A Prediction Model Of Rotavirus Vaccine Performance Based On Meta-Analysis And Variable Selection

Yueqi Li
liyueqi0131@gmail.com

Follow this and additional works at: https://elischolar.library.yale.edu/ysphtdl

Part of the Public Health Commons

Recommended Citation
https://elischolar.library.yale.edu/ysphtdl/2174

This Open Access Thesis is brought to you for free and open access by the School of Public Health at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Public Health Theses by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.
A Prediction Model of Rotavirus Vaccine Performance Based on Meta-analysis and Variable Selection

Yueqi Li

Year of Completion: 2022
Year Degree Awarded: 2022
Degree Awarded: Master of Public Health
Yale School of Public Health

Committee Chair: Virginia Pitzer, ScD
Committee Members: Daniel Weinberger, PhD
Abstract

Background: Rotavirus is the leading cause of diarrheal diseases worldwide and is responsible for substantial morbidity and mortality among children younger than 5 years of age. The introduction of national immunization programs against rotavirus has greatly reduced hospital admissions and deaths from rotavirus gastroenteritis, but vaccine efficacy and effectiveness vary substantially between high- and low-income settings. Given the disparity of vaccine performance across countries, we built a model to predict how well the vaccines would work in settings without vaccine introduction.

Methods: Data on vaccine performance and relevant predictor variables were extracted from the literature and repositories. We performed a meta-analysis and subgroup analysis to evaluate vaccine efficacy and effectiveness and to examine potential sources of heterogeneity. Finally, we implemented a meta-regression, testing several variable selection methods, and cross-validation models to identify the best set of predictors of vaccine efficacy and effectiveness.

Results: The visualization of vaccine performance disparity among countries was consistent with our prior understanding. Using subgroup analysis, we found that heterogeneity in vaccine efficacy and effectiveness could be explained by income level. Based on our meta-regression analysis, the best predictor combination at the study level was under-five mortality, Global Burden of Disease region (GBD), and their interaction. At the country level, diarrhea prevalence, GDP per capita, and study indicator were the best predictor combination.

Conclusion: Variability in published estimates of rotavirus vaccine efficacy and effectiveness can be explained by country income level, diarrhea prevalence, and under-five mortality. Our prediction model is a critical tool to understand how a future rotavirus vaccine program would help to decrease mortality and morbidity in countries with high diseases burdens and no vaccination implemented.
Acknowledgements

I would like to thank my thesis advisors, Dr. Virginia Pitzer and Dr. Daniel Weinberger, for their guidance and mentorship throughout the process. In addition, I would like to express my deepest appreciation to Ottavia Prunas for her support in method consultation and proofreading. I’d also like to express my gratitude to all colleagues in Pitzer and Weinberger lab, and professors at the Yale School of Public Health who have given me the tools and knowledge to be a public health professional. Finally, I would like to acknowledge my family and friends for their continued support throughout my MPH program.
Table of Contents

Introduction .................................................................................................................................4

Methods ........................................................................................................................................6
  Data Source ...............................................................................................................................6
  Statistical Analysis ...................................................................................................................7

Results ........................................................................................................................................10
  1. Descriptive analysis of included studies .............................................................................10
     Figure 1. Study inclusion and selection process ......................................................................10
     Figure 2. Disparity of vaccine performances in HIC and LMIC settings ...............................12
     Figure 3. Forest plot of relative risk estimates from RCTs of rotavirus vaccine efficacy. ........13
     Figure 4. Forest Plot of Meta-analysis of Case-control studies ..............................................14
  2. Subgroup Analysis stratified by income level .......................................................................15

3. Predictor selection at study level .........................................................................................16
   Figure 5. Collinearity of predictor variables in Meta-regression .............................................17
   Table 1. Meta-regression outcome statistics ..........................................................................18

4. Predictor selection at country level .....................................................................................18
   Table 2. Selected predictor variables with different methods ..................................................19
   Table 3. Cross validation output of the best prediction model ................................................20
   Figure 6. Model-based vaccine prediction ..............................................................................22

Discussion ................................................................................................................................23

Reference ..................................................................................................................................26

