The Causal Relationship between Sanitation and Child Height

Effects at Household and Community Levels

Susanna Makela

May 11, 2017

1 Introduction

Height is an important measure of population health and human capital, with links to labor market outcomes (Case and Paxson, 2006; Vogl, 2012), mortality (Jousilahti et al., 2000), and cognitive achievement (Spears, 2012). While genetics play a major role in determining maximum height, early life nutrition and disease environment strongly influence the extent to which this height potential is achieved (JH, 1999). One major determinant of the disease environment, particularly in less-developed countries, is exposure to poor sanitation in the form of open defecation. India has one of the highest rates of open defecation, with over 60% of its rural population and 44% of the overall population lacking toilet or latrine facilities in 2015 (WHO/UNICEF JMP, 2015).

India also has high rates of stunting, with 38% of children in 2015 under the age of five being too short for their age (National Family Health Survey-4, 2016). Indeed, children in India are shorter, on average, than children in Africa, who are poorer, on average, and one proposed explanation for this "Asian enigma" is the high rates of open defecation in India (Spears, 2013). Recent evidence in the form of a randomized field experiment (Hammer and Spears, 2016) and an ecological analysis at the district level (Spears et al., 2013) suggests that there is a causal relationship between the sanitation environment at the neighborhood level and height-for-age among children in India. This study uses data from three household surveys to study the causal effects of open defecation at both the household and neighborhood levels in India using a sample of over XXX children from three rounds of a national household survey.

1.1 Causal questions

In this study, we examine the effects of open defecation (hereafter OD) at the household and neighborhood levels on height for age of Indian children aged three and younger. Specifically, we ask and answer the following questions:

1. What is the effect of living in a household with a latrine in neighborhoods where the OD rate is low?

2. What is the effect of living in a household with a latrine in neighborhoods where the OD rate is high?

3. What is the effect of a high versus low neighborhood-level OD rate on children who are likely to live in a household with a latrine?

Note that we use the terms OD and latrine use interchangeably; a household practicing OD is assumed to not use a latrine or toilet, and vice versa. The first three of the above questions closely parallel those asked by Hong and Raudenbush (2006) in a study of the effect of kindergarten retention on reading outcomes. Our analysis uses their analytical framework as described in Section 2 below.

1.2 Data

The data we use come from three rounds of the National Family Health Survey (NFHS). The NFHS is a cross-sectional survey conducted in 1992-93, 1998-99, and 2005-06 intended to yield estimates of indicators of maternal and child health, nutrition, and reproductive health valid for the rural and urban parts of Indian states. The NFHS includes

anthropometric measurements of height for children born to interviewed mothers in the three years prior to the survey, as well as numerous mother- and household-level covariates. While children are naturally nested within mothers and households, we ignore this level of nesting in our analysis. Because heights are measured only for children born in the three years prior to the survey, 90% of mothers and 89% of households in our sample have data for only one child. Conversely, 81% of the children in our sample are the only measured child of their mother and 89% are the only measured child from their household.

For the purposes of this study, we define "neighborhood" to be the primary sampling unit (PSU) to which each household belongs. In rural areas, the PSUs are generally villages of up to 500 households (larger villages are split in two), and in urban areas the PSUs are census enumeration blocks (CEBs) consisting of 150-200 households (technically, the urban PSUs are census wards in which the CEBs are nested, but since one CEB is selected per ward, we can think of the CEB as the PSU (Roy and Acharya, 2016)). The outcome variable is the height-for-age z-score: a child's height converted into a z-score relative to the 2006 WHO healthy reference population. Households reporting that their toilet facility is "no facility/bush/field" are considered to practice OD. To ensure adequate PSU-level sample sizes, we discard PSUs with fewer than 15 sampled households.

