# Causal Inference and Indian Undernutrition

**Saul Richardson** saulrichardson44@gmail.com
*February 3, 2023* *Williams College*

## 1. Introduction

The "Asian Enigma" refers to a conundrum in South Asia where children perform worse on nutritional indicators relative to children in Sub-Saharan Africa despite living in economies of equal, if not larger, size. The country at the focal point of child undernutrition is India. Despite the country's GDP quadrupling from 470 billion to 2.1 trillion USD between 2000 and 2015 (World Bank), Indian children still experience below-average nutritional status. Nutritional health outcomes can be measured via Height-for-Age and Weight-for-Age Z-scores. These z-scores indicate how many standard deviations the height or weight of a child for a given age is away from the mean of a reference population. The reference population used comes from the World Health Organization's growth standards (Spears and Behrman, 2019). For example, the average height-for-age z-score for Indian children in the 2015-16 National Family Health Survey was $-1.53$, and the average height-for-age z-score was $-1.48$ (International Institute for Population Sciences, 2015-16)—meaning that Indian children have weight and height measurements well below what is expected of a healthy child. To address these deficits, the Indian government has invested vast sums of money into supplemental nutrition programs over the last few decades, the primary being Integrative Child Development Services (ICDS). The program provides supplemental nutrition to any child aged zero to six years old attending one of 1.3 million anganwadi centers (preschools) across India. However, the program has experienced infrastructure issues, and attempts to quantify the impact of ICDS centers on attendees have been limited, often only pointing to broad trends in the data without ascertaining causality (Sood, 2008). Many of these attempts don't attempt causality because the only representative, national dataset with information on child health and ICDS take-up is observational, making it difficult to construct a true counterfactual. In this paper, I use causal inference methods in an attempt to identify the average causal effect of attending an ICDS center using the 2015-16 round of the National Family Health Survey International Institute for Population Sciences (2015-16). I find that there is little to no evidence that attending an ICDS center has a positive impact on the Indian children nutritional outcomes. To provide robustness, I use a mix of linear and random forest implementations of augmented inverse probability weighting and backdoor adjustment to confirm the absence of any positive impact of ICDS centers on child beneficiaries.

## 2. Preliminaries

One way around the issues with observational data, such as negative selection bias, is Augmented Inverse Probability Weighting (AIPW). Giving unlikely responses higher weights, we can produce an appropriate counterfactual to compute the average causal effect (Horvitz, 1952). A benefit to using AIPW specifically is its double robustness against misspecification. We may have a lack of confidence in our ability to correctly model the outcome. But, we also don't know with full confidence whether our model for the propensity score used to reweigh the data is correctly specified. AIPW alleviates this by "augmenting" the regular IPW to include the outcome regression. AIPW converges to the truth if either model is correctly, or close to being correctly, specified (Glynn, 2009). To determine these weights, we need a propensity score predicting the probability of a certain treatment given a valid backdoor adjustment set. Finding such a set relies on a DAG (Directed Acyclic Graph) that encodes potential causal relations between variables. Each variable can be interpreted as a function of its parents, allowing us to use $p(V) = \prod_{V_i \in V} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i))$ to describe the distribution of $p(V)$ that factorizes according to the DAG $G$. From the graph, we can characterize causal relations between nodes. A directed edge implies a variable is potentially a cause of another, and we can use d-separation to see if some set $Z$ exists such that the two variables are independent given $Z$ (Pearl, 2009). For example, if $X \perp\!\!\!\perp Y \mid Z_{\text{d-sep}}$, then $X \perp\!\!\!\perp Y \mid Z_{\text{in } p(V)}$. I will restrict my analysis to the set of faithful distributions where $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}} \iff (X \perp\!\!\!\perp Y \mid Z)_{\text{in } p(V)}$ to facilitate structural learning. When constructing a valid backdoor adjustment set, I find the optimal set $O$ by including the factual parents for each potential outcome that lies on a causal path between treatment and outcome. I then obtain the optimal minimal adjustment set by pruning $O$ , which then provides the highest precision among all possible minimal adjustment sets (Rotnitzky, 2020). As for the relationship between variables, I'm careful to not rely on a linearity assumption by using non-parametric methods to perform causal discovery and compute the average causal effect, but I include linear implementations for reference.