Appendix ..................................................................................................................................33
  1. Multi-collinearity of variables included in country level analysis .......................................33
  2. Principal Component Analysis .............................................................................................34
  3. Cross validation output of different predictive models ..........................................................36
Introduction

Diarrheal diseases were the third leading cause of death among children younger than 5 in 2017 [1], accounting for approximately 8% of all deaths worldwide. That translates to 70.6 deaths per 100,000 population, or about 525,000 children a year, despite the availability of prevention methods. Rotavirus, which is responsible for nearly 30% of all diarrheal cases, is the single largest causative agent of children's diarrhea and diarrhea-related death. Rotavirus infection was responsible for more than 258 million episodes of diarrhea among children under age 5 in 2016 and causes 122,000 to 215,000 deaths worldwide each year [2].

Fortunately, the mortality of Rotavirus Gastroenteritis (RVGE) decreased by 43.6% from 2005 to 2015, which is most likely attributable to the introduction of rotavirus vaccines [3]. Four live-attenuated rotavirus vaccines are prequalified by World Health Organization: Rotarix (a 2-dose monovalent vaccine; prequalified in 2009), RotaTeq (a 3-dose pentavalent vaccine; prequalified in 2008), Rotavac (prequalified in 2018) and Rotasiil (prequalified in 2018). Except the most recently prequalified vaccines -- Rotavac and Rotasiil -- currently only in use in India and Palestine, Rotarix and RotaTeq are implemented globally. The rotavirus vaccine program has been introduced in 114 countries as of January 2022 [4]. Despite the high coverage of vaccination programs, vaccine protection showed a huge gap between high-income countries (HICs) and low- and middle-income countries (LMICs) [5]. The vaccine performance measurements include vaccine efficacy and vaccine effectiveness. Vaccine efficacy is measured in a controlled clinical trial while vaccine effectiveness is a measure of how well vaccines work in the real world [6]. The vaccine efficacy in HICs is 98-100% and the real-world vaccine impact in HICs is also substantial. In contrast, the vaccine only showed 43-64% protection in LMICs settings, where the majority of rotavirus deaths occur.
While post-implementation vaccine effectiveness data and surveillance systems showed a substantial decrease in severe RVGE and rotavirus hospitalization in HICs, there is little evidence regarding post-introduction changes in low-income settings. Given the randomized controlled trials showed lower vaccine efficacy in low-income settings, we aim to better understand the real-world performance after vaccine implementation in those countries. Meanwhile, for countries with high diarrhea burden and currently without vaccine introduction, we want to predict the vaccine performance by identifying predictors for extrapolation and estimating the benefits of vaccine introduction in those settings.
Methods

Data Source

Search strategy and inclusion criteria

We updated previous systematic reviews [1] by applying the search criteria and searching articles published until October 1, 2021. We included randomized controlled trials (RCTs) [7–36] for rotavirus vaccine efficacy studies and case-control designs [29, 36–78] for vaccine effectiveness studies. We extracted data on case counts, person-time of follow-up, and details of study design. Studies with fewer than 100 enrolled participants were excluded. We extracted the outcome information including vaccine efficacy against severe RVGE (Vesikari Score ≥ 11), and vaccine effectiveness against rotavirus associated hospitalization.

Selected predictor variables

We gathered data on widely available possible predictors of rotavirus vaccine efficacy and effectiveness from publicly available databases (Institute for Health Metrics and Evaluation, World Bank Open Data, World Health Organization Data). Potential predictors were selected based on the relevance to disease transmission, including indicators of environmental characteristics and socioeconomic development level. We identified ten different variables, including Global Burden of Disease region (GBD) super region classification, diarrhea prevalence among children age <5, population density, under 5 mortality, crude birth rate, gross domestic product (GDP) per capita, undernourishment prevalence, Gini index, income share held by lowest 20%, urban population (% of total). We collected the time when each study starts and ends, and pulled the predictor variable that matches the year in the middle to account for the changes of each predictor variable during the study period. The value of study level covariates was calculated as the average of this predictor variable among all included countries.
Statistical Analysis

Data visualization

To examine potential geographic clustering pattern of vaccine performance, we applied the R ggplot2 package to visualize the vaccine efficacy and effectiveness differences among regions and countries.