2 Methods

2.1 Causal inference framework

We use the potential outcomes framework of Rubin (1978) to define causal effects and the estimands. In a population of N children, $i = 1, \ldots, N$, we let z_i denote the treatment status of the household in which child i lives: $z_i = 1$ if the household practices OD and $z_i = 0$ if it does not. We let $\mathbf{z} = (z_1, \ldots, z_N)$ denote the vector of treatment assignments for all N children. Since the treatment is binary, $z_i \in \{0, 1\}$ and $\mathbf{z} \in \{0, 1\}^{2N}$. The potential outcome of child i, were their household's treatment status z_i and the other households' treatments $\mathbf{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N)$, is $Y_i(\mathbf{z})$.

A fundamental assumption that makes causal identification possible in the potential outcomes framework is the stable unit treatment value assumption (SUTVA). SUTVA states that there is a single version of the treatment and that the potential outcome of one unit is independent of the treatment assignment of the other units (Rubin, 1986). In other words, SUTVA allows us to simplify the definition of potential outcome and write $Y_i(\mathbf{z}) \equiv Y_i(z_i, \mathbf{z}_{-i}) = Y_i(z_i)$. In our context, SUTVA implies that the potential outcome for a child depends only on the sanitation status of their own household and not on that of their neighbors.

A key methodological challenge in this study is that SUTVA is an unreasonable assumption. Because OD by definition occurs outdoors, the practice of OD by one household necessarily exposes others in the neighborhood to possible contact with human waste. It is therefore entirely plausible that the sanitation behavior of the neighbors of a given household can affect the potential outcome of a child in that household, regardless of the sanitation status of that household.

2.2 Relaxing SUTVA

Our analytical strategy for relaxing STUVA is inspired by the approach taken by Hong and Raudenbush (2006), hereafter HR, in a study of the causal effect of kindergarten retention (i.e. repeating kindergarten) on reading outcomes, a context in which SUTVA is also an implausible assumption. Hong and Raudenbush (2006) handle this complication by assuming that the collective effects of $\mathbf{z}_{-i}$ on $Y_i(\mathbf{z})$ can be summarized by z_i and a scalar function $v(\mathbf{z})$, so that

$$Y_i(\mathbf{z}) \equiv Y_i(z_i, \mathbf{z}_{-i}) = Y_i(z_i, v(\mathbf{z})). \tag{1}$$

Following HR, we restrict $v(\mathbf{z})$ to be binary, $v(\mathbf{z}) = 0$ for neighborhoods with low OD rates and $v(\mathbf{z}) = 1$ for neighborhoods with high rates. Given two sets of possible treatment assignments for a single unit and all units, $(z, \mathbf{z})$ and $(z', \mathbf{z}')$, a generic causal estimand is of the form $\mathbb{E}[Y(z, \mathbf{z}) - Y(z', \mathbf{z}')]$. This formulation relies on three main assumptions as described by HR: a) causal inferences are generalized conditional on current neighborhood residence; b) there is no interference between neighborhoods, and c) treatment assignment at both the household and neighborhood levels are strongly ignorable conditional on observed covariates. See Appendix A for a more detailed description of these assumptions.

2.3 Estimable causal quantities

An important feature of HR's analysis of kindergarten retention rates is their division of children into subpopulations depending on their risks of retention under high and low school-level retention rates. They make this division because some potential outcomes are not defined for some children: some children may only be retained under a high school-level retention rate but not under a low one, while other may never be retained regardless of the school-level rate.

The identification of subpopulations is relevant in our context as well. In rural north India, notions of purity, pollution, caste, and religion play an integral role in deep-seated sanitation preferences that make ending OD particularly challenging (Coffey et al., 2014, 2015). OD rates in India as a whole are starkly different by both rural vs. urban residency and by religion. Data from NFHS-3 in 2005-06 show that 17% of urban households practiced OD, compared to 74% of rural households. The same data also show that nearly 60% of Hindu households practiced OD, while only 40% of the relatively poorer Muslim households did.

