Economics Bulletin

Volume 39, Issue 1

India's Universal Immunization Program: a lesson from Machine Learning

Dweepobotee Brahma
Western Michigan University

Debasri Mukherjee
Western Michigan University

Abstract

This paper examines the predictors of immunization coverages of children across Indian states and evaluates the role of Universal Immunization Program (UIP) - a comprehensive policy of the Indian government in that light. Employing Machine Learning methods such as LASSO and hierarchical LASSO, we find that not the UIP expenditure by itself, but health infrastructure turns out to be a robust predictor of immunization coverage. The policy prescription that follows from our study is that the immunization program should focus on promoting the required health infrastructure in addition to monitoring the usage of funds closely for facilitating effective usage of the money. We also scrutinize performances of 'BIMARU', states that are considered traditionally underperforming states in terms of health and education.

The authors acknowledge useful comments from Allied Social Sciences Conference Annual Meeting, Philadelphia, 2018.
Citation: Dweepobotee Brahma and Debasri Mukherjee, (2019) "India's Universal Immunization Program: a lesson from Machine Learning", *Economics Bulletin*, Volume 39, Issue 1, pages 581-591
Contact: Dweepobotee Brahma - dweepobotee.brahma@wmich.edu, Debasri Mukherjee - debasri.mukherjee@wmich.edu.
Submitted: March 23, 2018. **Published:** March 16, 2019.

1. Introduction:

India has a startlingly high incidence of infant and maternal mortality resulting from vaccine-preventable diseases and alone contributes to about one-fifth of child mortality worldwide, mostly from such diseases.[1] The Universal Immunization Program (UIP) launched in 1985 is India's policy response to tackling this problem. However, as of 2015, the infant-mortality rate in India is still quite high, at about 38 per 1000 live births with a wide variation in the coverage of vaccinations as well as disbursement of funds under UIP across different states in India.[2] Several studies using household survey data have looked at the effects of household characteristics and UIP on immunization or the resulting anthropometric measures. *However, to the best of our knowledge, no empirical work has attempted to find the determinants of immunization at a macro (state) level and our work contributes to fill that gap. This is important for policy making. Although UIP funds are disbursed by the central government to the state governments, achieving targeted goals and policy implementations at the grass root levels are largely bestowed on state administrations. Indian states do display alarmingly high variances in terms of their socio-economic indicators* including *health and education, and that is precisely the reason why so called economically backward states (BIMARU) have been given special attention by policy makers over the years.*

We use a state level panel dataset spanning six years and 31 states and capture state specific unobserved heterogeneities (in terms of their socio-economic-cultural variations) using state dummies. We also use several state level observed covariates such as state level output per capita, child population under six, health infrastructure in states and UIP funds allocated to various states. Additionally, we include a time trend variable. Instead of treating all these regressors as being strictly independent of each other we further consider their *interaction terms to capture explicitly any possible complementarities among covariates*, *as well as state level or time wise variations in the effectiveness of the observed covariates on immunizations*.[3] This makes our number of regressors larger than the sample size, necessitating the use of Hierarchical LASSO regression (detailed to be explained in Section 3.2) - a new Machine Learning technique.

We consider number of people being vaccinated/covered (instead of vaccination rates) as dependent variable for each vaccine considered in our study. This is because our interest lies in the volume or level as opposed to rate which is often of interest to policy makers in health related studies - for example, number of deaths, number of people affected by an epidemic, number of people benefitted from a policy intervention etc. Our empirical ML based data mining exercises robustly indicate that although health infrastructure (to be defined in detail) is a crucial predictor for achieving immunizations, funds disbursed through the Universal Immunization Program is not. The upshots are that policy makers need to focus on building health infrastructure and can possibly use part of the fund for that purpose. Funds disbursed also need closer monitoring to prevent leakages. We also find variations across states in terms of *effectiveness of these policy variables*,

[1] See http://unicef.in/Whatwedo/3/Immunization

[2] While states like Karnataka and Andhra Pradesh report nearly 100% coverage under UIP, other states like Madhya Pradesh and Rajasthan report only about 75% coverage. There is also a wide variation in the funds received by states under UIP. Chandigarh received about one lakh under UIP while its neighboring state Himachal Pradesh received more than 200 lakh in a year and the difference can't be explained by population size alone.