Meta-analysis

For the meta-analysis, we included different effect sizes to evaluate the vaccine performance in RCTs versus case-control studies. Relative risks (RRs) were calculated in the RCTs and used as the effect size for vaccine efficacy estimation. The RRs were calculated using the number of participants and the number of cases in the vaccinated and unvaccinated population, respectively. Odds ratios (ORs) were calculated in the case-control studies and as the effect size for vaccine effectiveness estimation. The ORs were extracted from the original paper. We used the per-protocol estimates in RCTs and combined control groups in case-control studies and applied both the fixed and random-effects models. We included 29 studies in the RCTs meta-analysis, 45 studies in the case-controls meta-analysis, and 74 studies in the combined meta-analysis. The fixed-effect model applied the inverse variance weighted method and assumed the effect sizes of included studies were homogeneous, while the random effects model assumed the effect size among studies followed a distribution and is a better option when the heterogeneity within studies is large. The overall RR and OR were calculated in the Mantel-Haenszel method and the heterogeneity between studies was evaluated by $I^2$. The $I^2$ describes the percentage of variation across studies.

To investigate the pattern of heterogeneity in our included studies, we further conducted subgroup analysis. We stratified the countries into High-income, Middle-income, and Low-
Income countries based on World Bank country classifications. We combined the Upper-middle income and Lower-middle income countries in the World Bank classification as Middle-income countries in our income subgroup. Subgroup analysis was performed with the R *metafor* package.

All the statistical analyses were performed using R (version 4.0.2).

**Meta-regression**

Vaccine efficacy (RR estimates) and vaccine effectiveness (OR estimates) at the study level were the outcome of interest for the meta-regression analysis. The meta-regression analysis was performed with the R *meta* package.

**Principle component analysis**

Given many of our predictor variables are highly correlated, we conducted principal component analysis (PCA) to reduce the number of variables while preserving as much information as possible.

**Predictor selection at country level**

We implemented several predictor selection methods: step-wise selection from both directions based on Akaike information criterion (AIC), Least Absolute Shrinkage and Selection Operator (Lasso) regression, spike-and-slab regression, and random forest. The full model is a linear regression of the outcome (on the logarithm scale) with all predictor variables included, while the null model includes none of the variables. Step-wise selection deals with the trade-off between the goodness of fit and the simplicity of the model. Lasso regression uses a shrinkage penalty to encourage a simpler/sparse model. It is suitable for models with highly correlated predictors or when only a few predictors influence the response. Spike-and-slab regression is a Bayesian variable selection method that is useful when the number of possible predictors is
larger than the number of observations. Random forest is an ensemble learning method that constructs decision trees for regression. The variable selection criterion evaluates how much each feature decreases the variances. Features that are selected at the top of decision trees are more important than features at the bottom nodes of decision trees. We also performed the variable selection method on the combined dataset, adding study-type indicator as a binary variable to describe the observation type (i.e. RCT or case-control study).

**Cross validation**

We implemented cross-validation to assess the performance of different predictive models and to evaluate their performance on an independent data set outside the sample data used for model fitting. We included different variable combinations to predict the vaccine efficacy and effectiveness measures in the combined dataset. We applied three cross-validation methods: Leave One Out Cross-Validation (LOOCV), cross-validation (10 fold), and repeated cross-validation (10 fold, repeated 5 times). LOOCV considers each observation as the validation set and the remaining observations as the training set. Ten-fold cross-validation randomly shuffles the dataset and splits it into 10 groups; each group is treated as the test dataset and the remaining group as the training set. A model is fit to the training set and then evaluated on the test set.
Results

1. Descriptive analysis of included studies

Study selection and characteristics

The study inclusion and selection process are shown in Figure 1. 74 studies (29 RCTs and 45 case-control studies) were included after the selection process. As for vaccine types, 54 studies for Rotarix and 35 studies for RotaTeq were included.

Figure 1. Study inclusion and selection process.
95 literatures were extracted and reviewed; studies with fewer than 100 participants were excluded. After the selection process, 29 RCTs and 45 case-control studies were included.

