Gavi Vaccination Impacts: A Propensity Score Study

Jessica Spencer

Applied Statistics for Social Science Research : New York University

APSTA-GE 2018: Advanced Causal Inference

Joseph Cimpian

May 12, 2020

Abstract

In 2000, a pooling of public and private donors created a partnership called the Global Alliance for Vaccines and Immunisation (aka the GaviAlliance), whose mission was to increase immunization rates in poor countries. Gavi began to support low-income countries by providing them with bulk purchases of vaccines for diphtheria, pertussis and tetanus (DPT), hepatitis B, pneumococcal disease, rotavirus, and Haemophilus influenza type B. The alliance used a strict cut-off to determine eligibility of aid, which researchers from the Center for Global Development exploited to create a causal estimate using a regression discontinuity design. They found Gavi's provision of vaccines to have little or no impact on vaccination rates for hepatitis B and DPT or mortality rates. There is some evidence of positive impact for the other, newer and more expensive, vaccines, but they are small and statistically insignificant. However, the regression discontinuity design only produces a local average treatment effect, which cannot be generalized to countries far from the $1000 GNI cutoff. The following paper attempts to produce more relevant estimates to the population of interest – poor countries, which on average have low vaccination rates. Due to reproducibility issues, only the effect of Gavi on mortality rates were examined. A doubly-robust propensity score design is used. The sample size was very small, and balance was not achieved in the matching step – although this was compensated for with the doubly robust design. Still, consistent near-zero and very uncertain results were found, with little power.

Literature Review

Research into Gavi was motivated in part by a gap in literature on aid effectiveness. In the early 2000s, researchers were beginning to question where the financing from aid organizations was being put to (Burnside, 2000). Many were worried that aid packages were fungible-- were aid packages meant for vaccination being used for something else? In 2006, Lu et al. found evidence of positive and significant impact due to Gavi funding on DPT vaccination rates where only 65% of children under five were originally receiving the vaccine. Not much later, in 2010, Hulls et al. extended this analysis to include more years of data, and found no such effect on those countries for DPT coverage. They did, however, find a significant impact on countries that had between 65% and 85% coverage of vaccines for children under five. Both these studies used a General Method of Moments design (GMM) from software packages, which raised concerns about causal inference. Dykstra, Glassman, Kenny and Sandefur created an independent measure of the causal impact of Gavi support in a 2015 working paper published by the Center for Global Development, that also investigated fungibility. This paper, and its reproducibility materials, form the basis for this study.

The working paper leveraged the strict cut-off for Gavi eligibility to create a fuzzy regression discontinuity design. This cutoff was decided at Gavi's first board meeting:

> Countries with small resources and a lack of purchasing power have been considered to be in greatest need of financial support for the new vaccines. It is proposed that this be initially interpreted as those with a GNP/ capita equal or less that 1,000 USD

Dykstra et al. measured the effect of Gavi in participating countries for the five phases of program rollout. They looked at outcomes of child mortality and aid received per infant, both on the country level.

The authors discuss the limits of their data, which I agree could have weakened their analysis. In the first phase of Gavi support, the $1000 GNI acts appropriately as an instrument. However, in subsequent phases, changes in GNI do not correspond with a loss in eligibility. Over the five phases of Gavi, using GNI data from 1998, 2003, 2009, 2010 and 2011, twenty-one countries experienced a change in eligibility. This was either because the country's GNI changed, or the eligibility criteria from Gavi changed around them. Out of these countries, seventeen originally qualified for aid, and then became ineligible. However, in fourteen of these countries, Gavi continued to disburse aid even after the country technically became ineligible. Gavi has a graduation policy that lets countries continue to receive aid from their program for an agreed upon period – usually about 5 years. This explains many of the mentioned cases. Therefore, countries that were eligible in phase 1 were persistently eligible, despite any gains in GNI. This shows that the true instrument for eligibility is primarily GNI per capita during phase 1. Nevertheless, the authors decided to use GNI from the above-mentioned time points, adjusting for any changes in Gavi policy, as their instrument for eligibility for all the phases of rollout they examined. Furthermore, the authors discuss a pattern that they spotted, where larger countries receive less Gavi aid per capita. An interaction of the country's infant population and the eligibility indicator were used as an additional instrument to account for this effect. This interaction would take cases into consideration where smaller countries experience a bigger jump at the threshold due to larger aid.