Let q_1 denote the probability of a household practicing OD under a high neighborhood OD rate and q_0 denote the probability under a low rate. Assume that $q_1 \geq q_0$ for all households. In parallel with HR, we identify three subpopulations of households (equivalently, children, as we ignore the clustering of children in households):

(A) Households at risk of OD under a low neighborhood OD rate ($q_1 \geq q_0 > 0$)

(B) Households at risk of OD under a high neighborhood OD rate but not a low rate ($q_1 \geq q_0 = 0$)

(C) Households at no risk even under a high neighborhood OD rate ($q_1 = q_0 = 0$)

See Table 1 of Appendix B for more details on the potential outcomes and causal effects defined for these subpopulations.

Because of the observational nature of our data, only certain causal quantities are estimable. Specifically, our estimands parallel those in HR and are as follows (see Appendix B for details). First, we estimate the conditional effect of household OD for children in subpopulation (A) who live in low-OD neighborhoods,

$$\delta_{Z0} = \mathbb{E}[Y_A(1,0) - Y_A(0,0) \mid V = 0, \mathbf{X}, \mathbf{W}], \tag{2}$$

where $\mathbf{X}$ and $\mathbf{W}$ are child- and neighborhood-level covariates, respectively. Second, we consider the union of subpopulations (A) and (B), which we denote (AB). Children in subpopulation (AB) are those who are ever at risk for living in households that practice OD ($q_1 > 0$). For this ever-at-risk subpopulation, we estimate the conditional effect of household OD in high-OD neighborhoods for children actually living in high-OD neighborhoods:

$$\delta_{Z1} = \mathbb{E}[Y_{AB}(1,1) - Y_{AB}(0,1) \mid V = 1, \mathbf{X}, \mathbf{W}]. \tag{3}$$

The third estimand is the average causal effect of high vs. low neighborhood OD level for children in subpopulation (C), that is, the effect of neighborhood OD for children at no risk of living in households that practice OD:

$$\delta_{V0} = \mathbb{E}[Y_C(1,1) - Y_C(0,1) \mid V = 0, \mathbf{W}]. \tag{4}$$

2.4 Propensity score models

Our first task is to identify which neighborhoods have high OD rates and which have low rates. Due to the large variation in OD rates over time and space, we use the average OD rate by survey round and rural-urban residency to define high and low OD neighborhoods. See Appendix C for the average rates.

Next, we model the propensity Q for a neighborhood to have a high OD rate given neighborhood covariates $\mathbf{W}$, $Q = \mathbb{P}(V = 1 \mid \mathbf{W})$. We use a hierarchical logistic regression that accounts for the nesting of neighborhoods within states:

$$\begin{aligned}\text{logit}(\mathbb{P}(V_j = 1 \mid \mathbf{W}) = \alpha_0 + \alpha_1 \mathbf{1}(NFHS2) + \alpha_2 \mathbf{1}(NFHS3) + \\ \beta_1 \mathbf{1}(rural_j) + \beta_2 assets_j + \beta_3 water_j + \beta_4 muslim_j + \beta_5 hindu_j + u_j \\ u_j \sim N(0, \sigma_u^2)\end{aligned}$$

3 Results

4 Discussion

See Appendix D for a discussion of the limitations of this study.

We have dichotomized neighborhood OD rates into low and high. Future research will allow the neighborhood rate to vary continuously from 0 to 1. This extension will require the use of the generalized propensity score (Hirano and Imbens, 2005), a generalization of the propensity score to nonbinary treatments to estimate the neighborhood OD propensity Q.

References

A. Case and C. Paxson. Stature and status: Height, ability, and labor market outcomes. Working Paper 12466, National Bureau of Economic Research, August 2006. URL http://www.nber.org/papers/w12466.

