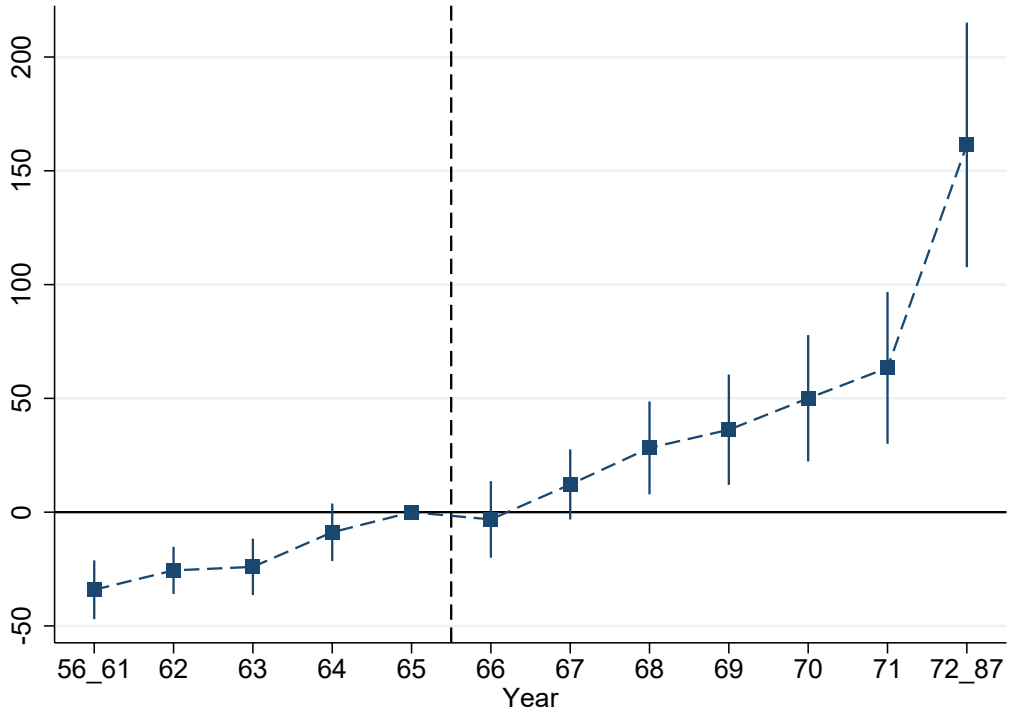
[38]The package developed by David A. Hirshberg can be found here: https://synth-inference.github.io/synthdid/

production in groundwater-rich districts; in fact, there is some evidence of a dip immediately after the Green Revolution.

Table C1 provides the regression results and standard errors from the SDID analysis. We see clear evidence that rice and wheat production rose significantly after the Green Revolution in water-abundant districts (Column 1), relative to water-scarce districts. We also find a significant decrease in the production of pulses (Column 2) and sugar (Column 3).