In addition to a weak instrument, the working paper also shared concerns of their variable estimates. Estimates of vaccine coverage, used as an outcome variable in many of the working paper's analyses, could be subject to under reporting, and errors in construction. Gavi doesn't report on the number of doses purchased per country in the phases examined. Instead, the authors used data from UNICEF, which procures vaccines on behalf of Gavi. UNICEF reports on the total number of doses purchased worldwide, and the amount spent per vaccine. This is used to calculate the average cost per dose of vaccine. Average cost per vaccine is then used to estimate how many doses of each vaccine was purchased by Gavi per country per year. Some of these vaccines were very likely used for 'catch-up' vaccination, which is vaccination given to children above the age of one, who are not accounted for in the WHO vaccination rates. The authors identified that this could potentially lead to an underestimation of Gavi's impact, if Gavi aid was often given to children older than one. They do not observe catch-up vaccination in their data, although it is well documented in the literature, and thus still a concern (Clark, 2009). Progress before Gavi's launch also suggests a convergence of coverage between low-income and higher-income well-vaccinated countries, which may explain some of the gains in immunization coverage seen in their analysis. (Gelman & Imbens, 2019)

The working paper is robust in its estimates and methods, even in light of new publications. The working paper modeled Gavi's impact with a fuzzy regression discontinuity design. In keeping with Gelman and Imbens (2014)[i], they use first and second order polynomials to estimate effects on either side of the cut-off. In 2019, Gelman and Imbens published another paper on the same topic, (high-order polynomials for causal inference), where they reaffirm this method, showing that any more than linear and quadratic terms could lead to "noisy estimates,

sensitivity to the degree of the polynomial, and poor coverage of confidence intervals" (Gelman & Imbens, 2019).

The functional form chosen then is still recommended practice, but there are some other standards of practice that have change. To check for robustness, the bandwidth around the threshold was varied. First the optimal bandwidth was chosen to minimize mean-square error (Imbens, 2012), and then half and twice this bandwidth was also shown. The results found remain persistent across these bandwidths. At the time, it was thought that this bandwidth selection method allows for a trade-off between bias and variance that would come from choosing a bandwidth that is very close to the threshold, or too far away (Cattaneo, Titiunik, & Vazquez-Bare, 2019). But we now know that this bandwidth selection method is problematic. In a 2019 paper, Cattaneo et al. recommend bandwidth selection that minimizes on the coverage error probability in cases of causal inference. This bandwidth selection algorithm chooses smaller bandwidths than mean-squared error bandwidth selection, at least in large samples. In fact, the MSE-optimal bandwidth was shown to be too large for conducting inference in OLS approximations. Perhaps the author's choice of bandwidth selection algorithm, which was standard and recommended practice at the time, was large enough that it diluted their results.

To improve upon their study, which they acknowledge has a weak instrument, the authors could consider reevaluating their analysis considering the results by Feir, Lemieux and Marmer from 2016. "Weak Identification in Fuzzy Regression Discontinuity" was published in 2016, after the publication of the working paper, but going forward it's results could strengthen the analysis of Gavi treatment effects. Feir et al. found that weak identification in fuzzy regression discontinuity designs could lead to size distortions. Therefore, confidence intervals constructions

(+-constant x standard error) are invalid. They discuss a new t-statistic that accounts for these special circumstances.

Falsification tests were discussed by the authors, but rarely shown. The Gavi study's regression discontinuity design assumes that the potential outcomes are continuous to the forcing variable. To check this assumption, the authors discussed that there was no discontinuity in pre-treatment trends in vaccination across the threshold. The weighting method of the design was not discussed, and it was never shown that the distribution of the t-statistic was nearly standard normal, as discussed in the Cattaneo et al. regression discontinuity introduction. Showing falsification test results would have only strengthened the study's design, and made it easier for readers to understand and critique their analysis. Covariate balance mentioned, and the authors do mention that there was no discontinuous jump at the threshold for covariates other than the outcome-- indicating that there is no evidence of confounders. However, this was never shown. The density of the running variable was said to be tested for continuity, but again, the exact results were not part of the published paper. One other puzzling element of their design was the decision to include the log GNI as a control in both the first and second stage of the regression discontinuity design. Additionally, it would have been interesting to see alternative cutoffs explored on this dataset.