Vaccine performance disparity

Figure 2 visualizes the worldwide vaccine performance among the included studies. Vaccine efficacy is evaluated as one minus the relative risk reduction against severe RVGE during the follow-up period. Vaccine effectiveness is measured as one minus the odds ratio of rotavirus hospitalization rate among vaccinated children against unvaccinated children during the follow-up period. Both vaccine efficacy and vaccine effectiveness exhibit the same trend,
showing higher protection in high-income countries, with over 90% reduction in severe RVGE occurrence or rotavirus hospitalization among the vaccinated group of North American and European countries. On the contrary, the vaccines showed poor protection against RVGE in low-income settings, with vaccine efficacy and effectiveness estimates only 30-50% in Africa. The data was incomplete, as we did not have data for vaccine efficacy and effectiveness in most part of Africa, Middle East and Central Asia.

1) Vaccine Efficacy against severe RVGE
Figure 2. Disparity of vaccine performances in HIC and LMIC settings.
Plot 1) shows the vaccine efficacy estimates available for different countries throughout the world according to the color bar; similarly, plot 2) shows the vaccine effectiveness estimates. Countries in grey do not have vaccine efficacy/effectiveness estimates available in the literature.

Meta-analysis of included studies

First, we conducted a meta-analysis of vaccine efficacy estimates from RCTs (Figure 3).

For the 29 studies included in our analysis, the overall estimate of the relative risk for the fixed effect model was 0.31 (95% confidence interval (CI): 0.29, 0.34) with a p-value <0.0001, resulting in a vaccine efficacy of 68.9% (95% CI: 66.4%, 71.2%). The estimation of relative risk in the random effect model was 0.22 (95% CI: 0.16, 0.31) with a p-value <0.0001, resulting in a vaccine efficacy of 77.8% (95% CI: 68.7%, 84.3%). There was a high variation across the 29 studies due to heterogeneity rather than chance ($I^2=88.9\%$). Given the high heterogeneity within our analysis, the random-effects model is the preferred option.
Figure 3. Forest plot of relative risk estimates from RCTs of rotavirus vaccine efficacy.
Relative risk of severe RVGE across the 29 RCTs, and the calculated overall relative risk among all studies, with associated \( I^2 \) value.

The meta-analysis of case-control studies was showed in Figure 4. The estimation of OR in the fixed effect model was 0.14 (95% CI: 0.12, 0.15) with a p-value <0.0001, resulting in a vaccine effectiveness of 86.4% (95% CI: 85.0%, 87.8%). The estimation of OR in the random-effects model was 0.19 (95% CI: 0.14, 0.23) with a p-value <0.0001, resulting in a vaccine effectiveness of 81.5% (95% CI: 76.9%, 86.1%). The \( I^2 \) among all 45 studies was 80.0%. Similar
to the RCT analysis, the random-effect model is a more suitable approach since it takes the high heterogeneity among studies into account.

![Forest Plot of Meta-analysis of Case-control studies](image)

**Figure 4. Forest Plot of Meta-analysis of Case-control studies.**

Vaccine effectiveness across 45 studies, and the calculated overall vaccine effectiveness among all studies with associated $I^2$ value.
We combined RCTs and case-control studies and conducted a meta-analysis for all the 74 included studies. We pooled the study together assuming OR is a good approximation to the RR, as the risk in both case and controls are small. The estimation of effect size in the fixed effect model was 0.13 (95% CI: 0.12, 0.14) with a p-value <0.0001, resulting in a vaccine effect estimate of 87.2% (95% CI: 86.1%, 88.3%). The estimation of risk difference in the random-effect model was 0.20 (95% CI: 0.16, 0.24) with a p-value <0.0001, resulting in a vaccine effect estimate of 79.8% (95% CI: 75.8%, 83.8%). The I$^2$ among all 74 studies was 84.3%.