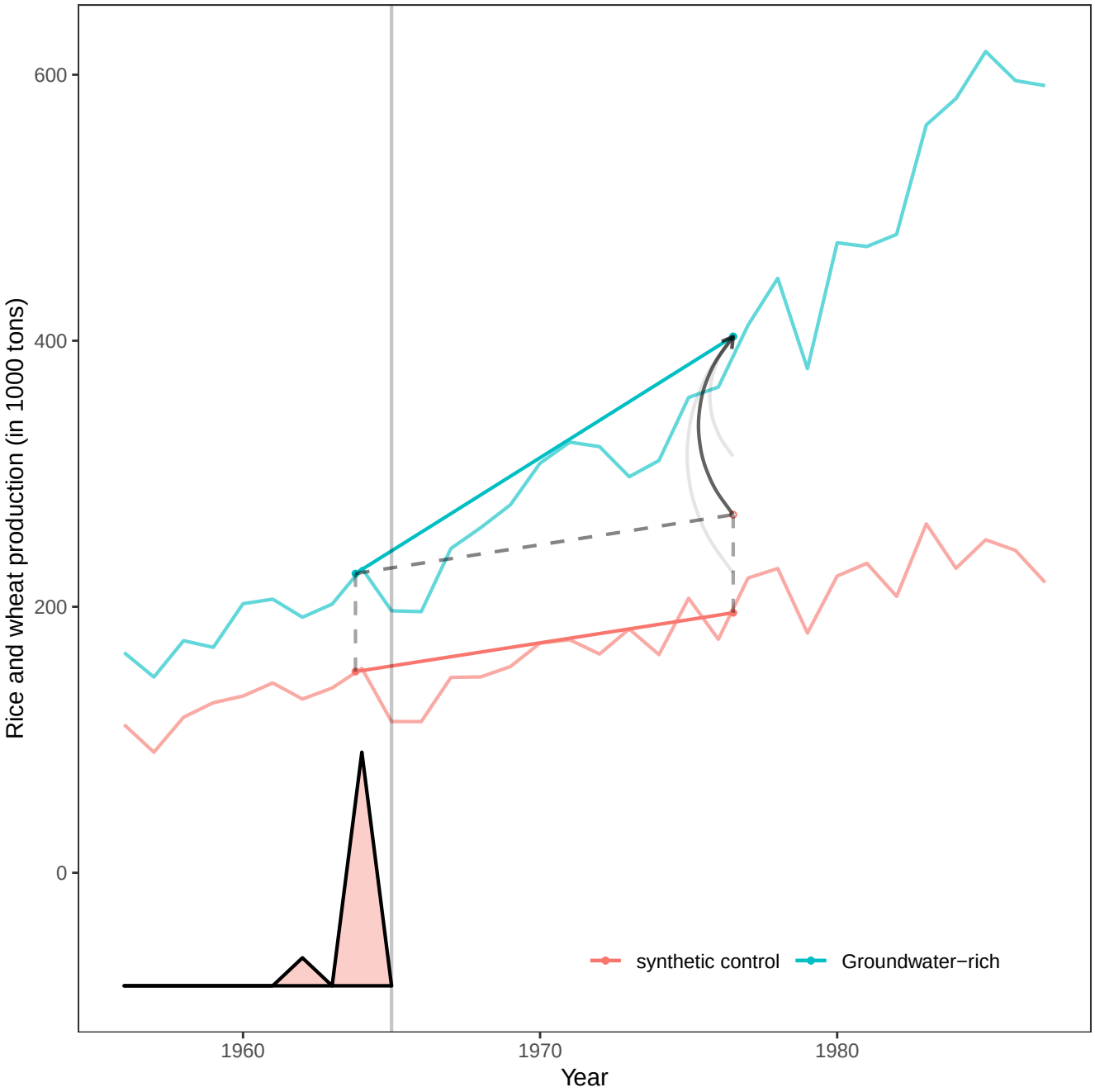
Finally, we illustrate the effect of the Green Revolution on diabetes using the SDID approach and district weights and get very similar results. Table C2 uses the district weights from the SDID analysis of rice and wheat production (Panel A), pulses production (Panel B), and sugar production (Panel C). We do not use the time weights since the SDID method only puts positive weights on two cohorts, reducing our statistical power when analyzing health outcomes significantly.

FIGURE C1: Event-study estimates for the impact on rice and wheat production

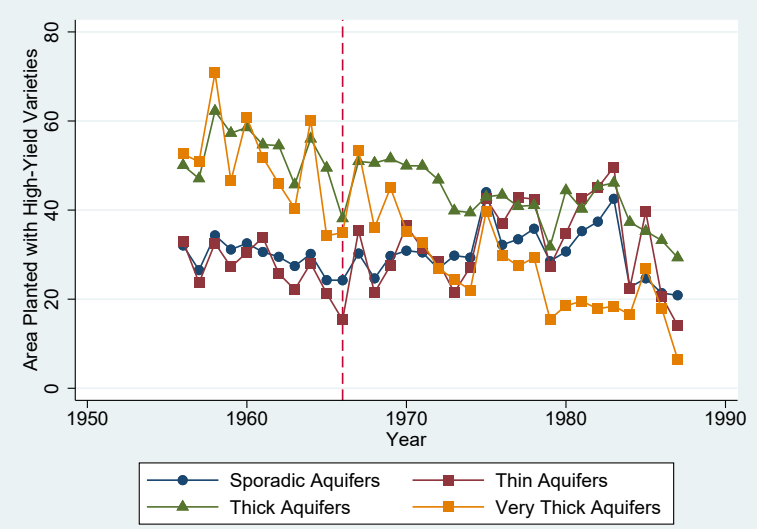


Notes: This figure plots the coefficients from estimating equation (C1) using the quantity of rice and wheat produced (in 1000 tons) as the dependent variable. 1965 is the omitted year. The regression includes district and year fixed effects. Vertical bars indicate 95% confidence intervals.