Research Design

The estimates in the Center and Global Development working paper were interesting, but incredibly narrow, due to the Regression Discontinuity design. As the authors themselves point out:

> The local average treatment effect we estimate is specific to the immediate vicinity of the Gavi eligibility cut-off, i.e., for the wealthiest Gavi recipients. There is *a priori* justification to believe the impact of subsidized vaccines will be smallest for these recipients, both because baseline immunization rates are higher, and because these middle-income countries have the greatest capacity to purchase new vaccines as they emerge

The original regression discontinuity design only includes countries that were close to the $1000 GNI per capita threshold. Therefore, the population of the study does not align with the countries they are most concerned about. The true population of interest are lower income countries, that have more need and less resources for vaccination, that have also received Gavi aid. The purpose of the following propensity score study design is to create a causal estimate of the impact of the Gavi program on mortality rates on these participating countries, which will include those further from this $1000 GNI per capita cut-off. Like the authors of the above study, I think that it's worth investigating whether the impact of the Gavi program will be pronounced further from the threshold. Unfortunately, the dose-response model from the original study could not be replicated with the materials that the authors provided. Only outcomes of mortality remained useable, and only the first phase of Gavi rollout is examined, in comparison to 5 years of data before the implementation of the program. The research question is as follows: does participation in the Gavi Vaccination program have any effect on mortality rates?

Methods: Propensity Score Design Overview

Propensity score matching is a method that can be used to obtain estimates in non-experimental settings, where there is non-random assignment of the treatment. This fits well with the Gavi vaccination program, which is only available to countries that have a per capita GNI less than $1000 USD. Propensity score matching will be used to choose the sample for the analysis, and then estimates of the impact of Gavi will be taken from this sample. The "control" observations are from countries that did not receive any Gavi support. The "treatment" observations are from countries that did. The control and treatment observations must have "sufficiently" similar measurements on a group of pretreatment variables, X, in this subset, for the effect estimate to be believable. This, along with other assumptions, will be detailed in a later section.

Two methods of propensity score matching are used to ensure robust results, and assumptions are checked on both. The first method used to estimate propensity scores is a logistic model that then uses nearest-neighbor matching. The second is a Generalized Boosted Model (GBM) estimate. Both methods first create a 'propensity score', or a single number summary of an observation's covariates (for the covariates given to it, listed below). Then, these methods match similar control and treatment estimates using that score to create a new sample.

The Logistic regression approach generates propensity scores using this equation:

$$Treatment = \alpha + \beta_0 * aveIMR * INC + \beta_1 * avePop1 + \beta_{2*} aveU5MR + \beta_3 * aveMalnourished$$

Where *aveIMR* is average infant mortality rate, *INC* is income, *avePop1* is population under the age of 1, *aveU5MR* is mortality rate for children under the age of 5, and *aveMalnourished* is the percent of children who are malnourished in each country included in the analysis. More specific information on these covariates can be found in the next section.

The Generalized Boosted Matching (GBM) used this equation:

$$Treatment = \alpha + \beta_0 * aveMalnourished + \beta_1 * aveU5mr + \beta_2 * aveIMR + \beta_3 * avePop1 + \beta_4 * avePop5 + \beta_5 * INC$$

Where all the covariates retain the same meaning as above, and *avePop5* is population of children under the age of 5.

These procedures generated propensity scores for each observation, which are then used to create a sample where the treatment and control observations are 'sufficiently' similar. Similar observations should have similar propensity scores, and the new samples for treatment and control observations should have small differences in means, and ratios of standard deviation that are close to 1 (which would indicate that they are identical). Once this similarity is assessed, by looking for overlap in propensity scores, differences in means, and standard deviation ratios, a treatment effect can be estimated like so:

$$Y_{IMR} = \propto + \beta * treatment$$

Here, *Y_IMR* is the change in infant mortality rate from 1995-1999 to 2000-2005. The value $\propto$ is the average change in Y_IMR for all included countries, and the value β is the estimated change in infant mortality rate for those who received Gavi support. β is also known as the treatment effect on the treated, or TOT.

The above estimation would be enough, if it could be shown that the new propensity score samples were sufficiently similar. However, in both cases of propensity score estimation, "balance" was not reached. This means that the observations in the sample were not found to be sufficiently similar in the matching set. A 'doubly robust' propensity score matching design was then used to 'soak up' the variance left by poor matching. The equation for this effect estimation was then two-fold:

For the logistic estimate:

$$Y_{imr} = \alpha + \beta_0 * Treatment + \beta_1 * avePop1 + \beta_2 * aveMalnourished + \beta_3 * aveWasted + \beta_4 * aveU5mr + \beta_5 * aveLGNIpc + \beta_6 * aveIMR * INC$$