## 3. Methods

### 3.1 Data

The data comes from the 2014-15 round of the NFHS (International Institute for Population Sciences, 2015-16). The data collects information on a representative sample of households across all states in India. The survey collects information from a mother's height to what type of toilet facility is in the household. The initial question of interest is whether receiving supplemental nutrition from an ICDS center leads to an improvement in a child's nutritional status. The survey asked households whether or not a child attended an ICDS center and received supplemental nutrition benefits in the last 12 months. One might think to simply pass the entire NFHS dataset to a causal discovery algorithm to find the major determinants of child nutrition. But, a generalized algorithm that doesn't rely on a linear test would never finish as there are over $200,000$ observations and over 500 variables measured for each one. So, I restrict my analysis to only variables the literature has identified as determining either nutritional status or access to ICDS care. These include baseline characteristics, such as child age and gender, as well household demographics, such as mother's height, mother's years of education, overall wealth level, and whether or not the household is in a rural area (Jayachandran and Pande, 2017). Additionally, birth order and number of siblings have been shown to affect a child's access to nutrition (Spears and

Behrman, 2019), as well as exposure to open defecation, i.e. whether or not human waste is flushed away (Spears, 2020). Social status, as determined by caste, can also determine nutritional outcomes and access to ICDS care (Deshpande and Ramachandran, 2020).

I dropped any observation with that didn't contain a response to at least one variable of interest. I attempted to include state variables in my causal discovery phase by creating a binary indicator for each state. But, the problem is that it adds over 25 variables. Coupled with the fact that there are over 200,000 observations after cleaning, the causal discovery algorithm failed to finish even after 4 hours of running [1]. So, I will restrict the causal discovery phase to a single state—Uttar Pradesh—and randomly sample a set observations (10%) to construct a DAG. Then, I will use the full dataset when computing the average causal effect.

### 3.2 Causal Discovery

In order to justify the relations between variables beyond just background knowledge, I used Tetrad to execute a search algorithm over the data Glymour (1988). For background knowledge, I put household-specific factors, such as age, sex, and mother's height, in a lower tier relative to more demographic-related factors, such as rural and lowerCaste, so that the direction of causality is accurate. For instance, changing the mother's height won't cause her caste to change, but being a part of a given caste may have consequences on a mother's height. I then used the PC algorithm to search over the data and create causal model. To account for non-linear relationships between variables, I use the Conditional Correlation Independence (CCI) test which is "fairly general independence test" that models each variable as "a possibly nonlinear function of its parents, plus some additive noise, where the noise may be arbitrarily distributed" (Carnegie Mellon University, 2020). Using CCI within the PC algorithm helps elicit a causal model without making a strong assumption about linearity that an algorithm like FGES does when using a Gaussian BIC Score [2]. I also use the Fast Conditional Independence Test (FCIT)[3], a non-parametric method, to test the absence and presence of certain key edges in the initial graph produced by Tetrad. FCIT tests the null hypothesis $X \perp\!\!\!\perp Y \mid Z$ and a gives a p-value. If the p-value is below some specified significance level, we reject the null hypothesis, implying that an edge between $X$ and $Y$ should exist (Krzysztof et. al., 2018).

### 3.3 Identification

As for estimating the average causal effect of receiving nutrition benefits from an ICDS center, I look at both Uttar Pradesh and the country-wide level. For country-wide analysis, a few assumptions are required, but I don't believe they will significantly bias results. For both linear AIPW estimations, I used the formula below to calculate the average causal effect using my own implementation of AIPW[4]. A logistic model is fit for predicting treatment and a linear model for outcome, both using the same backdoor ad-

1. Using a more general, non-linear test
2. This, however, meant that the causal discovery algorithim did took over four hours to finish even after only running it on a random sample...
3. The average p-value of 10 FCIT tests was was used for testing edges
4. The random seed was set to 42

justment for $Z$. I also compute ACE using machine learning implementations of backdoor adjustment and augmented inverse probability weighting, each of which don't operate under a parametric assumption like linearity. Both rely on random forests from the scikit library to make predictions based on votes from each decision tree (Pedregosa et al., 2011).