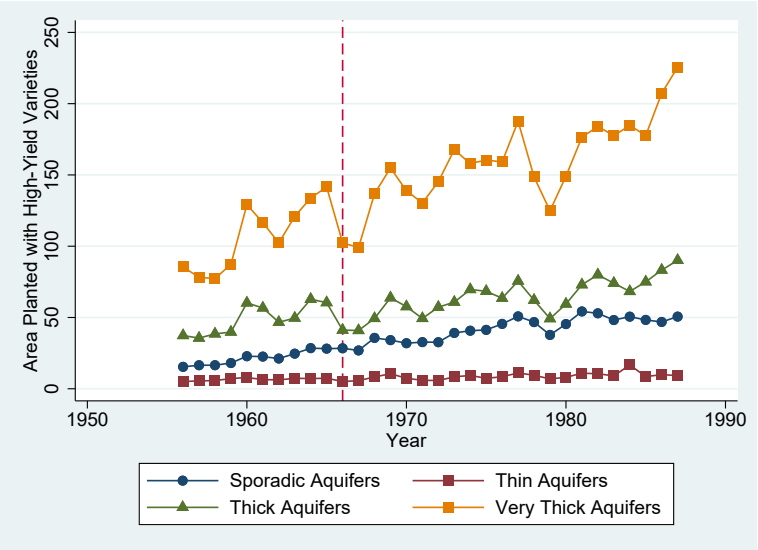
FIGURE C2: SDID estimates for the impact on rice and wheat production



Notes: This figure plots the evolution of the quantity of rice and wheat produced (in 1000 tons) in water-abundant districts and a synthetic control group of districts using the SDID method described in (Arkhangelsky et al. 2021).