D. Coffey, A. Gupta, P. Hathi, N. Khurana, D. Spears, N. Srivastav, , and S. Vyas. Revealed preference for open defecation: evidence from a new survey in rural north india. Working Paper 1, research institute for compassionate economics, June 2014. URL http://riceinstitute.org/research/revealed-preference-for-open-defecation-evidence-

D. Coffey, A. Gupta, P. Hathi, D. Spears, N. Srivastav, and S. Vyas. Culture and health transition: Understanding sanitation behaviour in rural north india. Working paper, International Growth Centre, April 2015. URL http://www.theigc.org/wp-content/uploads/2015/04/Coffey-et-al-2015-Working-Paper.pdf

J. Hammer and D. Spears. Village sanitation and child health: Effects and external validity in a randomized field experiment in rural india. *Journal of Health Economics*, 48(C):135–148, 2016. URL http://EconPapers.repec.org/RePEc:eee:jhecon:v:48:y:2016:i:c:p:135-148.

P. Hathi, S. Haque, L. Pant, D. Coffey, and D. Spears. Place and child health: The interaction of population density and sanitation in developing countries. *Demography*, 54(1):337–360, 2017. ISSN 1533-7790. doi: 10.1007/s13524-016-0538-y. URL http://dx.doi.org/10.1007/s13524-016-0538-y.

K. Hirano and G. W. Imbens. *The Propensity Score with Continuous Treatments*, pages 73–84. John Wiley & Sons, Ltd, 2005. ISBN 9780470090459. doi: 10.1002/0470090456.ch7. URL http://dx.doi.org/10.1002/0470090456.ch7.

G. Hong and S. W. Raudenbush. Evaluating kindergarten retention policy. *Journal of the American Statistical Association*, 101 (475):901–910, 2006. doi: 10.1198/016214506000000447. URL http://dx.doi.org/10.1198/016214506000000447.

C. Jarquin, B. F. Arnold, F. Muñoz, B. Lopez, V. M. Cuéllar, A. Thornton, J. Patel, L. Reyes, S. L. Roy, J. P. Bryan, J. P. McCracken, and J. M. Colford. Population density, poor sanitation, and enteric infections in nueva santa rosa, guatemala. *The American Journal of Tropical Medicine and Hygiene*, 94(4):912–919, 2016. doi: https://doi.org/10.4269/ajtmh.15-0555. URL http://www.ajtmh.org/content/journals/10.4269/ajtmh.15-0555.

H. JH. Child undernutrition, tropical enteropathy, toilets, and handwashing. *Lancet*, 374 (9694):1032–1035, September 1999.

P. Jousilahti, J. Tuomilehto, E. Vartiainen, J. Eriksson, and P. Puska. Relation of adult height to cmause-specific and total mortality: a prospective follow-up study of 31,199 middle-aged men and women in finland. *American Journal of Epidemiology*, 151(11): 1112–1120, 2000.

National Family Health Survey-4. India fact sheet, 2016. URL http://rchiips.org/NFHS/pdf/NFHS4/India.pdf.

T. Roy and R. Acharya. *Statistical Survey Design and Evaluating Impact*. Cambridge University Press, 2016. ISBN 9781107146457. URL https://books.google.com/books?id=re4UjwEACAAJ.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *Ann. Statist.*, 6(1):34–58, 01 1978. doi: 10.1214/aos/1176344064. URL http://dx.doi.org/10.1214/aos/1176344064.

D. B. Rubin. Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

D. Spears. Height and cognitive achievement among indian children. *Economics & Human Biology*, 10(2):210–219, 2012. URL http://EconPapers.repec.org/RePEc:eee:ehbiol:v:10:y:2012:i:2:p:210-219.

D. Spears. How much international variation in child height can sanitation explain? Policy Research Working Paper Series 6351, The World Bank, 2013. URL http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-6351.

D. Spears, A. Ghosh, and O. Cumming. Open defecation and childhood stunting in india: An ecological analysis of new data from 112 districts. *PLOS ONE*, 8(9), 09 2013. doi: 10.1371/journal.pone.0073784. URL https://doi.org/10.1371/journal.pone.0073784.