$$\mathbb{E}\left[\frac{\mathbb{I}(A=a)}{p(A \mid Z)} \times (Y - \mathbb{E}[Y \mid A, Z]) + \mathbb{E}[Y \mid A=a, Z]\right] - \mathbb{E}\left[\frac{\mathbb{I}(A=a')}{p(A \mid Z)} \times (Y - \mathbb{E}[Y \mid A, Z]) + \mathbb{E}[Y \mid A=a', Z]\right]$$

## 4. Results

### 4.1 FCIT Sensitivity

Before I calculated the average causal effect, I tested the presence/absence of certain key edges using the FCIT, all of which can be seen in Table 1. The color scheme for the final DAG is as follows: orange edges were not produced by Tetrad but were suggested by FCIT, blue edges were edges produced by Tetrad but not confirmed by FCIT, purple nodes indicate the initial optimal minimal adjustment set according to the DAG first produced by Tetrad before any modifications, and the outcomes of interest are highlighted in red. To focus my sensitivity analysis, I first tested the absence of edges as the absence of an edge implies no direct effect and, if wrong, can lead to invalid backdoor adjustment sets. Tetrad produced a graph without any edges emanating from wealth, birth order, years of mother education, number of children under five, and mother height to either nutritional health outcome. This was concerning as these variables have extensive support from the literature as being prominent drivers of child health outcomes. FCIT test rejected the absence of edges between all of these variables and the two nutritional outcomes with p-values far below my significance level of $\alpha = 0.05$. This means that the initial optimal minimal backdoor adjustment set given by Tetrad would be susceptible to spurious correlation as multiple backdoor paths from "recievedBenefits" to both outcomes would have been open. So, I added directed edges from these variables to the each z-score variable.[5] It is interesting as to why Tetrad didn't output these edges, especially since the Conditional Correlation Independence and FCIT tests are both general tests without any strong linearity assumptions. Nonetheless, they are different tests and we should account for the FCIT results given the significantly low p-values. Edges that FCIT did confirm the absence of were those from "numberChildrenunderFive" and "motherHeight" to each of the nutritional z-scores. Both Tetrad and FCIT agreed on the absence of any direct causality. This is a bit surprising as mother's height may be an indicator of a genetic disposition and having more siblings may spread food supplies thin in a household. However, FCIT reveals that, conditioning on the appropriate set leads to no direct causal effect so I will not include the edges.

Interestingly, Tetrad put a directed edge from the treatment "recievedBenefits" to each of the nutritional outcomes "HeightforAgeZ" and "WeightforAgeZ." But, FCIT failed to reject the null hypothesis with p-values above the significance level, suggesting that neither edge should be present. I choose to keep both edges in place because although FCIT may have not concluded any direct dependency between nutritional status and receiving ICDS benefits *on average*, certainly, on an individual level, taking part in an ICDS center's

5. This did not create a cycle as both Height-for-AgeZ and Weight-for-AgeZ are sink vertices

supplemental nutrition program can affect a child's height and weight, and thus impact their height-for-age or weight-for-age z-score. Receiving benefits from an ICDS center is still a potential cause of nutritional outcomes so each edge is kept.