(a) Pulses

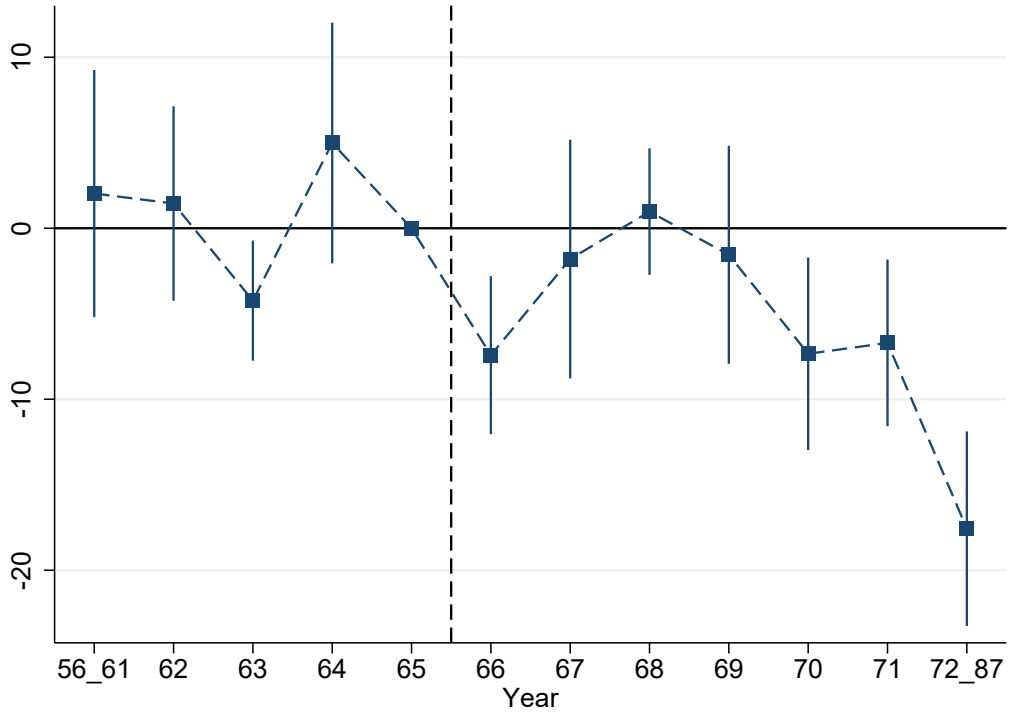


(b) Sugar

FIGURE C3: Production of pulses and sugar

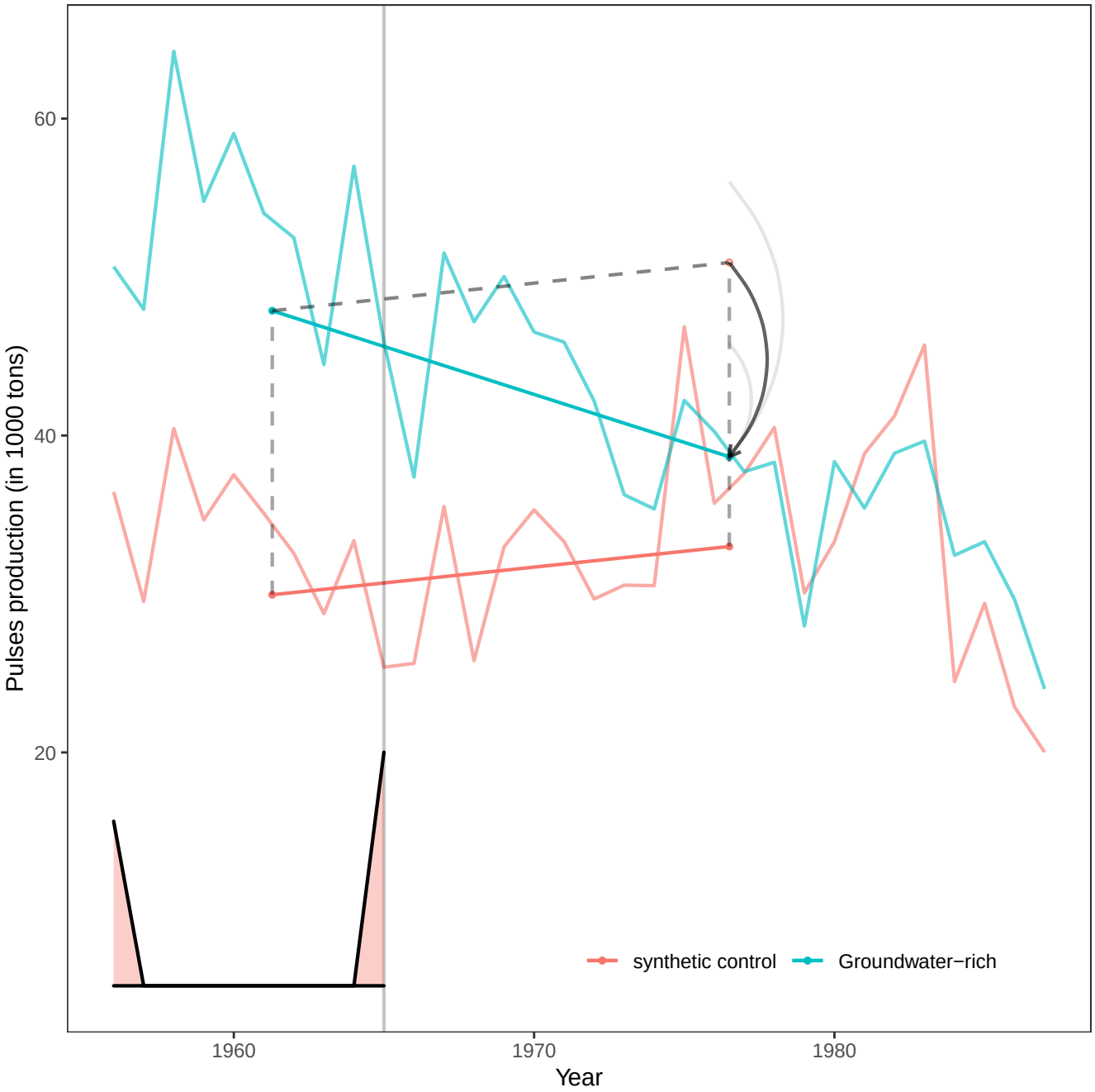
Notes: These figures graph the production of pulses (Panel a) and sugar (Panel b) in 1000s of tons separately for districts with sporadic aquifers (blue), just thick aquifers (red), medium thick aquifers (green) and the thickest aquifers (yellow).