2. Subgroup Analysis stratified by income level

Given the high heterogeneity in the pooled studies, we conducted a subgroup analysis to investigate the pattern of heterogeneity in our included studies. There were vaccine effect estimates from 30 high-income, 28 middle-income, and 16 low-income countries included in the analysis. We use the random-effects estimation as the included studies were from a pooled effect size estimation which is a distribution rather than a true estimation. Also, the study variation is large thus random effects model was a better option. The random-effects model showed a standardized mean difference of 0.08, 0.24, 0.46, which corresponds to vaccine protection of 91.94%, 76.28%, and 53.94% for high-income, middle-income, and low-income countries, respectively. The test statistics of between groups score was calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, which was $Q = 192.22$ (p <0.0001), indicating the effect sizes were significantly different between the three income-stratified groups. The $I^2$ value was 75.7% and 77.7% for high-income and middle-income subgroups. The $I^2$ was smaller than the value from pooled studies, which indicates income could explain some of the heterogeneity. However, the remaining heterogeneity was still
considerably large, implying other factors also contributing to the variation between different studies. Our results indicate that income level is an important source that leads to our effect size heterogeneity.

3. Predictor selection at study level

*Predictor selection with Meta-regression*

Following our meta-analysis of the included studies, we investigated the most important predictors for the model prediction step. We implemented a meta-regression analysis to synthesize multiple studies while adjusting for the effects of available covariates. We explored how much heterogeneity could be explained by the predictor variables and looked at their interactions. We first checked the collinearity of continuous predictors (Figure 5). The plot showed a high percentage of correlation existed between predictor variable pairs.

We then evaluated how each predictor variable accounted for the between-study variation. We ran the meta-regression model with a single variable at a time, we reported the test statistics and the residual heterogeneity (Table 1). The top three variables that accounted for total variability were under 5 mortality, GBD region, and urban population, which explained 80.7%, 73.4%, and 71.7% of total heterogeneity, respectively.
Figure 5. Collinearity of predictor variables in Meta-regression.
The histogram shows the distribution of each predictor variable. The scatter plot with the red line indicates data distribution and the correlation coefficients show the correlation of predictor variable pairs.

After considering the importance of a single predictor, we tested several meta-regression models with different variable combinations and with interaction terms between the predictors. Multi-collinearity of our predictors could lead to overfitting and would threaten the validity of the predictor selection. We reduced the multicollinearity by removing the highly-correlated variables ($r \geq 0.8$). The dropped variables were crude birth rate, undernourishment prevalence, and percentage of urban population. The best model from the meta-regression approach was to incorporate under 5 mortality, GBD region, and their interaction. This predictor combination could account for 85.7% of the total heterogeneity.
Table 1. Meta-regression outcome statistics

<table>
<thead>
<tr>
<th>Included variable</th>
<th>$\text{tau}^2$ (SE = 0.0017)</th>
<th>$\text{i}^2$ (%)</th>
<th>$\text{R}^2$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBD region</td>
<td>0.0058</td>
<td>69.46%</td>
<td>73.44%</td>
</tr>
<tr>
<td>Under 5 Mortality</td>
<td>0.0037</td>
<td>60.09%</td>
<td>80.71%</td>
</tr>
<tr>
<td>Diarrhea</td>
<td>0.0078</td>
<td>75.60%</td>
<td>58.21%</td>
</tr>
<tr>
<td>GDP</td>
<td>0.0077</td>
<td>75.38%</td>
<td>59.96%</td>
</tr>
<tr>
<td>Urban Population</td>
<td>0.0055</td>
<td>68.62%</td>
<td>71.77%</td>
</tr>
<tr>
<td>Population Density</td>
<td>0.0174</td>
<td>87.28%</td>
<td>9.86%</td>
</tr>
<tr>
<td>Gini</td>
<td>0.0133</td>
<td>83.63%</td>
<td>29.64%</td>
</tr>
<tr>
<td>Income Lower</td>
<td>0.0159</td>
<td>85.94%</td>
<td>13.79%</td>
</tr>
<tr>
<td>Crude Birth Rate</td>
<td>0.0078</td>
<td>75.40%</td>
<td>59.82%</td>
</tr>
<tr>
<td>Undernourishment</td>
<td>0.0092</td>
<td>79.43%</td>
<td>54.51%</td>
</tr>
<tr>
<td>Under 5 Mortality + GBD region</td>
<td>0.0036</td>
<td>58.64%</td>
<td>81.59%</td>
</tr>
<tr>
<td>Under 5 Mortality * GBD region</td>
<td>0.0028</td>
<td>51.65%</td>
<td>85.68%</td>
</tr>
</tbody>
</table>

4. Predictor selection at country level

Instead of using the whole study as our analysis unit, we separated each study into different observations based on their study location (i.e., country). Eighty-two observations were included in the vaccine efficacy analysis, while 66 observations were included in the vaccine effectiveness analysis. Before implementing various variable selection methods, we first checked the multi-collinearity of the included variables (Appendix 1).