### 4.2 FCIT Results

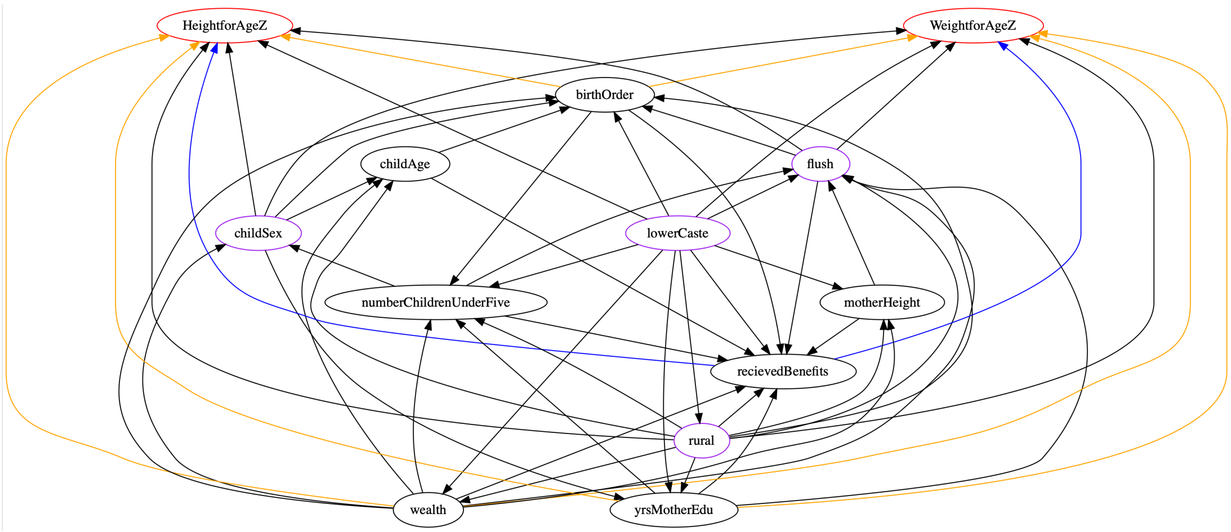
Testing for Presence

| EDGE | Z | p-value |
|---|---|---|
| wealth → WforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 0.000237 |
| wealth → HforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 3.97e-07 |
| birthOrder → WforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 1.45e-05 |
| birthOrder → HforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 7.87e-05 |
| yrsMotherEdu → WforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 4.72e-06 |
| yrsMotherEdu → HforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 1.57e-06 |
| numberChildunderFive → WforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 0.63 |
| numberChildunderFive → HforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 0.79 |
| motherHeight → WforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 0.63 |
| motherHeight → HforHZ | flush, lowerCaste, rural, childSex, recievedBenefits | 0.79 |
| Testing for Absence | | |
| receivedBenefits → WforHZ | flush, lowerCaste, rural, childSex | 0.37 |
| receivedBenefits → HforHZ | flush, lowerCaste, rural, childSex | 0.23 |

The final DAG after adding the edges in orange is below. The optimal adjustment set for computing the ACE of attending an ICDS center for both nutritional outcomes (height and weight) is the same: {wealth, yrsMotherEdu, rural, childSex, lowerCaste, birthOrder, flush}. This is expected as both outcomes are quite similar measures of nutritional status and thus will have similar determinants. The set is also minimal as removing any variable makes the set no longer a valid backdoor adjustment set. This is the adjustment set used for estimating the ACE of attending ICDS at the state level but will require a modification for looking at the country level. [6]

6. Note that full dataset (whether it be at the Uttar Pradesh level or country wide) is used to estimate ACE, rather than the random sample taken during the causal discovery phase.

### 4.3 Final DAG




### 4.4 Estimates

Below are the linear estimates of the average causal effect of having received benefits from an ICDS center in the last year in Uttar Pradesh with confidence intervals computed via bootstrapping a 100 times.[7] The estimates were calculated using the all observations from Uttar Pradesh.

#### 4.4.1 Uttar Pradesh: Linear Estimation

Linear ACE: Receiving ICDS benefits for children in Uttar Pradesh

| | (1) | (2) |
|---|---|---|
| Z-Scores (WHO) | Weight-for-Age | Height-for-Age |
| AIPW (no trim) | -0.0532 | -0.0109 |
| Confidence Intervals | (-0.0775, -0.0313) | (-0.0361, 0.0186) |
| AIPW (with trim) | -0.0531 | -0.0109 |
| Confidence Intervals | (-0.0766, -0.0323) | (-0.0425, 0.0279) |