FIGURE C4: Event-study estimates for the impact on pulses production



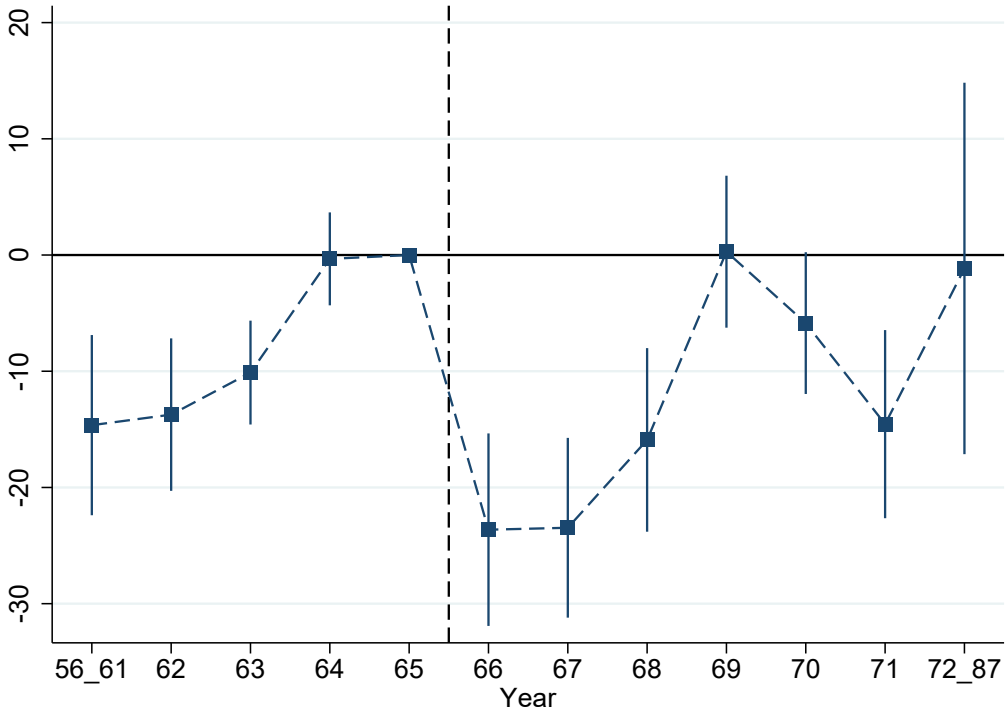
Notes: This figure plots the coefficients from estimating equation (C1) using the quantity of pulses produced (in 1000 tons) as the dependent variable. 1965 is the omitted year. The regression includes district and year fixed effects. Vertical bars indicate 95% confidence intervals.

FIGURE C5: SDID estimates for pulses production



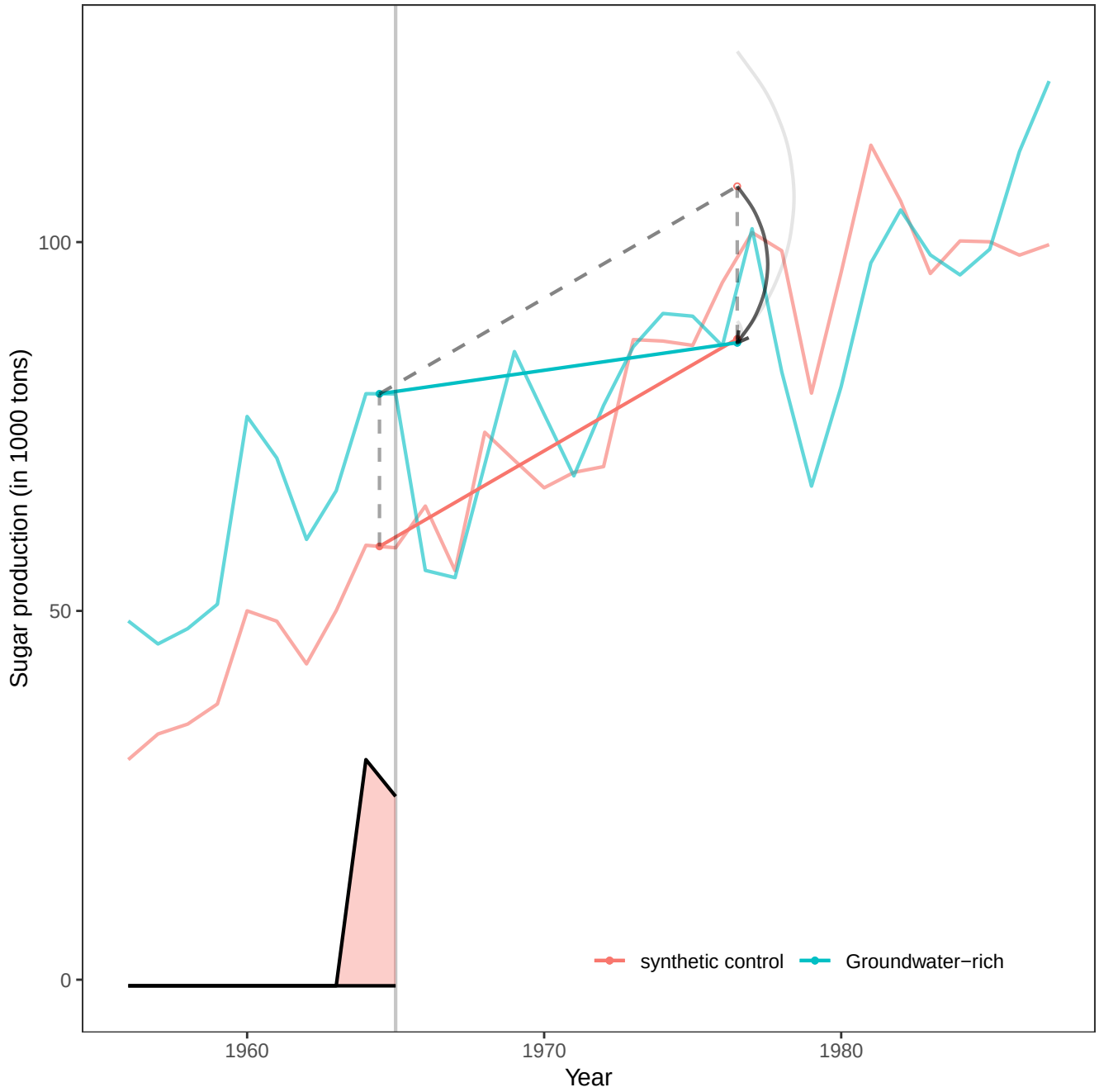
Notes: This figure plots the evolution of the quantity of pulses produced (in 1000 tons) in water-abundant districts and a synthetic control group of districts using the SDID method described in (Arkhangelsky et al. 2021).