As the RRs and ORs were not normally distributed, we transformed the outcome variables to a log scale. We then conducted the data cleaning procedure by scaling the predictor variables to make sure they were in the same measurable range, and we excluded several
predictor variables. In particular, we excluded progression to secondary school (female %), the percentages of population using safely managed drinking water and using safely managed sanitation, and the prevalence of HIV, as the percentage of null value in these covariates were over 40%, and this would jeopardize the variable selection process.

**Variable selection through different methods**

The selected variables with different methods are shown in Table 2. With various methods implemented, the selected predictor variables were consistent between vaccine efficacy and effectiveness studies (Table 2). GDP per capita was most frequently selected across different methods. However, GBD super-region also showed significant importance using the spike and slab method, while not being selected by other methods. The study type was an important binary indicator for the model to distinguish between vaccine efficacy and effectiveness. The results of the combined study were mostly consistent with the results from the separate efficacy and effectiveness analysis.

**Table 2. Selected predictor variables with different methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Vaccine efficacy</th>
<th>Vaccine effectiveness</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AIC</strong></td>
<td>GDP</td>
<td>GDP</td>
<td>GDP, Study Indicator</td>
</tr>
<tr>
<td><strong>LASSO</strong></td>
<td>GDP, Diarrhea</td>
<td>GDP</td>
<td>Study Indicator, Diarrhea, GDP</td>
</tr>
<tr>
<td><strong>Spike and slab</strong></td>
<td>GBD_HighIncome, GDP</td>
<td>GDP</td>
<td>GBD_HighIncome, Study Indicator</td>
</tr>
<tr>
<td><strong>Random forest</strong></td>
<td>GDP, Under 5 Mortality, Population Density</td>
<td>Under 5 Mortality, Crude Birth Rate, GDP, Diarrhea</td>
<td>Diarrhea, Under 5 Mortality</td>
</tr>
</tbody>
</table>
**Model building and validation**

To build the country-level regression model, we included the union of selected variables by different methods. In particular, we included GDP per capita, study indicator, diarrhea prevalence among children age <5, GBD super region classification, and under 5 mortality. We built 6 models ranging from including all five predictors, to different combination of three predictor variables. The result of different models and cross-validation statistics are shown in Appendix 2. The best model included diarrhea prevalence, GDP and the study indicator as predictors. This model was selected based on a relatively high variance explained by the model (i.e., $R^2$), low prediction error which is reflected by the root mean square error (RMSE), and fewer variables included to reduce the bias. The cross-validation output of the best model is shown in Table 3.

**Table 3. Cross validation output of the best prediction model**

Best model: $\text{Ln(Outcome)} \sim \text{Diarrhea} + \text{GDP} + \text{Study Indicator}$

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>$R^2$</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOCV</td>
<td>18.76</td>
<td>0.035</td>
<td>8.71</td>
</tr>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>15.91</td>
<td>0.40</td>
<td>9.01</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)</td>
<td>14.88</td>
<td>0.40</td>
<td>8.72</td>
</tr>
</tbody>
</table>

*Study indicator = 0 when the observation is efficacy study, study indicator = 1 when the observation is effectiveness study

**Posterior model predictions**

We used our final model at the country level, with prevalence of diarrhea in children under 5, GDP, and study indicator as predictors, for our prediction step. We then built a country-
level generalized linear regression model, and we found that the coefficients of each predictor variable were 2.809, -4.363 and -2.062 for diarrhea, GDP and study indicator, respectively. The predicted vaccine efficacy and effectiveness for each country was showed in Figure 6. Countries that have a negative vaccine performance indicate the vaccine didn’t provide protection, which is hard to interpret in real-world settings. We replaced the negative number in the prediction model with zero and generated the geographic plot. The model filled the information gaps for countries without related data, and the estimated vaccine performance was in accordance with our observed pattern.
Figure 6. Model-based vaccine prediction

Plot 1) shows the model estimated vaccine efficacy globally; similarly, plot 2) shows the vaccine effectiveness estimates. The negative prediction number is showed as 0 in the plot.
Discussion

The result of our meta-analysis indicated substantial protection of rotavirus vaccine (RV1 and RV5), especially for severe RVGE and rotavirus-related hospitalization. However, huge variation existed in our pooled studies as the vaccine protection showed a similar pattern of high protection in North America and Europe while lower protection in Africa and South-east Asia in both vaccine efficacy and effectiveness estimates. We found the variability among different regions was partially attributed to income level and our result of subgroup analysis was in line with previously reported data [1-3].