95% confidence intervals produced via bootstrapping 100 times

#### 4.4.2 Country Level: Linear Estimation

Below are the linear estimates of the average causal effect of having received benefits from an ICDS center in the last year for all Indian children. Since the causal discovery phase could only be run on a random sample of observations from Uttar Pradesh, we will need to discuss how to construct a valid backdoor adjustment set without necessarily knowing what the DAG would have looked like had we had a supercomputer to search through the country-level dataset. First, are the causal relations encoded in the Uttar Pradesh DAG applicable to all of India? Uttar Pradesh is the most populated state in the country and many of the

7. A random seed of 42 was used throughout

variables in the DAG, i.e. wealth, flush, lowerCaste, will affect child nutritional outcomes in some fashion regardless of which state a child is located in. So, we can assume that the DAG is largely applicable to all Indian children. To be safe, we will forgo precision in favor of ensuring a valid adjustment set by simply conditioning on all the parents of the treatment "recievedBenefits." However, the country-wide dataset includes a state dummies for each state. Different states in India will implement their ICDS program differently, meaning that access to supplemental nutritional programs will vary according to state. Thus, we can assume state is a determinant of attending an ICDS center and is thus a parent included in our adjustment set. By conditioning on all the parents of "recievedBenefits," including state, we can be confident that we've constructed a valid backdoor adjustment set for computing the average causal effect of receiving ICDS care for all Indian children[8].

Linear ACE: Receiving ICDS benefits for all children in India

| | (1) | (2) |
|---|---|---|
| Z-Scores (WHO) | Weight-for-Age | Height-for-Age |
| AIPW (no trim) | -0.133 | -0.128 |
| Confidence Intervals | (-0.141, -0.126) | (-0.162, -0.087) |
| AIPW (with trim) | -0.133 | -0.129 |
| Confidence Intervals | (-0.143, -0.117) | (-0.161, -0.87) |

95% confidence intervals produced via bootstrapping 100 times

### 4.4.3 Non-linearity Sensitivity using Machine Learning

In order to fully estimate the average causal effect of attending an ICDS center without linearity assumptions, I used random forests to implement backdoor adjustment and AIPW. I used the same adjustment sets as I did in the linear estimations.[9]. Below are the results from computing ACE from non-parametric machine learning methods.

Uttar Pradesh

| | (1) | (2) |
|---|---|---|
| Z-Scores (WHO) | Weight-for-Age | Height-for-Age |
| Backdoor_ML | -0.0955 | -0.00191 |
| Confidence Intervals | (-0.121, 0.00653) | (-0.0884, 0.0903) |
| AIPW_ML | 0.00904 | 0.0509 |
| Confidence Intervals | (-0.142, 0.184) | (-0.246, 0.120) |

Country-Wide

| | (1) | (2) |
|---|---|---|
| Z-Scores (WHO) | Weight-for-Age | Height-for-Age |
| Backdoor_ML | -0.0881 | -0.0674 |
| Confidence Intervals | (-0.125, -0.0646) | (-0.107, -0.0206) |
| AIPW_ML | -0.150 | 0.142 |
| Confidence Intervals | (-0.231, 0.0126) | (-0.0451, 0.272) |

95% confidence intervals produced via bootstrapping 25 times

8. Country-wide adjustment set: { flush, lowerCaste, rural, child sex, birth order, wealth, yrs mother edu, child age, mother's height, state}
9. Uttar Pradesh Optimal Backdoor Adjustment Set: {wealth, yrsMotherEdu, rural, childSex, lowerCaste, birthOrder, flush}