FIGURE C6: Event-study estimates for the impact on sugar production



Notes: This figure plots the coefficients from estimating equation (C1) using the quantity of sugar produced (in 1000 tons) as the dependent variable. 1965 is the omitted year. The regression includes district and year fixed effects. Vertical bars indicate 95% confidence intervals.

FIGURE C7: SDID estimates for Sugar production



Notes: This figure plots the evolution of the quantity of sugar produced (in 1000 tons) in water-abundant districts and a synthetic control group of districts using the SDID method described in (Arkhangelsky et al. 2021).

TABLE C1: Synthetic DID estimates on agricultural production

VARIABLES	(1) Rice & wheat production	(2) Pulses production	(3) Sugar production
Post-1966 x groundwater-rich	133.95***	-12.26***	-21.21**
	(22.27)	(2.59)	(9.21)
Observations	6,480	6,480	6,480
R-squared	0.78	0.75	0.87

Note: Each cell presents the results from estimating equation (1) using an agricultural outcome as the dependent variable and the weights generated from the Synthetic DID method described in Arkhangelsky et al. (2021). All regressions include district and year fixed effects. Standard errors clustered at the district level are in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

TABLE C2: DID impact on long-term health using SDID weights

VARIABLES	(1) Metabolic syndrome index	(2) Metabolic syndrome index	(3) Diabetes	(4) Diabetes
Panel A: Rice/wheat production weights				
Born $\geq$ 1966 x med. or thickest	0.079**	0.10*	0.014**	0.037**
	(0.040)	(0.057)	(0.0056)	(0.016)
Observations	24,245	24,245	24,245	24,245
R-squared	0.066	0.108	0.049	0.073
Panel B: Pulses production weights				
Born $\geq$ 1966 x med. or thickest	0.078**	0.11*	0.014**	0.037**
	(0.039)	(0.056)	(0.0056)	(0.016)
Observations	24,245	24,245	24,245	24,245
R-squared	0.066	0.109	0.050	0.075
Panel C: Sugar production weights				
Born $\geq$ 1966 x med. or thickest	0.091**	0.13**	0.018***	0.042**
	(0.040)	(0.058)	(0.0059)	(0.019)
Observations	24,245	24,245	24,245	24,245
R-squared	0.070	0.113	0.045	0.069

Note: Each column presents the results from estimating equation (3) using a health outcome from the IHDS 2005 as the dependent variable. The specification is as in Table 5 Panel A (odd columns) or Panel B (even columns), except that Panel A uses weights from the SDID analysis of rice and wheat production and Panel B uses weights from the SDID analysis of pulses production. *** p<0.01, ** p<0.05, * p<0.1.