[3] State dummies and time trend have been interacted with observed covariates in the regression.

and make interesting observations regarding the performance of BIMARU (so called economically backward) states.

2. Brief Background and Literature on Immunization

2.1: Brief Background of UIP in India

In 1978 India launched the Extended Program on Immunization (EPI) with the aim of immunizing all children with DPT, OPV, BCG and Typhoid by the first year of their lives. However, it was largely unsuccessful and managed to reach only urban areas. In 1985 India revamped and relaunched its immunization program - the Universal Immunization Program (UIP), this time in a phased manner aiming to cover all districts by 1989-90. This program also established the cold chain logistic system for proper storage and transportation of vaccines, and aimed to reduce mortality and morbidity from six vaccine-preventable diseases. Currently, UIP is run by the Immunization division under the National Rural Health Mission (NRHM) under the Ministry of Health and Family Welfare (MoHFW). All states submit a program implementation plan (PIP) which includes the projected requirement for funds to the MoHFW. The MoHFW disburses these funds to the states along with providing technical guidance about vaccination through the Immunization Technical Support Unit (ITSU). [4] The funds are used for vaccine logistics and cold chain system logistics, program monitoring and evaluation, monitoring adverse events following immunization and strategic communication. The utilization of these funds is supervised by the states through the State Health Societies under NRHM [5]. Thus, it is conceivable to find variations across the states in the implementation of the program. State level dummies and their interactions with infrastructure and funds variables have been used to capture this, while no other study (to our knowledge) has addressed these effects in such detailed manner.

2.2. Brief Literature on Immunization

The existing literature on UIP in the fields of economics, public policy and health policy concentrate on the short-term and long-term health effects on immunization exploiting survey data at the household or individual level. Patra (2006) uses household data from the National Family Health Survey-2 (1998-99) and estimates the effect of various demographic and socio-economic factors like income, education level, religion along with if the mother received any ante-natal care (binary variable) on the likelihood of child's immunization using a logistic regression. Datar, Mukherji, and Sood (2005) use the same survey dataset to estimate the effects of a constructed measure of health infrastructure (nearest primary health infrastructure facility available to a family and the presence of community health workers relevant to immunization) and community outreach programs on the vaccination status of a child. They found a positive and significant effect of the availability of health infrastructure on the coverage in certain situations. Using district level survey data, Anekwe and Kumar (2012) estimate the effect of being exposed to UIP (using a categorical variable) during the first year of a child's life on anthropometric outcomes and vaccination status.

[4] For a more detailed description of the policy implementation plan see the following reports by the Ministry of Health and Family Welfare at http://www.who.int/immunization/programmes_systems/financing/countries/cmyp/india_cmyp_2013-17.pdf?ua=1 and also at https://www.nhp.gov.in/sites/default/files/pdf/immunization_uip.pdf

[5] See the discussion in section 2.2 in https://www.who.int/immunization/programmes_systems/financing/countries/cmyp/india_cmyp_2013-17.pdf?ua=1

They find a positive effect of UIP on an anthropometric measure. Banerjee et al. (2010) conduct a clustered randomized controlled trial in 134 villages in rural Rajasthan and find that in a low-immunization coverage setting, immunization camps (their measure of infrastructure) combined with small incentives such as raw lentils and metal dishes have a positive impact on immunization coverage. Thus, micro data finds positive impact of vaccination on anthropometric measures of children, modest effect of UIP funds, and favorable effect of incentives and health infrastructure on the likelihood of child being vaccinated (fully or partly). These studies mostly focus on overall vaccination status of children, whereas we consider immunization coverage of individual vaccines under the program. Thus our study is at a more disaggregated vaccine level compared to the existing ones. We consider five vaccines which the program started with (refer to Table II for more details), namely BCG (Baccilus Calmette Guerin), OPV (Oral Polio Vaccine), Measles, DPT (diphtheria, tetanus and pertussis (whooping cough)) and TT (Tetanus Toxoid for pregnant women). We run separate regression for each vaccination coverage as a dependent variable. Following Datar et al. (2005) we include the total number of operating health centers in each state as a measure for health infrastructure in our regression. Note that our study complements the aforementioned micro household level studies. While the Banerjee et al. (2010) measure of health infrastructure is the temporary immunization camps at the village level (as part of their experiment), our measure of health infrastructure include the state-level primary health centers and community health centers as in Datar et al. (2005). Also note that while the Banerjee et al. (2010) finds effectiveness of temporary health infrastructure in conjunction with small incentives in a localized setting, our paper finds the effectiveness of permanent state level public health infrastructure (similar to findings in Datar et al. (2005)). To the best of our knowledge we do not find any macro-state level study on UIP and vaccination coverage. Our macro state level study corroborates some of these finding from micro/regional studies. Additionally, it also assesses the effectiveness of policies in various states and draws interesting conclusions regarding BIMARU (economically backward) states.