We applied a meta-regression model at the study level and found the best covariates to explain the variation of vaccine outcome were under-five mortality, GBD super region classification, and their interaction. As the diarrheal disease was the third leading cause of death among children <5, our finding of under-five mortality could explain the major part of the study variation was consistent with our prior guess. Besides, we built a prediction model by including diarrhea prevalence in children age <5 and GDP per capita as the predictor variable and indicating the study type (RCT or case-control) to estimate the vaccine efficacy and effectiveness for a specific country. The predictor coefficient indicated diarrheal prevalence was negatively associated with vaccine protection. With higher background diarrheal prevalence, the vaccine showed poor protection against severe RVGE and rotavirus-related hospitalization. The poor vaccine uptake in those settings could be explained by the interference with other gastrointestinal pathogens and chronic gut inflammation in children. Also, children are more likely to be naturally exposed before their first vaccine dose in a setting with high diarrheal prevalence, such that the vaccination would not be able to elicit a desirable extent of protection. GDP per capita was positively associated with vaccine protection. This could be explained as children in high-

23
income settings are less likely to experience malnutrition which leads to gut inflammation and they have better water and sanitation infrastructures.

The prediction model can be applied to estimate the vaccine performance in a country without current immunization program. The model is a good source for estimating the vaccine performance when taking local diarrheal prevalence into account, and calculating how many rotavirus related hospitalization and death could be averted. We could also predict the vaccine effectiveness in a setting with recent vaccine introduction but without surveillance system to capture how well the rotavirus vaccine actually performs in real-world settings.

Our analysis has two main limitations. First, the number of studies stratified by income and at the country level was limited. Given we only have around 140 observations at the country level with a large proportion of high-income countries, the small dataset may involve bias in extrapolating vaccine protection in middle- and low-income settings. Besides, the further partition into training and validation sets results in even smaller numbers in the validation set and thus increase the validation bias. Secondly, we found most of our interested predictors had high correlation which may pose threats to our model building process, as the $R^2$ and RMSE didn’t show a huge improvement between models. However, such correlation would also have a positive effect when we don’t have information for a specific predictor and we could substitute it with other predictors without much decrease of prediction power. The additional flexibility of this model is beneficial especially when disease prevalence data and development indicators are hard to obtain in low income settings.

In conclusion, we built a model based on the most important predictors for vaccine efficacy and vaccine effectiveness evaluation. Since the value of each predictor variable from different countries can be obtained from publicly available database, it becomes straightforward
to estimate the vaccine performance in countries without vaccine implementation and to evaluate how much benefit could a vaccine program add to the current scenario. The next step of our study is to validate our model performance by collecting data from countries with new surveillance systems and comparing the model predicted vaccine effectiveness with post-introduction surveillance data.
Reference


[23] Vesikari T, Karvonen A, Ferrante SA, Kuter BJ, Ciarlet M. Sustained efficacy of the pentavalent rotavirus vaccine, RV5, up to 3.1 years following the last dose of vaccine.


Appendix

1. Multi-collinearity of variables included in country level analysis

The plot shows the correlation between 13 predictor variables that are included in the analysis. The scatter plot and the correlation coefficient indicate many variables are highly correlated (r ≥0.8). The histogram of the variable distribution indicates most of the variables are not normally distributed.
2. Principal Component Analysis
   As we included 15 predictor variables and some of those were highly correlated, we applied the dimensionality-reduction method by reducing the number of variables while preserving as much information as possible. We implemented a principal component analysis (PCA) to our outcome and predictor variables. We ranked the first 10 principal components with the percentage of the variability they explained by scree plot, and we plotted how each predictor contributes to the first and second principal component (Figure 6). We found that under 5 mortality, extreme poverty, urban population, GDP, and undernourishment had the highest influence on the first two principal components.
Figure 6. Principal component analysis of continuous predictor variables. Plot 1) is the scree plot and the histogram of percentage of variance that the first 10 principal components explained; Plot 2) is a score plot of the first two principal components.
3. Cross validation output of different predictive models