For the GBM estimate:

$$Y_{imr} = \alpha + \beta_0 * Treatment + \beta_1 * Inc + \beta_2 * avePop1 + \beta_3 * avePop5 + \beta_4 * aveMalnourished + \beta_5 * aveWasted + \beta_6 * aveU5mr + \beta_7 * aveLGNIpc + \beta_8 * aveIMR$$

The coefficient "β_0" retains the same meaning in this equation. The other covariates are taken from the equation used to estimate the propensity scores. None of the other coefficients are interpretable – they are only there to soak up left over variance that could not be accounted for in the matching. The population of causal inference is very small for propensity score studies. This treatment effect can by generalized only to countries that were included in the sample for that estimation.

Propensity Estimation Detail

There are a few technical assumptions that must be satisfied if a researcher is to use propensity scores to generate a causal estimate:

1) Unconfoundedness – potential outcomes are independent of the treatment assignment given a set of variables
2) Overlap- any observations with the same *P* values have a positive probability of being in both the treatment and control group

Both these assumptions together are referred to as 'strong ignorability' (Caliendo & Kopeinig, 2008). The first assumption is dependent on the quality of the data, and is assessed at the discretion of the researcher. The second can be satisfied by looking at the overlap of treatment and control observations by propensity score, and by examining the balance of pretreatment covariates. Looking at overlap is as simple as creating a histogram of the frequency of observations per propensity score for both the treated and control groups, and overlaying these

histograms. These histograms shouldn't have much separation. In fact, they should be overlapping enough that a reasonable subset can be made from overlapping treatment and control observations. Balance is examined by comparing moments in this matched sample. The matched sample's difference in means should be assessed, and should ideally be less than 0.10, says Linden, while describing standard practice (Linden & Yarnold, 2016). This design also looks at the ratio of standard deviations, which should be between 0.9-1.1 for good balance to be reached, as suggested by Jennifer Hill in her Introduction to Causal Inference course at NYU in Fall 2019. If the propensity score model meets its assumptions, then outcomes from these well selected cohorts can be attributed to the effects from the treatment for the countries in that matched subset (Caliendo & Kopeinig, 2008).

Logistic regression was used to create propensity scores, a standard choice in propensity score designs when there is a binary treatment indicator. The exact functional form of this matching step is not very important, and can be adjusted to try and achieve better overlap and balance. Nearest-neighbors (NN) matching without replacement was then used to match treated and control observations using the propensity scores. NN matching choses control observations by a treatment observation that has the closest propensity score. Because replacement was not used when matching, the bias is lower and the variance is higher than what might have been found with replacement. Balance is examined by looking at the difference in means, and the ratio of standard deviations.

As detailed by McCaffrey in a 2011 paper on the technique, Generalized Boosted Models (GBMs) are a multivariate, nonparametric regression technique that estimate propensity scores in an automated, data-adaptive fashion, which allows for flexible and non-linear relationships between the treatment and covariates (McCaffrey, Ridgeway, & Morral, 2004). Other estimation

techniques are less flexible and can sometimes require variable selection, which risks biasing those estimates. As with the logit model, the exact functional form can be adjusted to try and achieve better overlap and balance. The Twang package from CRAN is used for my GBM, and the balance is assessed by looking at the standardized difference in means, also known as the effect size. Balance is assessed by looking at the box-and-whisker plot produced by different stopping rules. If balance is reached, then the rule with the most overlap is the rule used to create matches.

Analysis

The Dataset, Variables & Data Transformation

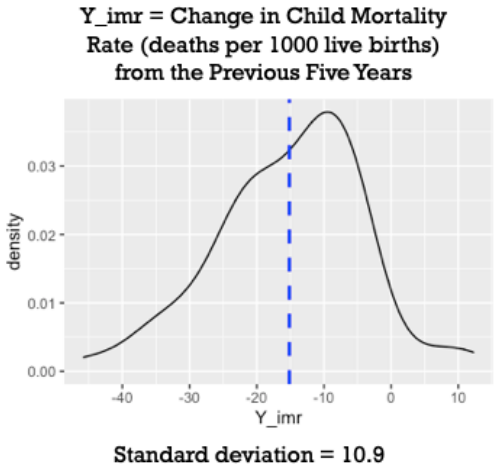
The dataset to be analyzed is derived from the originally published study on the Gavi vaccination program. It contains observations on 127 countries, all of whom are 2 log points from the threshold of $1000 GNI per capita, from five years prior to the study to five years after the implementation of its first phase (1995-2005). The data from the five years prior to the study were summarized, and these data will be referred to as 'phase 0'. The data from 2000-2005 will be referred to as 'phase 1', or the phase when some countries began to receive Gavi aid. The following covariates are used:

Variable Name	Variable Description	Data Source + Measurement Info
Phase	Levels = 1,0 : where 0 is the years 1995-1999, or the four years prior to the start of Gavi, and when 1 is 2000-2005, which was the first phase of the Gavi vaccination roll-out	Calculated from Year

treatment	Levels = 1,0: whether the country in question received Gavi support during phase 1	Gavi
Inc	Low, low medium, or upper medium. Countries with High income levels were deleted from the dataset by the authors of the original paper	Historical World Bank income group classification
Ave_pop1	population of the country under the age of one, average from phase 0 and phase 1	World Bank
Ave_pop5	Population of the country under the age of five, average from phase 0 and phase 1	World Bank
Ave_lpop1	population of the country under the age of one, log of the average from phase 0 and phase 1	World Bank
Ave_lpop5	Population of the country under the age of five, log average from phase 0 and phase 1	World Bank
Ave_malnourished	Percent of children under 5 who were 2 standard deviations under the weight-for-age median, WHO estimate, average from phase 0 and phase 1	WHO
Ave_wasted	Percent of children under 5 who were 2 standard deviations under the weight-for-height median, WHO estimate, average from phase 0 and phase 1	WHO
Ave_u5mr	Mortality rate for children under age of five for that year, average from phase 0 and phase 1	WHO
Ave_lgni_pc	The log of the Gross National Income per capita, average from phase 0 and phase 1	World Bank Development Income Indicator. Imputed by the original authors in low income countries lacking this data

Ave_imr	Infant mortality rate, average from phase 0 and phase 1	WHO – data available at 5 year increments

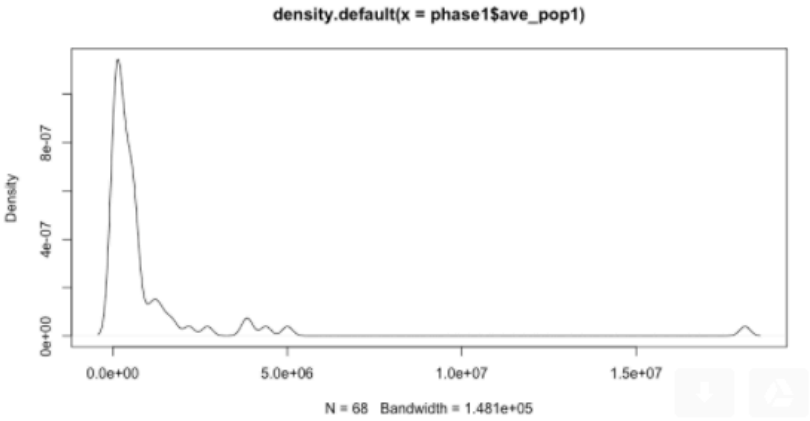
The outcome variable of interest is Y_imr, or the change in child mortality rate from the previous phase. These estimates were made by the WHO, and refer to the number of deaths per 1,000 live births. Looking at the distribution of this outcome in figure 1, Y_imr, one can see that it's mean is -15. This can be interpreted as "on average, the countries included in this study saw a decrease of 15 deaths per 1000 live births from the years 1995-1999 to 200-2005". The infant mortality rate is estimated every 5 years by the World Health Organization.



There were a few other variables that needed to be transformed or summarized other than by taking the average. The variable Inc, a factor variable with the levels Low, Low Medium, or Upper Medium, used the designation from 1999 for phase 0, and 2005 for phase 1. The designation is given anew every year, but the Gavi eligibility would have been assessed in 1999,

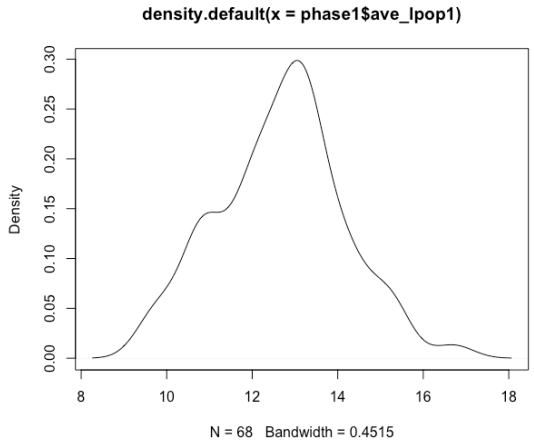
and then reassessed in 2005 at the end of the first phase, which is why those years were chosen. Additionally, the average population variables were transformed using a logarithm, because they were both very skewed, as shown below:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
14949	135811	356922	963002	648426	18116457