3. Data and Econometric Analysis

3.1. Data

We use state-wise annual longitudinal data on coverage of five vaccines (BCG, DPT, Measles, OPV and TT) and total funds disbursed under UIP for the years 2005-06 to 2007-08 and 2009-10 to 2011-12 from Indiastat.com.[6] The Union territories of Andaman and Nicobar Islands, Lakshadweep, Dadra and Nagar Haveli, and Daman and Diu were dropped due to unavailability of data. Information on funds for the year 2008-09 was unavailable for all the states. A few other observations where the value of funds disbursed under UIP were missing were dropped from the analysis. The main analysis was conducted using 31 states and six years. For TT, the coverage refers to the total number of pregnant women whereas for other vaccines it refers to the total number of children vaccinated. The data on state-wise release of funds under UIP (measured in rupees), is adjusted for inflation using the price level in 2011. We also include the state-wise per capita net state domestic product (SDP) at factor cost at constant prices. A key covariate in this study is our measure of health infrastructure. It is the total number of Primary Health Centers (PHCs) and Community Health Centers (CHCs) in a state in a given year since both PHCs and

[6] We do not find data on more recent years, or specific funds disbursed for various vaccines.

CHCs are responsible for administering vaccines under UIP. Another important covariate is the size of the population of the state. While there is a wide variation of population across the states, state-level data for this variable is not collected on a yearly basis. The most relevant and closest measure that we find for our purpose is the number of children under 6 years which was available only for the census year, 2011. We use the values for 2011 as repeated observations for each state to control for their population.[7] A brief summary statistics of our variables are provided in Table I.

3.2 Econometric Analysis: Challenge and Methodology

As noted earlier, we are interested in state level panel analysis. We have a total of 175 observations (owing to many missing values in the 31 states, 6 years panel). The short panel nature of our data makes it inadequate for any typical non-parametric or semi-parametric regression approach such as that in Bhaumik et al. (2018). A researcher might use a standard fixed or random effects panel regression without accounting for all possible interaction effects in the covariates. But, ignoring those interactions will essentially mean *a priori* that they are independent when they are actually not. We not only include state dummies (as is done in a typical fixed effects model) and time trend but also consider their interactions with observed covariates. Interactions of the observed covariates with state dummies and time trend also allow us to capture the variations in their effectiveness across states and over time.

Why LASSO? Note that, consideration of all such interactions lands us in a 'high-dimensional setting' of number of regressors exceeding the sample size. As a result the traditional regression methods fail and new Machine Learning (ML) techniques are called for. LASSO (Least Absolute Selection and Shrinkage Operator), a sparsity-based ML method is *tailor-made* to deal with such high-dimensional problem where K (number of regressors) exceeds or is very close to the sample size (n). LASSO also selects regressors by using out-of-sample prediction performance maximization criterion.[8] In addition, Hierarchical LASSO (a modified version of LASSO) is further tailor-made to capture *covariate interaction effects* along with base covariate effects appropriately when degrees of freedom can be negative.[9] We use ordinary LASSO results simply as benchmark but focus on the results from Hierarchical LASSO. See Bien, Taylor, and Tibshirani (2013) and Hastie, Tibshirani, and Friedman (2009) for details of the algorithms on Hierarchical LASSO and LASSO respectively. In our Hierarchical LASSO (results in Table III), *while sample size is 175, the number of covariates in the model is 195* to begin with, including all base covariates (35 dummies and other variables) plus the interaction terms (observed variables interacted among themselves and interacted with state dummies and time trend). But then this sparsity algorithm highly shrinks the set of covariates selecting only few as reported in Table III.