T. Vogl. Height, skills, and labor market outcomes in mexico. Working Paper 18318, National Bureau of Economic Research, August 2012. URL http://www.nber.org/papers/w18318.

WHO/UNICEF JMP. Progress on sanitation and drinking water access – 2015 update and mdg assessment. Technical report, WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation, 2015. URL https://www.wssinfo.org.

Appendix A Causal assumptions

In this section, we provide details on the main assumptions underlying our causal framework as established by Hong and Raudenbush (2006).

Intact neighborhoods. The generic estimand above implicitly ignores the assignment of individuals to neighborhoods. If we were to take this assignment into account, we would include the vector of neighborhood assignments $\mathbf{s}$ in the potential outcome, where $\mathbf{s} = (s_1, \ldots, s_N)$ and $s_i \in \{1, \ldots, J\}$ denotes which of the J neighborhoods unit i is assigned to. The estimand would then be

$$\mathbb{E}[Y(z, \mathbf{z}, \mathbf{s}) - Y(z', \mathbf{z}', \mathbf{s}')],$$

where $\mathbf{s}$ and $\mathbf{s}'$ are two vectors of possible neighborhood assignments. However, we know that in the real world, people are not assigned to neighborhoods at random, and we wish to estimate causal effects conditional on actual neighborhood membership. We therefore condition our estimands on the observed neighborhood assignment $\mathbf{s}^*$:

$$\mathbb{E}[Y(z, \mathbf{z}, \mathbf{s}^*) - Y(z', \mathbf{z}', \mathbf{s}^*) \mid \mathbf{S} = \mathbf{s}^*].$$

For clarity of notation, we consider all estimands to be conditional on $\mathbf{S} = \mathbf{s}^*$ from this point forward and do not explicitly write out this conditioning.

No interference between neighborhoods. We assume that the potential outcome for a child can be affected by the treatment assignments of other children in the same neighborhood, but not by the treatment assignments of children from other neighborhoods. We therefore write $\mathbf{z}_j = (z_{1j}, \ldots, z_{n_j j})$ for the vector of treatment assignments for the n_j children in neighborhood j and $\mathbf{s}_j = (j, \ldots, j)$ for the corresponding neighborhood assignments. The potential outcome for a child is then

$$Y_i(z_i, v(\mathbf{z}), \mathbf{s}^*) = Y_i(z_i, v(\mathbf{v}_{j[i]}), \mathbf{s}^*_{j[i]}) \equiv Y_i(z_i, v),$$

where the notation $j[i]$ refers to the neighborhood j that unit i belongs to. We let $V = v(\mathbf{Z})$ denote the random variable taking on values $v(\mathbf{z}) = v \in [0, 1]$.

Strongly ignorable treatment assignment. We assume that the assignments of intact neighborhoods to open defecation level $V = v$ and children to household latrine use status $Z = z$ are ignorable conditional on observed neighborhood-level covariates $\mathbf{W}$ and child- and household-level covariates $\mathbf{X}$. Under this assumption, we can write

$$\mathbb{E}[Y(z, v) \mid Z = z, V = v, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] = \mathbb{E}[Y(z, v) \mid \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}],$$

allowing us to estimate the conditional average causal effect

$$\mathbb{E}[Y(z, v) \mid \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(z', v') \mid \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}]$$

with the observed quantity

$$\mathbb{E}[Y(z, v) \mid Z = z, V = v, \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}] - \mathbb{E}[Y(z', v') \mid Z = z', V = v', \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}]$$

Together, the assumptions of intact neighborhoods, no interference between neighborhoods, and strong ignorability allow us to conceive of our observational data as approximating a two-stage experiment. In the first stage, neighborhoods are randomly assigned an open defecation rate within blocks defined by $\mathbf{W}$. In the second stage, households within neighborhoods are randomly assigned to binary latrine use within blocks defined by $\mathbf{X}$ and $\mathbf{W}$.