Now with the transformation, it is more approximately a normal curve:

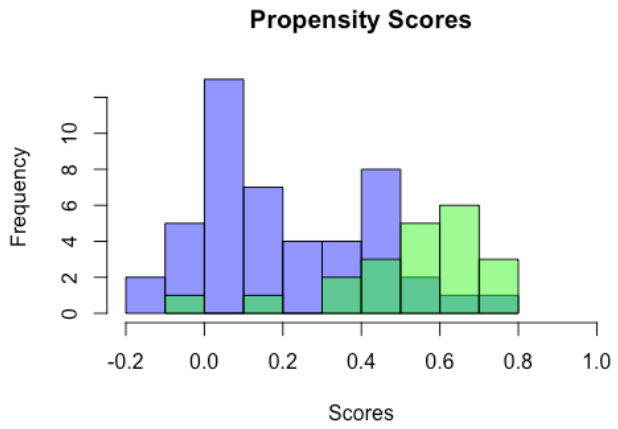
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
9.612	11.819	12.782	12.664	13.382	16.712



Results

Logistic Approach

First, propensity scores were generated using logistic regression, and after some testing, the equation above was settled on. The overlap generated by this matching design was ok, but still quite separated. The blue indicates control observations, and the green indicates treatment observations. It's clear that treatment observations tend to have higher propensity scores, between 0.3 and 0.8. Most the control observations fall below 0.2. However, there is workable overlap between 0.3 and 0.8.



Unfortunately, when balance is checked, there are some issues. Some of the differences in means between the control and treatment group have gotten worse after matching. The ideal threshold is less than a 0.1 difference in mean. The table can be read as follows:

Mn1 = mean of the unmatched treatment group

Mn0 = mean of the unmatched control group

Mn1.m = mean of the *matched* treatment group

Mn0.m = mean of the *matched* control group

Diff = difference in means before matching

Diff.m = difference in means after matching

Ratio = ratio of standard deviations before matching

Ratio.m = ratio of standard deviations after matching

	mn1	mn0	mn1.m	mn0.m	diff	diff.m	ratio	ratio.m
ave_lpop5	14.517	14.085	14.517	14.072	0.314	0.324	1.075	1.145
ave_lpop1	12.939	12.498	12.939	12.501	0.314	0.312	1.046	1.112
ave_lgni_pc	5.876	6.922	5.876	6.225	-2.489	-0.830	2.317	1.907
ave_u5mr	145.950	84.109	145.950	122.317	0.947	0.362	0.962	1.080
ave_imr	90.652	56.769	90.652	76.712	1.059	0.436	1.015	1.035
ave_malnourished	23.734	16.820	23.734	23.933	0.577	-0.017	1.029	1.076

For *ave_lpop5,* and *ave_lpop1,* the means are worse, or not much different from their unmatched means. They are not sufficiently close to 0.1. The rest of the covariates are improved, although many are still not close to 0.1. The ratios of standard deviation for the population variables have also gotten further apart after matching, and although the rest have improved, the *ave_lgni_pc* variable is still a bit far. There is very small sample size as well. Only 42 countries – 21 in the treatment and 21 in the control – were selected by the nearest-neighbors algorithm to be included in the matched sample. Without adding in covariates for robustness, the treatment effect was estimated as follows:

```
Call:
lm(formula = Y_imr ~ treatment, data = phase1_w)

Residuals:
     Min       1Q   Median       3Q      Max
-27.3310  -6.0060   0.6655   8.4798  29.3119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.062      2.532  -6.738 4.35e-08 ***
treatment     -1.357      3.581  -0.379    0.707
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.6 on 40 degrees of freedom
Multiple R-squared:  0.003578,  Adjusted R-squared:  -0.02133
F-statistic: 0.1436 on 1 and 40 DF,  p-value: 0.7067
```