We built 6 models ranging from including all five predictors (GDP, study indicator, diarrhea, GBD income category, and under 5 mortality) to different combination of three predictor variables. Model 1 includes all 5 predictors. Model 2 and 3 include 4 predictors by excluding GBD and under 5 mortality, respectively. Model 4, 5, and 6 include 3 predictor variables and assess the performance of different predictor combinations.

Three evaluation statistics were compared. Root Mean Square Error (RMSE) is the standard deviation of the prediction errors, and tells us how concentrated the data is around the line of best fit. R-squared ($R^2$) measures the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Mean Absolute Error (MAE) measures the average magnitude of the errors in a set of forecasts, which indicates the accuracy for continuous variables.

The best model was Model 4, which included diarrhea prevalence, GDP and the study indicator as predictors. This model was selected based on a relative high variance explained by the model (i.e., $R^2$), low prediction error which is reflected by the root mean square error, and fewer variables included to reduce the bias.

Model 1: $\ln(\text{Outcome}) \sim \text{GBD SuperRegion} + \text{Diarrhea} + \text{GDP} + \text{StudyIndicator} + \text{Under5Mor}$

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOCV</td>
<td>18.94</td>
<td>0.023</td>
<td>8.83</td>
</tr>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>14.99</td>
<td>0.31</td>
<td>9.24</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)*</td>
<td>15.21</td>
<td>0.34</td>
<td>8.89</td>
</tr>
</tbody>
</table>

*prediction from a rank-deficient fit may be misleading

Model 2: $\ln(\text{Outcome}) \sim \text{GBD SuperRegion} + \text{Diarrhea} + \text{GDP} + \text{StudyIndicator}$

<table>
<thead>
<tr>
<th></th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOOCV *</td>
<td>18.84</td>
<td>0.032</td>
<td>8.92</td>
</tr>
<tr>
<td>Cross-Validated (10 fold)*</td>
<td>14.12</td>
<td>0.42</td>
<td>9.61</td>
</tr>
</tbody>
</table>
Cross-Validated (10 fold, repeated 5 times)*

<table>
<thead>
<tr>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.95</td>
<td>0.36</td>
<td>8.90</td>
</tr>
</tbody>
</table>

### Model 3: Ln(Outcome) ~ Under5Mor + Diarrhea + GDP + StudyIndicator

<table>
<thead>
<tr>
<th>LOOCV</th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>18.80</td>
<td>0.03</td>
<td>8.82</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)</td>
<td>15.14</td>
<td>0.33</td>
<td>8.53</td>
</tr>
</tbody>
</table>

### Model 4: Ln(Outcome) ~ Diarrhea + GDP + Study Indicator

<table>
<thead>
<tr>
<th>LOOCV</th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>15.91</td>
<td>0.40</td>
<td>9.01</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)</td>
<td>14.88</td>
<td>0.40</td>
<td>8.72</td>
</tr>
</tbody>
</table>

### Model 5: Ln(Outcome) ~ Diarrhea + StudyIndicator + Under5Mor

<table>
<thead>
<tr>
<th>LOOCV</th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>14.18</td>
<td>0.25</td>
<td>8.58</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)*</td>
<td>15.12</td>
<td>0.22</td>
<td>8.69</td>
</tr>
</tbody>
</table>

### Model 6: Ln(Outcome) ~ GDP + StudyIndicator + U5Mor

<table>
<thead>
<tr>
<th>LOOCV</th>
<th>RMSE</th>
<th>R2</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-Validated (10 fold)</td>
<td>15.89</td>
<td>0.42</td>
<td>8.76</td>
</tr>
<tr>
<td>Cross-Validated (10 fold, repeated 5 times)*</td>
<td>15.24</td>
<td>0.42</td>
<td>8.36</td>
</tr>
</tbody>
</table>