Hierarchical LASSO solves the following constrained maximization problem with L_1 norms:

---

[7] For TT vaccine targeted to pregnant women we use both (i) total adult female population as well as (ii) children under 6 years for relevant 'population' measure, and obtain qualitatively similar results. More appropriate measures are not available.

[8] Other ML techniques such as Random Forest can't be used in our setting because they deal with very large N as opposed to large K.

[9] If there are two covariates, say *x1* and *x2*, for example and we consider the interaction term *x1*x2*, Hierarchical LASSO can read the *x1*x2* as it is, instead of reading it as a new variable, say *z*.

$$\text{Minimize} \sum_{i=1}^{n} \left(Y - \beta_0 - \sum_{j=1}^{p} \beta_j X_j - \frac{1}{2} \sum_{j \neq k}^{p} \theta_{jk} X_j X_k \right)^2 + \lambda \sum_{j=1}^{p} \left|\left|\beta\right|\right|_1 + \frac{\lambda}{2} \sum_{j \neq k}^{p} \left|\left|\theta\right|\right|_1 \qquad (1)$$

Here *Y* is the dependent variable measuring vaccination coverage, *X* is the matrix of covariates, *i(=1,..., n)* refers to the index representing sample observations, *j, or k (=1,...,p)* refers to the index representing covariates. β and θ are the coefficient estimates of the main effects and interaction terms respectively. λ, is the tuning parameter which controls the degree of regularization and is chosen through an out-of-sample prediction error minimization criterion known as cross-validation (explained later).

Hierarchical LASSO produces sparse estimates of β and θ honoring either one of the two hierarchy conditions.

(i) Strong hierarchy where $\theta_{jk} \neq 0$ when $\beta_j \neq 0$ *and* $\beta_k \neq 0$
(ii) Weak hierarchy where $\theta_{jk} \neq 0$ when $\beta_j \neq 0$ *or* $\beta_k \neq 0$

In studying macro-developmental issues, the strong hierarchy condition is too restrictive in terms of predictor selection. This is because there is no reason to assume that the interaction of two covariates is a good predictor *only* when both the main effects are individually good predictors. We employ the weak hierarchy condition which chooses an interaction term as a good predictor if *at least* one of the main effects are chosen as a good predictor. Thus, weak hierarchy is not only less restrictive but also more intuitive than strong hierarchy. For example, a state itself may not stand out in terms of performance (compared to average performance of other states) but it may show up as an important predictor while interacted with infrastructure, which would mean infrastructure is an important predictor for vaccination performance in that state. Weak hierarchy will pick this effect while strong hierarchy will not.

As stated before, although our interest lies in capturing interactions, we also run ordinary LASSO with no interaction terms (results in Table IV) as a baseline regression. Since there are no interaction terms, K for this is 35 (not a high dimensional setting for 175 observations). This is still presented as a baseline comparison to Hierarchical LASSO. This optimization equation (without the interactions) now is as follows:

$$\text{Minimize} \sum_{i=1}^{n} \left(\mathrm{Y} - \beta_0 - \sum_{j=1}^{p} \beta_\mathrm{j} \mathrm{X_j} \right)^2 + \lambda \sum_{\mathrm{j}=1}^{\mathrm{p}} \left|\left|\beta\right|\right|_1 \qquad (2)$$

The notations are similar to the ones used in Equation (1).