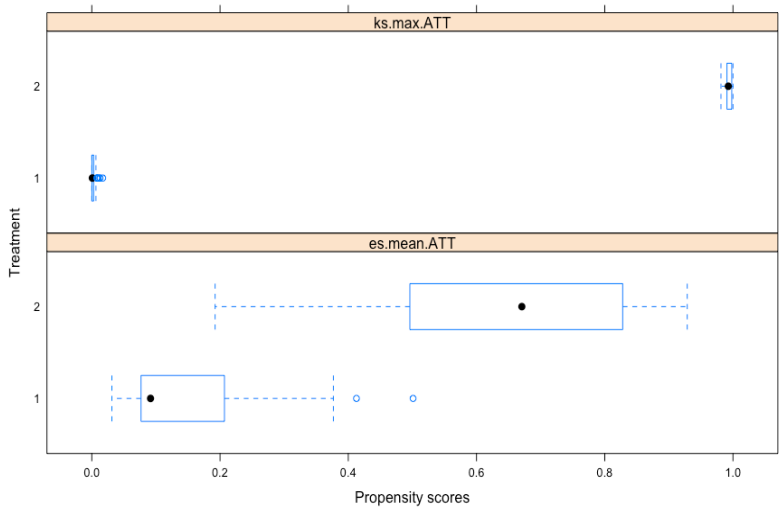
To compensate for this lack of balance in the propensity score matching step, a doubly-robust design was used. Adding in the covariates, the following estimation was produced:

```
Call:
lm(formula = Y_imr ~ treatment + ave_malnourished + ave_imr *
    inc + ave_lpop1 + ave_u5mr, data = phase1_w)

Residuals:
     Min       1Q   Median       3Q      Max
-27.9702  -6.6201   0.4664   6.1080  22.0904

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.0577    16.2491   0.496   0.6233
treatment        -1.8951     3.8660  -0.490   0.6272
ave_malnourished -0.4186     0.2094  -1.999   0.0539 .
ave_imr           0.1693     0.2627   0.644   0.5238
incLM            -6.1648    26.5630  -0.232   0.8179
incUM             5.0732    12.2005   0.416   0.6802
ave_lpop1        -1.6410     1.3375  -1.227   0.2286
ave_u5mr         -0.0610     0.1367  -0.446   0.6585
ave_imr:incLM     0.1880     0.7320   0.257   0.7989
ave_imr:incUM         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.96 on 33 degrees of freedom
Multiple R-squared:  0.2662,    Adjusted R-squared:  0.08835
F-statistic: 1.497 on 8 and 33 DF,  p-value: 0.1962
```

As you can see, the estimate is not much different. They are still close to zero, with a large standard deviation.

GGM Estimates

Next, the GBM estimates. The GBM generated propensity scores with the equation from above. It produced better overlap than the logistic method, when using the es.mean.ATT method, as seen here:



However, the balance experienced the same issues. The standardized effect size (std.eff.sz) is larger than 0.1 for the population variables, and for income levels Low and Low Medium.

```
$es.mean.ATT
                   tx.mn  tx.sd   ct.mn  ct.sd std.eff.sz  stat     p    ks ks.pval
ave_malnourished  23.734 11.985  23.522 13.567      0.018 0.053 0.958 0.230   0.497
ave_u5mr         145.950 65.282 129.539 75.848      0.251 0.763 0.448 0.232   0.497
ave_imr           90.652 31.982  80.396 35.667      0.321 1.028 0.308 0.261   0.358
ave_lpop1         12.939  1.404  12.365  1.381      0.408 1.329 0.189 0.345   0.120
ave_lpop5         14.517  1.375  13.930  1.367      0.427 1.390 0.169 0.335   0.147
inc:L              0.905  0.294   0.736  0.441      0.574 3.054 0.068 0.168   0.068
inc:LM             0.095  0.294   0.196  0.397     -0.344    NA    NA 0.101   0.068
inc:UM             0.000  0.000   0.067  0.251         NA    NA    NA 0.067   0.068
```

The sample size is slightly larger for this estimate, although not by much. There are 47 control observations and 21 treatment observations. Without correcting for balance, the first estimation of the treatment effect produced the following estimates, which were very similar to the previous logistic estimates:

```
Call:
svyglm(formula = Y_imr ~ treatment, design = design_gbm)

Survey design:
svydesign(ids = ~1, weights = ~weights, data = gbm_df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.2692     2.9689  -6.490 1.29e-08 ***
treatment     0.8502     3.7342   0.228    0.821
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 130.9376)

Number of Fisher Scoring iterations: 2
```