Appendix B Causal estimands and strong ignorability

Table 1 displays the potential outcomes and causal effects of interest for subpopulations of households defined by their probabilities of OD under high and low neighborhood-level OD rates. This table parallels Table 2 of HR.

Subpopulation	Probabilities of OD	Potential outcomes	Causal effects of interest
(A)	$q_1 \geq q_0 > 0$	$Y(1,1), Y(0,1), Y(1,0), Y(0,0)$	$\mathbb{E}[Y(1,1) - Y(0,1)]$
			$\mathbb{E}[Y(1,0) - Y(0,0)]$
(B)	$q_1 \geq q_0 = 0$	$Y(1,1), Y(0,1)$	$\mathbb{E}[Y(1,1) - Y(0,1)]$
(C)	$q_1 = q_0 = 0$	$Y(0,1), Y(0,0)$	$\mathbb{E}[Y(0,1) - Y(0,0)]$

Table 1: Potential outcomes and causal effects for subpopulations of households

To understand which causal effects are estimable under strong ignorability, we imagine that our observational data are in fact from two-stage randomized experiment. Neighborhoods are first randomized to a binary OD level $V = v \in [0, 1]$ within blocks defined by neighborhood covariates $\mathbf{W}$ with probability $Q = Q(\mathbf{W})$:

$$Q = \mathbb{P}(V = v \mid \mathbf{W}). \tag{5}$$

Given the neighborhood OD level, households with neighborhoods are randomly assigned an OD status, $Z = 1$ for practicing OD and $Z = 0$ for latrine use, within blocks defined by $\mathbf{X}$ and $\mathbf{W}$. The household-level probability of $Z = 1$ is then $q_v = q_v(\mathbf{X}, \mathbf{W})$, where

$$q_v = \mathbb{P}(Z = 1 \mid V = v, \mathbf{X}, \mathbf{W}), \tag{6}$$

and a child's probability of living in a household with OD status $Z = z$ and neighborhood with OD level $V = v$ is then

$$\mathbb{P}(Z = z, V = v \mid \mathbf{X}, \mathbf{W}) = \mathbb{P}(Z = z \mid V = v, \mathbf{X}, \mathbf{W})\mathbb{P}(V = v \mid \mathbf{W}). \tag{7}$$

Under the two-stage randomized experiment, we have the following

Appendix C OD rates by survey round and year

Table 2 displays the OD rates by survey round and rural-urban residence, with rates for all India shown for reference. NFHS-1 was conducted in 1992-93, NFHS-2 in 1998-99, and NFHS-3 in 2005-06.

	NFHS-1	NFHS-2	NFHS-3
Rural	87.1	80.8	74.0
Urban	24.1	19.2	16.8
All India	69.7	63.7	55.4

Table 2: OD rates (%) by survey round and rural-urban residency.

Appendix D Limitations

One limitation of the present study is that we have ignored the clustering of children within households, thus implicitly treating children in the same community as independent. Children within the same household are likely to be more similar than children in different households, so if the data for a community has many children clustered in only a few households, they are less informative than they would be for a community with the same number of children with one child per household. However, 89% of the children in our data belong to unique households and 81% to unique mothers.

We also do not know the geographic size of the PSUs. While the DHS manual tells us that urban PSUs are generally census enumeration blocks (roughly 150-200 households) and rural PSUs are villages of 500 or fewer households, there is likely to be variation in the geographic size and population density of the PSUs. In that case, our results are likely to be lower bounds, as the median sample size per PSU is 29 households (interquartile range 22 to 33). A recent study using both cross-national data from 172 countries and

subnational data from Bangladesh suggests that the negative effects of neighborhood-level open defecation are magnified by high population density (Hathi et al., 2017). However, an epidemiological study in Guatemala did not find that population density was an important determinant of intestinal infections or a strong effect modifier for household-level sanitation (Jarquin et al., 2016), so evidence remains mixed.