In general, the choice of the tuning parameter λ is a crucial decision in regularized models since λ controls the degree of regularization of each model and hence determines variable selection. The commonly used method for this purpose is cross-validation which is based on the principle of minimizing out-of-sample prediction error. We use standard 10-fold cross-validation

to choose the tuning parameter λ for both LASSO and Hierarchical LASSO algorithms. We run separate regressions for each of five vaccines. We then report the selected predictors with their respective signs for each vaccine following the one-standard error rule (as is common in the literature) in Tables III-IV and check if funds and health infrastructure (our policy variables) are identified as important predictors. We also check if any state underperforms (compared to median state) in any of the vaccination coverages or in terms of policy effectiveness.

Machine learning techniques like LASSO are designed to produce best out-of-sample predictions using sparsity by selecting important regressors and dropping other covariates. However, the estimated values of the marginal effects and their standard errors produced by LASSO can be biased and hence should not be used for causal inference. *But the results from LASSO can still prove useful for policy analysis by shifting the perspective from 'estimating marginal effects' to 'investigating whether the policy serves as an important predictor of the outcome' or not. See, Mullainathan and Spiess (2017) for a detailed discussion on this.* It is also well known that a good in-sample prediction (or any high measure of 'goodness-of-fit') based on in-sample explanatory power of the model does not necessarily imply a good out-of-sample prediction performance and the latter is much more important from a realistic point of view, though much harder to achieve. LASSO algorithms are based on the principle of minimizing out-of-sample prediction errors, and hence their practical advantage can't be overstated. Nevertheless, in line of standard econometric literature we also provide detailed illustration of the in-sample performance of the method we use by comparing the kernel densities of BCG coverage (observed values) and its in-sample predictions (emanating from our estimation). This is presented in Figure 1, displaying a good fit. We have very similar looking figures for other vaccines, but these are not reported for brevity.

4 Results and Concluding Discussions:

4.1 Results:

Our Hierarchical LASSO and ordinary LASSO results are reported in Tables III and IV respectively indicating important predictors selected and their respective signs.[10] The algorithms robustly choose health infrastructure as a good predictor of immunization across all vaccination coverages. However, the UIP funds variable almost never get selected except in case of TT vaccine under Hierarchical LASSO analysis. We also find population size to be a good predictor as expected. Even after controlling for population size, we find some populous states such as Uttar Pradesh (UP), Madhya Pradesh (MP) and Andhra Pradesh (AP) are selected with positive signs whereas a different but largely populous state named Bihar gets selected with negative signs (in most cases). Presumably, we can interpret it as: MP and AP, achieving relatively more vaccination coverages than the median achiever (state of Punjab) while Bihar achieving relatively less than Punjab, even after controlling for (child) population size variable.[11] Interactions of state specific effects with policy variables reveal some interesting facts in Hierarchical LASSO. We find that

[10] In LASSO estimation, the regression coefficient estimates are biased and therefore not used for inference purposes, hence not reported. The focal point is covariate selection and their signs which have been reported.

[11] Punjab is a median achiever for all vaccination coverages and dummy for this state has been dropped to avoid perfect collinearity problem (or so called dummy variable trap issue). Thus performance of other states can be assessed relative to this state.

both funds and infrastructure work well in UP in particular. Note, Bihar, MP, Rajasthan and UP have long been considered as backward states in terms of health and education indicators and special attention has therefore been provided by researchers and policy makers to those states. Our analysis reveals that MP is a relatively good performer for all vaccines, UP and Rajasthan are good performers for TT (vaccines given to pregnant women), while Bihar is a relatively bad performer for all vaccines. We also find that over time vaccination coverage has decreased as demonstrated by the selection of the time trend term with a negative sign. Therefore, more attention still needs to be provided for BIMARU states (except MP) and especially on Bihar.