The estimates remained fairly neutral, with large standard deviations even when the covariates were included:

```
Call:
svyglm(formula = Y_imr ~ treatment + ave_malnourished + ave_imr *
    inc + ave_lpop1 + ave_u5mr, design = design_gbm)

Survey design:
svydesign(ids = ~1, weights = ~weights, data = gbm_df)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.26047    9.84848   0.433  0.66691
treatment         2.33897    3.00309   0.779  0.43923
ave_malnourished -0.32934    0.16000  -2.058  0.04406 *
ave_imr           0.11721    0.22104   0.530  0.59795
incLM            13.53782    4.15015   3.262  0.00186 **
incUM             3.29660    4.89689   0.673  0.50349
ave_lpop1        -1.71622    0.96346  -1.781  0.08010 .
ave_u5mr         -0.04077    0.11570  -0.352  0.72583
ave_imr:incLM    -0.24957    0.08831  -2.826  0.00646 **
ave_imr:incUM     0.38130    0.08743   4.361 5.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 92.32778)

Number of Fisher Scoring iterations: 2
```

Discussion

Comparison to the original paper is difficult. Due to reproducibility challenges, coverage estimates in the dose-response model could not be reconstructed. Therefore, only mortality rates were examined. The vaccinations offered by Gavi only account for about 15% of child deaths world-wide. Therefore, the effect of Gavi on this outcome would be modest, even if it were to reduce even a fraction of these deaths. Additionally, this propensity score design ignores the difference between those countries who were eligible and accepted Gavi aid (compliers), and those who were eligible who did not accept aid (non compliers). It is possible that there is some systematic difference between these groups of countries that this causal design does not investigate.

The analysis itself is also fraught from a weak sample size, in both cases. There are very few control and treatment observations, which means that slight differences could be exaggerated, and the estimates could be incredibly biased towards certain observations. The large amount of uncertainty could be due to this small sample size. Additionally, since balance could not be achieved, it is difficult to have complete faith in the estimates generated. However, the estimates were persistently neutral, across all estimation methods, which implies that participation in the Gavi Vaccination program and no effect on child mortality rates in those countries. This is consistent with what the authors of the working paper found using a regression discontinuity design.

References

Burnside, C. a. (2000). Aid, policies and growth. 847-969.

Caliendo, M., & Kopeinig, S. (2008). Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys, 22*(1), 31-72.

Cattaneo, M., Titiunik, R., & Vazquez-Bare, G. (2019). *The Regression Discontinuity Design.*

Clark, A. a. (2009). Timing of children's vaccinations in 45 low-income and middle-income countries: An analysis of survey data. *The Lancet 373 (9674)*, 1543-1549.

Dykstra, S., Glassman, A., Kenny, C., & Sandefur, J. 2. (2015, 2). *Replication data for: The Impact of Gavi on Vaccination Rates: Regression Discontinuity Evidence*. Retrieved from Harvard Dataverse, V1: https://doi.org/10.7910/DVN/27921

Feir, D., Lemieux, T., & Marmer, V. (2016). Weak Identification in Fuzzy Regression Discontinuity Designs. *Journal of Business & Economic Statistics, 34*(2), 185-196.

Gavi. (1999). *First board meeting minutes.* Retrieved from http://libdoc.who.int/hq/1999/GAVI_99.02.pdf

Gelman, A. a. (2014). Why high-order polynomials should not be used in regression discontinuity designs. *NBER Working Paper 20405*.

Gelman, A., & Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business and Economic Statistics*, 447-456.

Hulls, D. P. (2010). *Second GAVI evaluation report.* Retrieved 5 2, 2020, from gavialliance.org: ww.gavialliance.org/library/gavi-documents/evaluation/second-gavi-evaluation-2006-2010

Imbens, G. a. (2012). Optimal bandwidth coice for the regression discontinuity estimator. *The Review of Economic Studies 79(3)*, 933-959.

Linden, A., & Yarnold, P. (2016). *Using machine learning to assess covariate balance in matching studies* (Vol. 22). Journal of Evaluation in Clinical Practice.

Lu, C. C. (2006). Effect of the Global alliance for Vaccines and Immunisation on diphtheria, tetanus, and pertussis vaccine coverage: An independent assessment. 1088-1095.

McCaffrey, D., Ridgeway, G., & Morral, A. (2004). *Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies* (Vol. 9). American Psychological Association.

Roodman, D. (2009). A note on the theme of too many instruments. *Oxford Bulletin of Economics and Statistics 71(1)*, 135-158.

Sarah Dykstra, A. G. (2015, 2). The Impact of Gavi on Vaccination Rates: Regression Discontinuity Evidence. *Center for Global Development 394*.