## 5. Discussion & Conclusion

On the whole, both linear and machine learning estimates of the average causal effect of attending an ICDS center in the last year show no evidence that the supplemental nutrition has resulted in significant, positive improvements in child health outcomes, as indicated by height-for-age and weight-for-age z-score. The confidence intervals for ACE using linear AIPW (both with and without trimming) for height-for-age z-score in Uttar Pradesh contain zero, suggesting that there is no causal effect of receiving ICDS supplemental nutrition benefits. The confidence intervals for weight-for-age z-score don't contain zero but the point estimates are negative and small with both roughly at $-0.05$. For the linear AIPW estimations at the country level, the condfience intervals don't contain zero but the point estimates are more negative at roughly $-0.133$ for weight-for-age and $-0.128$ for height-for-age. Testing the sensitivity of our results to the linearity assumption, I observe that even using non-parametric machine learning methods to compute ACE, there still is no evidence of a positive impact on child nutritional status from receiving benefits from an ICDS center. Each of the confidence intervals for Uttar Pradesh contain zero. On the country level, we see much of the same: Backdoor_ML suggests a negative but minuscule effect of receiving ICDS treatment and AIPW_ML suggests no impact with its confidence intervals including zero. Perhaps the most convincing evidence of ICDS having no positive impact on Indian children is FCIT's failure to reject the null hypothesis that "receivedBenefits" and either of the nutritional z-scores are independent when conditioning on the appropriate adjustment set, implying that there is no direct effect between ICDS benefits and nutritional outcomes. Along with the repeated instances of confidence intervals containing zero and the small negative point estimates above, there is no evidence of a positive average causal effect.

One assumption that I did not analyze the sensitivity of my results to, and that future work should explore, is the PC algorithm's assumption that "no two variables are caused by... an unmeasured variable" (Carnegie Mellon University, 2020). I did run an FCI algorithm over the dataset and unsurprisingly, there were many bidrected edges, including between the treatment and each outcome. Indeed, the real difficulty in computing an unbiased estimate of the ACE for ICDS centers would be to allow for there to be unmeasured confounding between treatment and outcome. In this case, we'd likely need an appropriate mediator between treatment and nutritional z-scores or a valid instrument. But, despite the size of the NFHS dataset, there are no self-evident variables to use for either front-door adjustment or the instrumental variable method. This should not keep us from attempting to access the effectiveness of ICDS, however. The best we can do in a policy landscape is block spurious correlation as best we can, compute estimates, and make policy from there. We should not let the difficulty of identification let the status quo persist, especially when it is evident that the current iteration of ICDS is not producing the expected impact of a billion dollar investment. In my analysis, I block all paths of spurious correlation according my DAG which was built on data containing information on the prominent determinants of child health and ICDS access documented in the literature. I also worked to ensure that my DAG was not susceptible to strong assumptions about linearity. My analysis shows that there is little evidence that ICDS centers have a positive ACE on Indian children, a result that is not so surprising when you account for the program's implementation issues (Timsit, 2019).

## References


Carnegie Mellon University. Tetrad manual. tetrad single html manual. Available at https://cmu-phil.github.io/tetrad/manual/ (2022/05/12), 2020.

A. Deshpande and R Ramachandran. Which indian children are short and why? social identity, childhood malnutrition, and cognitive outcomes. 2020.

Scheines R. & Spirtes P. Glymour, C. Exploring causal structure with the tetrad program. 1988.

Quinn Kevin. Glynn, Adam. An introduction to the augmented inverse propensity weighted estimator. 2009.

G. Donovan J Horvitz, Daniel. A generalization of sampling without replacement from a finite universe. 1952.

International Institute for Population Sciences. India national family health survey nfhs-4 2015-16. 2015-16.

S. Jayachandran and R Pande. "why are indians so short? the role of birth order and son preference",. 2017.

Krzysztof et. al. Fast conditional independence test for vector variables with large sample sizes. Available at https://doi.org/10.48550/arXiv.1804.02747, 2018.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Smucler Ez Rotnitzky, Andrea. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. 2020.

& Kumar A Sood, M. Research on icds an overview. *Journal of Sketchy Physics*, 3, 2008.

Coffey D. Spears, D. and Behrman. Birth order, fertility, and child height in india and africa. 2019.

D Spears. Exposure to open defecation can account for the enigma of indian child height. 2020.

A. Timsit. Inside india's ambitious effort to provide early care and education to 400 million kids. Available at https://qz.com/india/1584703/indias-icds-anganwadi-system-is-a-challenged-but-impressive-effort/ (2021/05/10), 2019.

World Bank. India gdp growth. Available at https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2017&locations=IN&start=2000 (2021/05/13).