4.2 Discussions:

While the existing studies mostly use micro household level data, standard regression techniques, and likelihood of overall vaccination coverage, we focus on macro-state level data, each vaccine under UIP separately, covariance interaction effects and use a new regression technique from Machine Learning literature that caters to our setting. Our analysis complements the existing micro studies and some clear policy prescriptions emerge. Our results suggest that there is a need to have more health infrastructure (health centers) in the states. While micro studies can't examine the effects of UIP funds disbursed to various states, we can. Our results clearly indicate that funds disbursed must be scrutinized more carefully. There is no denying the fact that in India much of aid money is wasted or misused due to lack of governance and corruption, and our results reiterate that problem in the context of vaccination funds. Various reports have found high rates of vaccine wastage (owing to lack of appropriate cold storage facilities in tropical India) as an important hindrance to vaccination.[12] Part of the funds should be spent towards such infrastructure building. Our measure of health infrastructure (PHCs and CHCs) are the last points in the cold chain logistic systems,[13] thereby acting as proper storage facilities. Thus, states that have a higher number of these PHCs and CHCs are better able to store these vaccines properly and to achieve better success with vaccination coverage compared to other states. *This explains why our measure of health infrastructure is consistently selected as a good predictor of coverage.* Combining these we can suggest that not only is a rigorous monitoring of funds needed, but a part of funding can also be spent for building more health infrastructure (ensuring proper storage of vaccines and trained personnel) to make UIP more effective.

Our analysis identifies critical states for future micro level studies (including RCTs) by indicating some underperforming states, with respect of specific (individual) vaccination coverages. For example, a detailed micro level study in Bihar can unmask the factors contributing to less vaccination coverage in this state. For Kerala and Tamil Nadu, an RCT can focus on identifying reasons behind poor coverage of TT vaccines (for the pregnant mothers). Given that our measure of infrastructure is consistently selected as an important predictor, policy makers must first identify the 'infrastructure deserts' in each state and then construct new health centers in such locations. A well-spaced network of PHCs and CHCs within each state will reduce vaccination wastage. Following the leads of multilateral agencies, the government of India may also provide disaggregated administrative data on UIP funds (used for various purposes) to researchers, in order to allow for the identification of leakage-sources.

---

[12] See one such report in https://www.mofa.go.jp/mofaj/gaiko/oda/seisaku/kanmin/chusho_h24/pdfs/a20-12.pdf

[13] See https://www.nhp.gov.in/sites/default/files/pdf/immunization_uip.pdf under "Components" , Part II "Cold Chain System, Vaccines, Logistics".

Acknowledgements: Authors are very grateful to the editor of the journal and two anonymous referees for bringing attention to many important issues. All remaining errors are ours.

References:

Anekwe, Tobenna D., and S. Kumar (2012) "The effect of a vaccination program on child anthropometry: evidence from India's Universal Immunization Program". *J Public Health* **34 (4)**, 489-497.

Banerjee, Abhijit Vinayak, Esther Duflo, Rachel Glennerster, and Dhruva Kothari. (2010). "Improving Immunisation Coverage in Rural India: Clustered Randomised Controlled Evaluation of Immunisation Campaigns with and without Incentives." *BMJ (Online)* **340 (7759): 1291**. https://doi.org/10.1136/bmj.c2220.

Bhaumik, Sumon Kumar, Ralitza Dimova, Subal C. Kumbhakar, and Kai Sun. (2018). "Is Tinkering with Institutional Quality a Panacea for Firm Performance? Insights from a Semiparametric Approach to Modeling Firm Performance." *Review of Development Economics* **22 (1)**, 1–22. https://doi.org/10.1111/rode.12311.

Bien, Jacob, Jonathan Taylor, and Robert Tibshirani. (2013). "A Lasso for Hierarchical Interactions." *Annals of Statistics* **41 (3)**, 1111–41. https://doi.org/10.1214/13-AOS1096.

Datar, Ashlesha, Arnab Mukherji, and Neeraj Sood. (2005). "Health Infrastructure and Immunization Coverage in Rural India Health Infrastructure and Immunization Coverage in Rural India." *RAND Health*. Vol. WR-294.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. (2009). *The Elements of Statistical Learning*. *Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7.

Mullainathan, Sendhil, and Jann Spiess. (2017). "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* **31 (2)**, 87–106. http://www.aeaweb.org/articles?id=10.1257/jep.31.2.87.

Patra, Nilanjan. (2006). "Universal Immunization Programme in India: The Determinants of Childhood Immunization." *Ssrn*. https://doi.org/10.2139/ssrn.881224.

APPENDIX

Table I : Descriptive statistics of all the variables

	Minimum	Maximum	Mean	Median	Standard Deviation
B.C.G	8045	5693227	834455.4	539523	1094819
Measles	8590	5330048	763465.6	510755	1002537
O.P.V	8105	5544695	781922.5	524453	1032407
T.T	7260	5689521	790056.1	519934	1048690
Real Funds (in INR)	63822	353538679	40562010	15583194	58903317
Per-capita SDP	7588	129397	39687	34096	22730
Infrastructure	1	4207	893	548	940.8
Under 6 pop	64111	30791331	5401406	3380721	6720341
Female pop	287507	95331831	19182168	12712303	21751093

Note: The sample size was 175. The target group for TT vaccine is pregnant women. The target group for all other vaccines are children.

Table II: Details of the vaccines considered in this study

Vaccine	Disease(s)	Schedule
BCG	Tuberculosis	at birth (up to one year)
DPT	Diphtheria, Pertussis, Tetanus	6,10,14 weeks
OPV	Poliomyelitis	6,10,14 weeks
Measles	Measles	9-12 months, 16-24 months
Tetanus Toxide (TT)	Tetanus (in pregnant women)	2 doses during the pregnancy

Table III: Predictors selected via Hierarchical LASSO for each vaccine

BCG	DPT	OPV	Measles	TT
Infrastructure (+)	Infrastructure (+)	Infrastructure (+)	Infrastructure (+)	Infrastructure(+)
Population (+)	Population (+)	Population (+)	Population (+)	Population (+)
Trend (-)	Trend (-)	Trend (-)	Trend (-)	Trend (-)
Infra*trend (-)	Infra*trend (-)	Infra*trend (-)	Infra*trend (-)	Infra*trend (-)
Andhra Pradesh (+)	Andhra Pradesh (+)	Andhra Pradesh (+)	Andhra Pradesh(+)	Andhra Pradesh (+)
Madhya Pradesh (+)	Madhya Pradesh (+)	Madhya Pradesh (+)	Madhya Pradesh (+)	Madhya Pradesh (+)
AP*SDP(+)	AP*SDP(+)	AP*SDP(+)	AP*SDP(+)	AP*SDP(+)
MP*SDP(-)	MP*SDP(-)	MP*SDP(-)	MP*SDP(-)	MP*SDP (-)
UP*population(+)	UP*population(+)	UP*population(+)	UP*population(+)	UP*Population (+)
UP*trend(-)	UP*trend(-)	UP*trend(-)		
				UP*funds (+)
				UP*infra (+)
Bihar*Population (-)	Bihar*Population (-)	Bihar*population (-)	Bihar*population (-)	
			Bihar*trend (+)	Bihar *trend (+)
				Bihar (-)
MP*Trend(-)				MP*trend (-)
West Bengal(+)				Chhattisgarh (+)
Population*trend (-)				Gujarat (+)
				Haryana (+)
				Kerala (-)
				Rajasthan (+)
				Tamil Nadu (-)
				Rajasthan *trend (-)
				Funds(+)

Note: AP, MP and UP stand for Andhra Pradesh, Madhya Pradesh and Uttar Pradesh respectively.

Table IV: Predictors selected via LASSO (without considering any interaction terms)

BCG	DPT	OPV	Measles	TT
Infrastructure(+)	Infrastructure(+)	Infrastructure(+)	Infrastructure(+)	Infrastructure(+)
Population(+)	Population(+)	Population(+)	Population(+)	Population(+)
Trend(-)	Trend(-)	Trend(-)		
Andhra Pradesh(+)	Andhra Pradesh(+)	Andhra Pradesh(+)	Andhra	Bihar(-)
Bihar(-)	Bihar(-)	Bihar(-)	Pradesh(+)	Madhya
Madhya	Madhya	Madhya	Bihar(-)	Pradesh(+)
Pradesh(+)	Pradesh(+)	Pradesh(+)	Madhya	Uttar Pradesh(+)
Uttar Pradesh (+)	Uttar Pradesh(+)	Uttar Pradesh(+)	Pradesh(+)	Rajasthan(+)
West Bengal(+)			Uttar Pradesh(+)	Kerala(-)

Figure 1: Kernel density plot comparing in sample prediction to observed values